1. **开发PySpark 所需准备环境**
2. 安装python环境

下载安装Anaconda。详细见文档“***Anaconda安装及使用.docx***”。

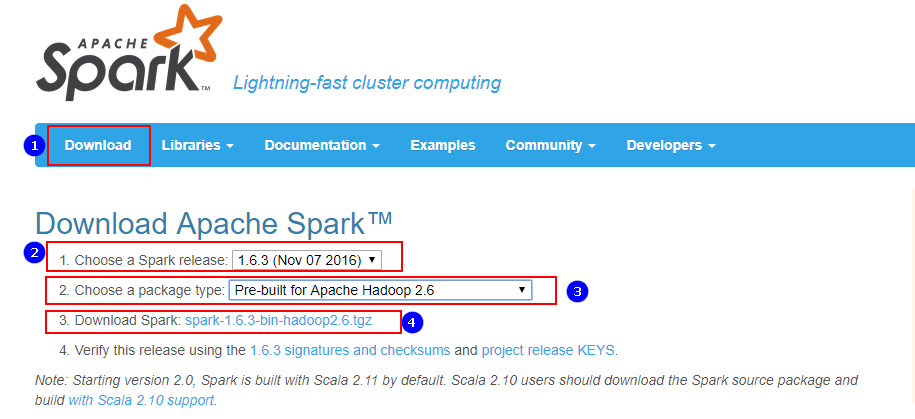
1. 安装PyCharm

下载安装PyCharm。

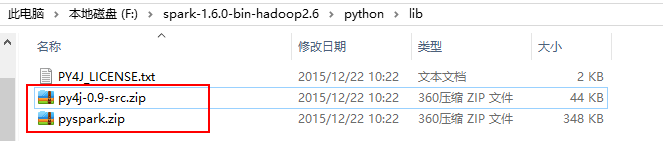
1. 官网下载Spark安装包

由于PyCharm开发spark 需要本地Spark环境，所以要在官网中下载Spark安装包到本地（这里我们下载Spark1.6版本）。

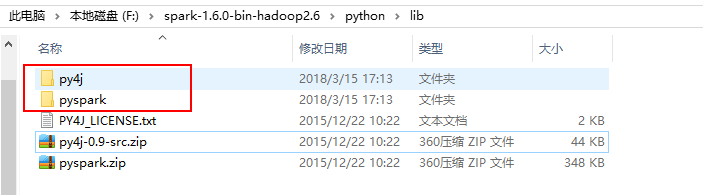
1. 进入Spark官网，找到对应的Spark版本下载。



1. 将下载好的安装包解压到本地某个路径（路径中不要有空格和中文）
2. 将解压的路径中的…spark….\python\lib下的以下两个压缩包解压



解压后：



**注意**：

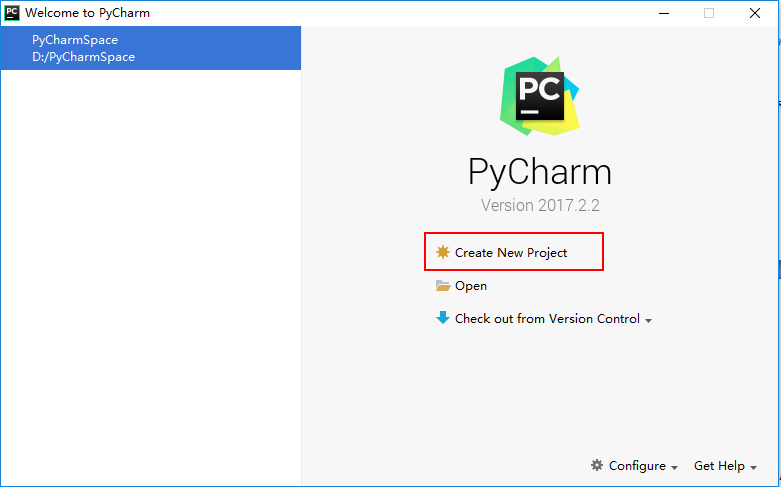
★py4j是一个用 Python和Java编写的库。通过Py4J，Python程序能够动态访问Java虚拟机中的Java对象，Java程序也能够回调Python对象。

