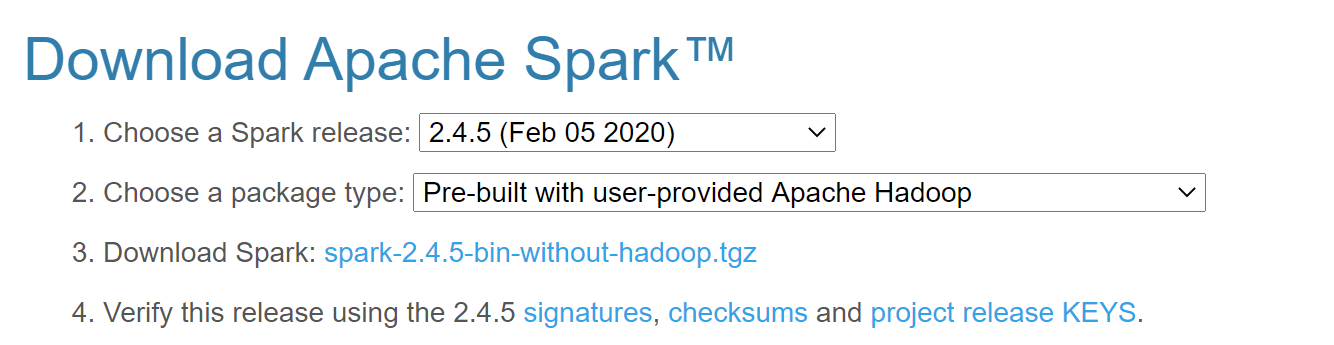
在Windows 10上安装Spark

## 一 先决条件

1. 已经成功装了Hadoop。HADOOP\_HOME环境变量已经正确设置（指向Hadoop的安装目录）；Hadoop的命令工具所在目录也已经被添加到Path环境变量之中（在%HADOOP\_HOME%\bin目录和%HADOOP\_HOME%\sbin目录）
2. 成功安装了Python环境（3.5以上版本）。没有安装的可以到Python的官网<https://www.python.org>自己下载安装，很简单。安装完成后检查一下Python的命令工具所在目录是否被添加到Path环境变量之中。（在命令行方式下运行python –V命令看是否成功。）

## 二 安装Spark

1. 到Spark的官网 <https://spark.apache.org/downloads.html> 下载Spark 2.4.5的安装包。按下图方式进行版本选择。

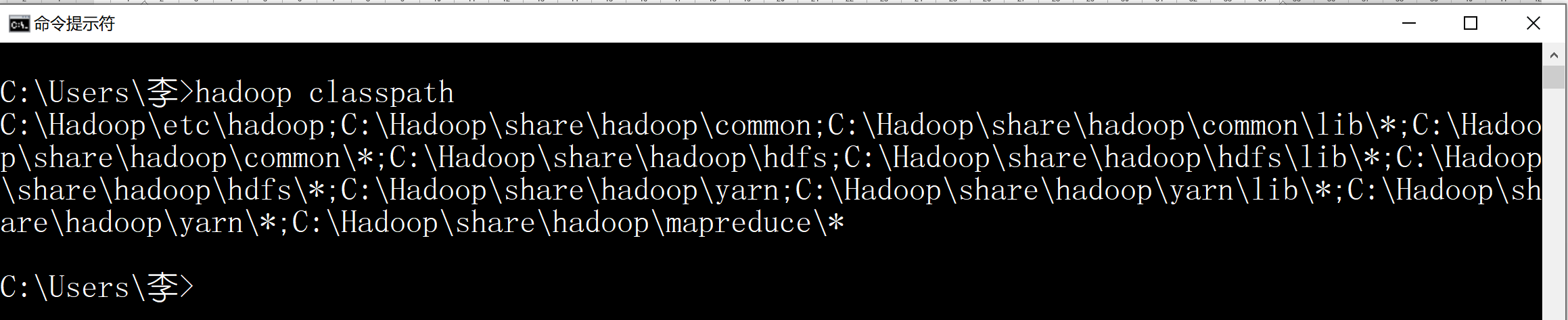


之所以要选“Pre-built with user provided Apache Hadoop”是因为我们之前已经安装好了Hadoop环境。