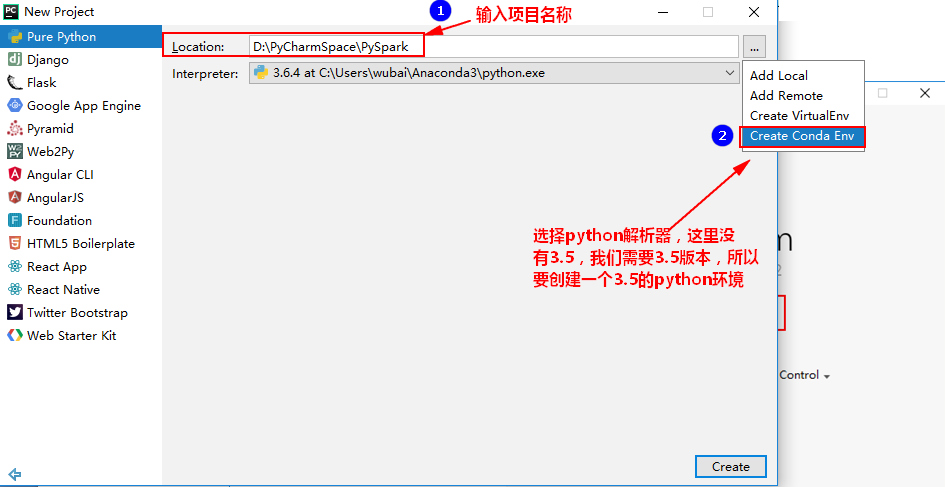
★pyspark 是 Spark 为 Python 开发者提供的 API。

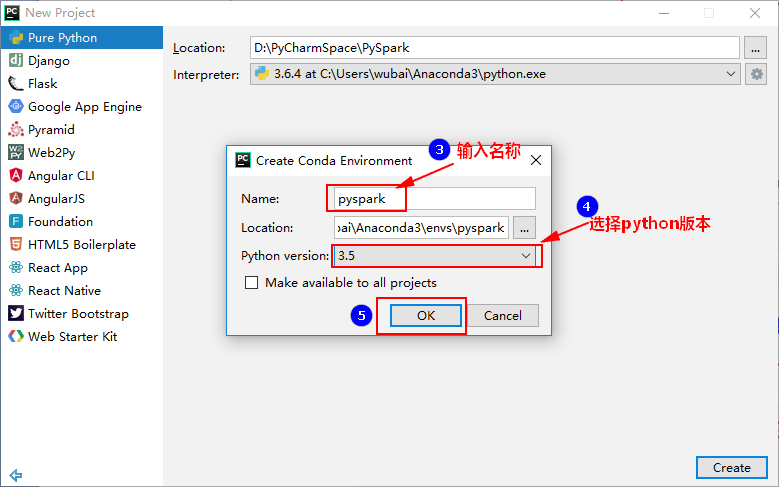
1. **PyCharm开发Spark应用程序**

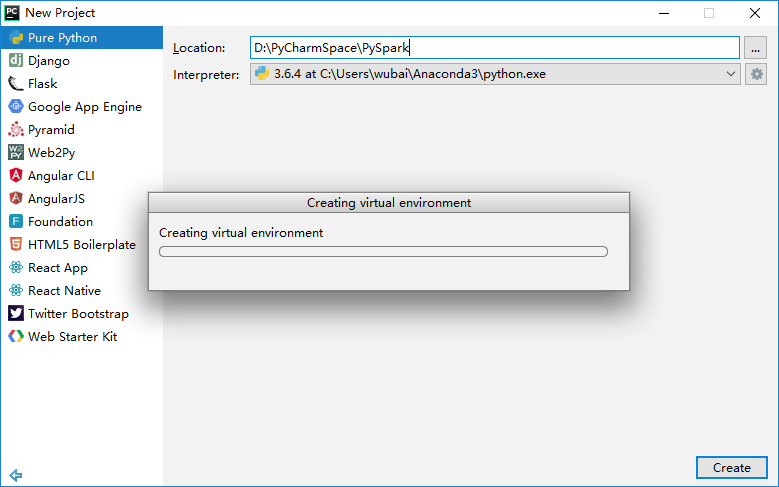
使用PyCharm开发Spark应用环境需要配置Python环境，我们使用的Spark1.6版本，Spark2.1以下版本不支持python3.6版本，所以我们使用兼容性比较好的Python3.5版本，需要在PyCharm中配置。具体如下：

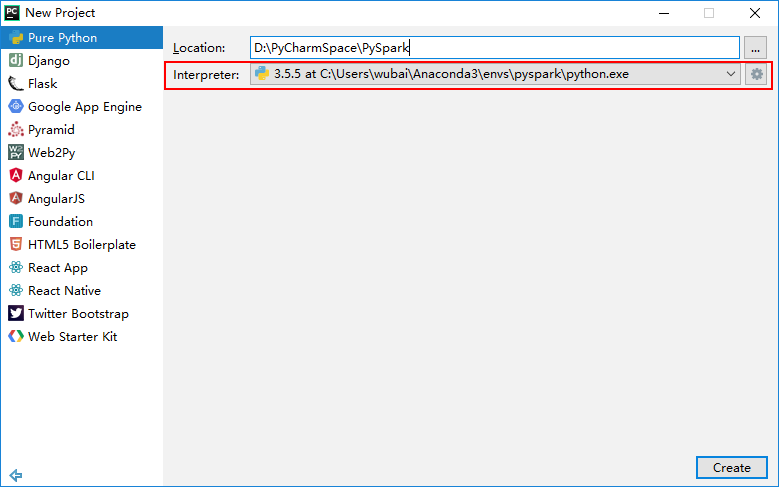
1. 打开PyCharm创建新的项目







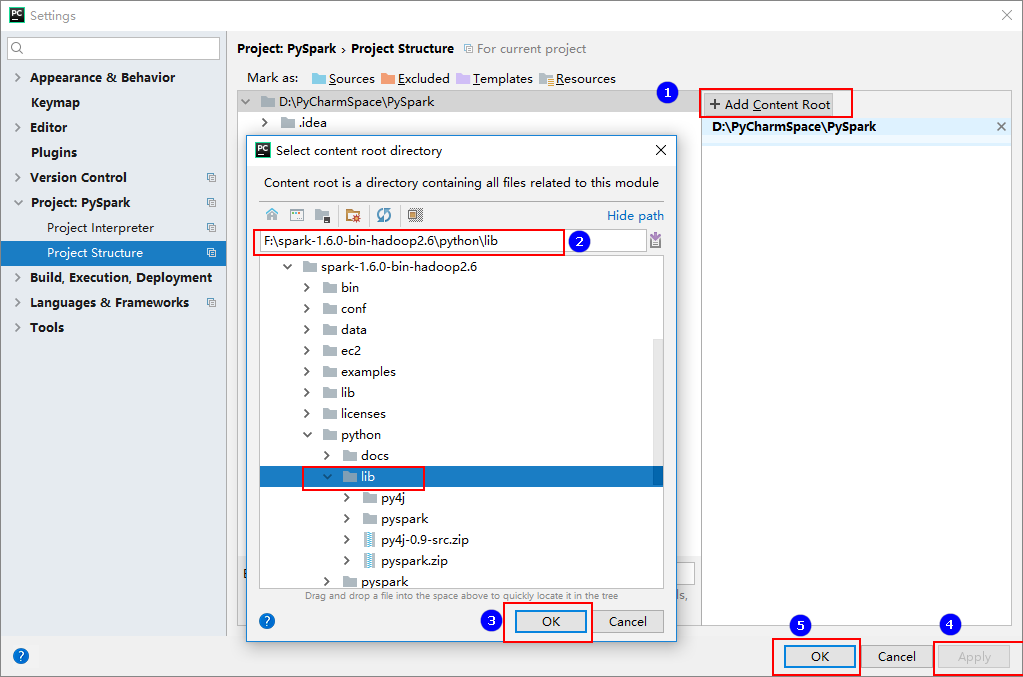




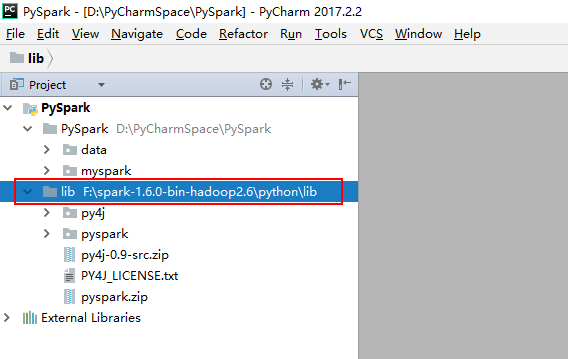
点击Create创建。

1. 配置项目的结构

选中项目,点击File->Settings…->点击 Project:xxx:->Project Structure

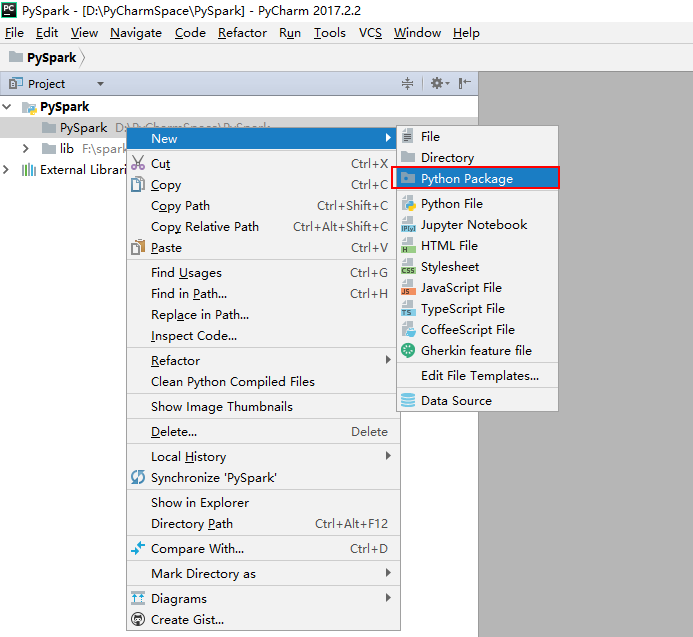


配置完成后，项目根目录下多了一个lib目录：

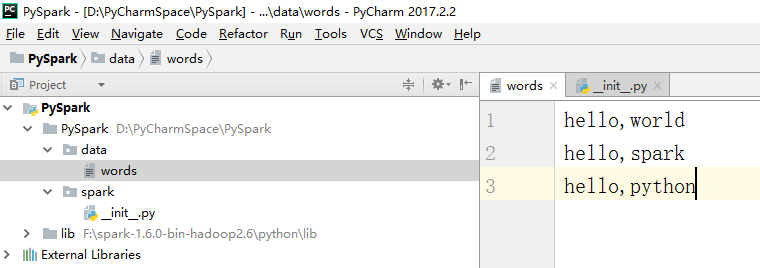


1. 创建一个Spark应用程序脚本（WordCount为例）

创建一个package：



右键创建一个目录,目录中创建一个文件，内容如下：



创建python脚本，内容如下:

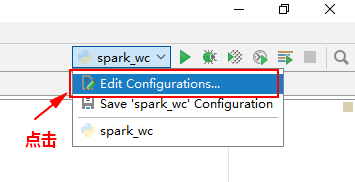
|  |
| --- |
| **from** pyspark **import** SparkConf **from** pyspark **import** SparkContext  **if** \_\_name\_\_ == **'\_\_main\_\_'**:  conf = SparkConf().setMaster(**"local"**).setAppName(**"wordcount"**)  sc = SparkContext(conf=conf)  lines = sc.textFile(**"../data/words"**)  words = lines.flatMap(**lambda** line:line.split(**","**))  pair\_words = words.map(**lambda** word:(word,1))  result = pair\_words.reduceByKey(**lambda** a,b:a+b)  print(result.collect()) |

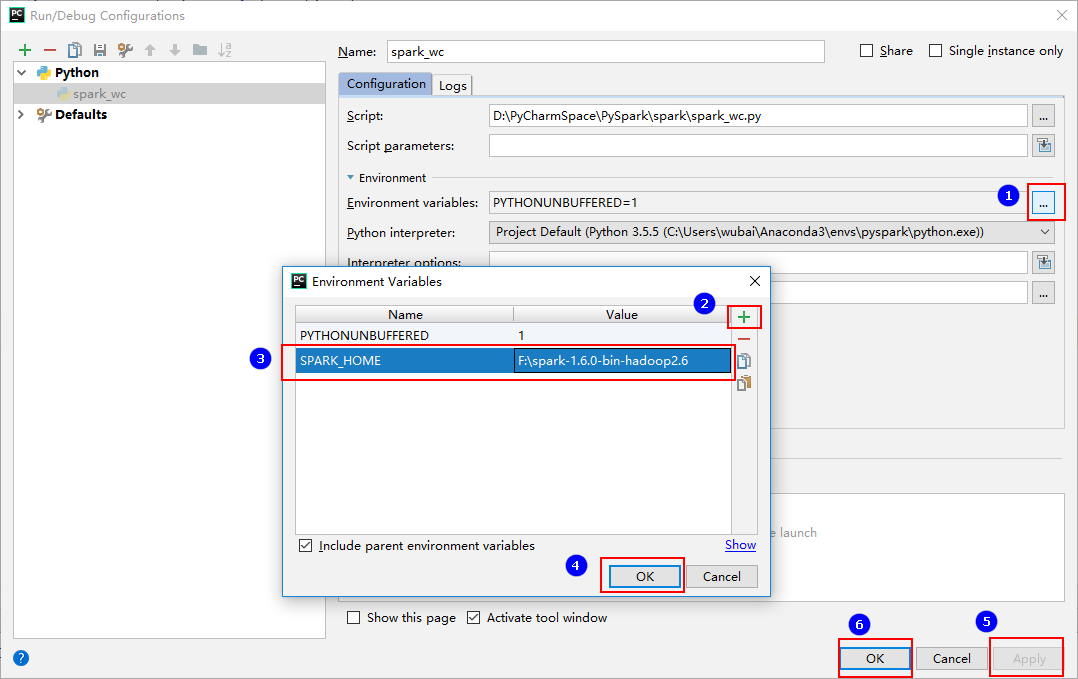
运行，报错如下，需要配置SPAKR\_HOME：



1. 配置SPARK\_HOME

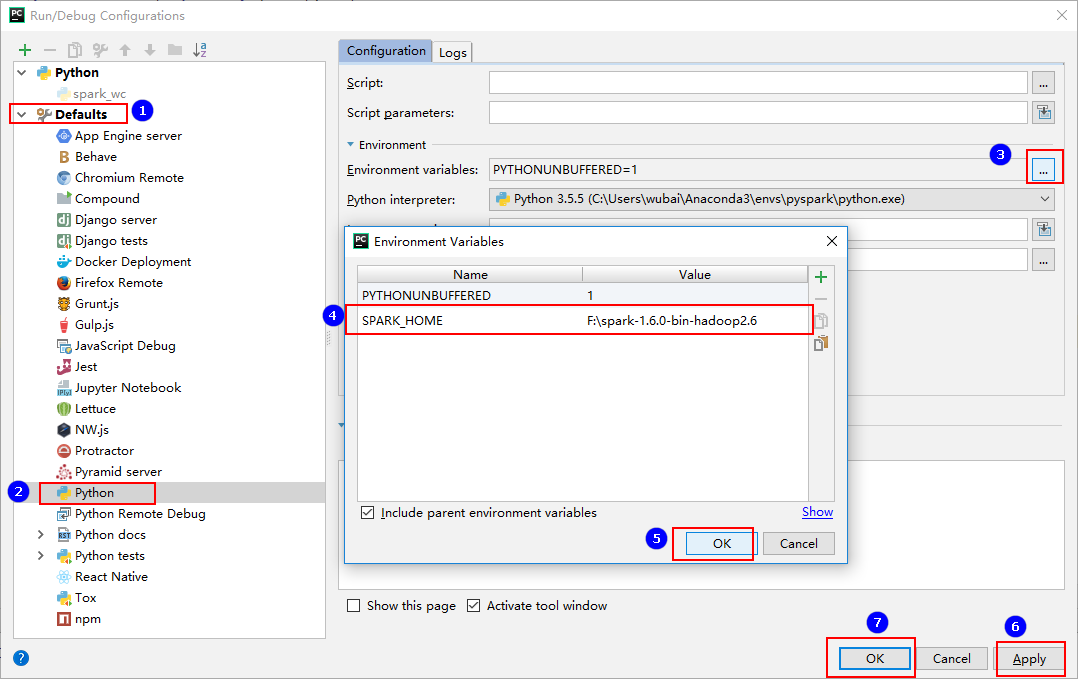
写好pyspark脚本后，在运行python脚本之前，要配置下SPARK\_HOME



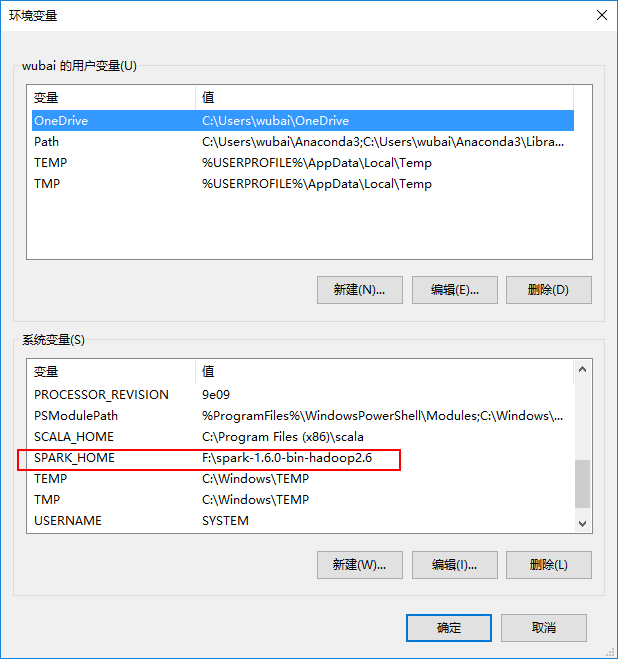


找到本地解压的spark安装包路径，配置SPARK\_HOME，完成即可。

也可以在Defaults中配置好SPARK\_HOME,每次创建pyspark脚本运行时，就不用手动来设置。



注意：这里的SPARK\_HOME也可以在本地系统中的环境变量中配置好，在pycharm中运行pyspark脚本时就不会找不到SPARK\_HOME,两种方式都是可以的。**在系统中配置好环境变量要重启pycharm。**



1. 运行脚本。

注意：运行脚本时有可能出错：

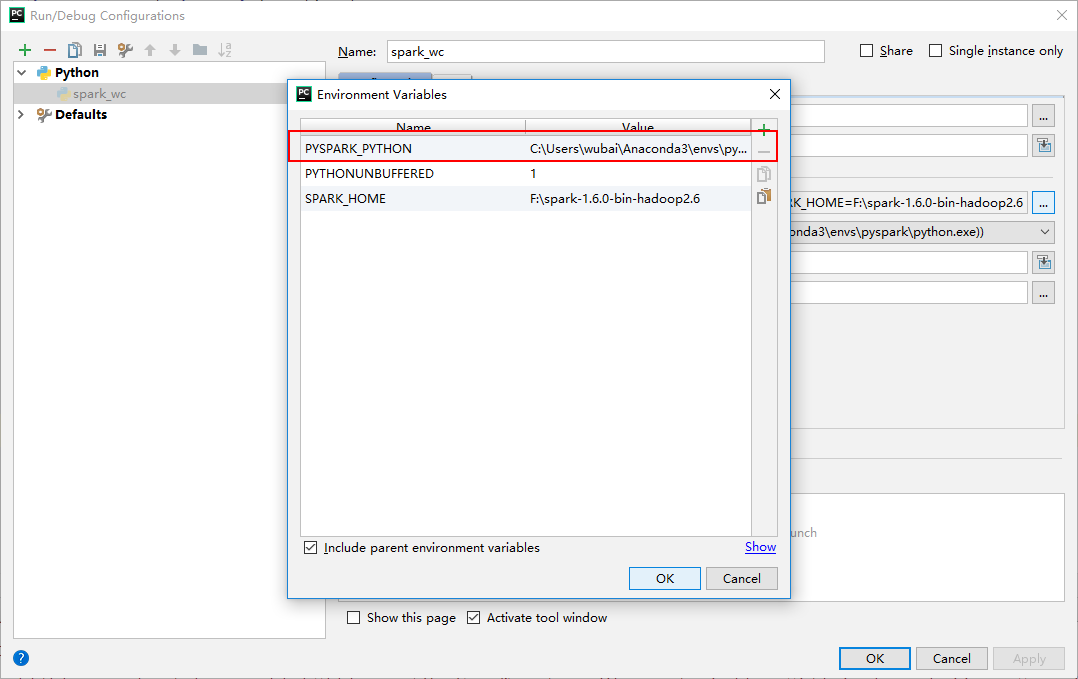


指定pyspark使用的python版本即可（与项目的Project Interpreter使用的python版本路径一致）：

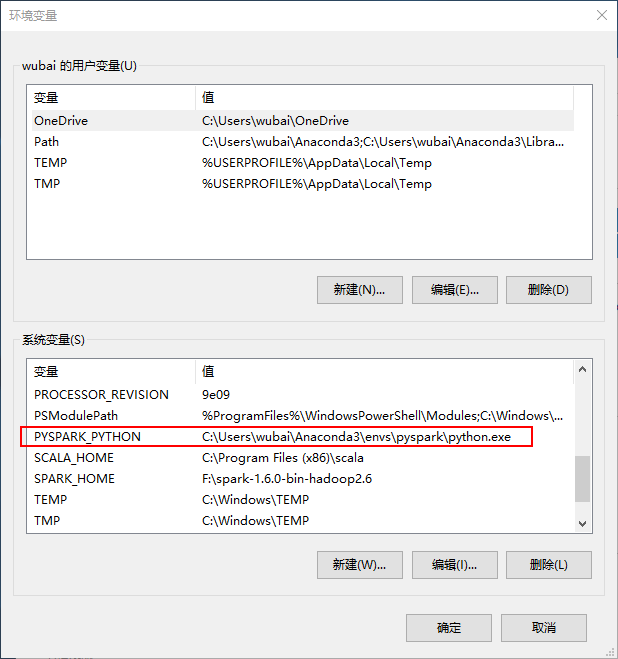
在代码中可以加入：

|  |
| --- |
| **import** os os.environ[**"PYSPARK\_PYTHON"**] = **"C:\\Users\\wubai\\Anaconda3\\envs\\pyspark\\python.exe"** |

也可以在Edit Configurations…中加入环境变量：



也可以在系统中配置环境变量（**要重启Pycharm**）：



以上三种方式任选其一即可。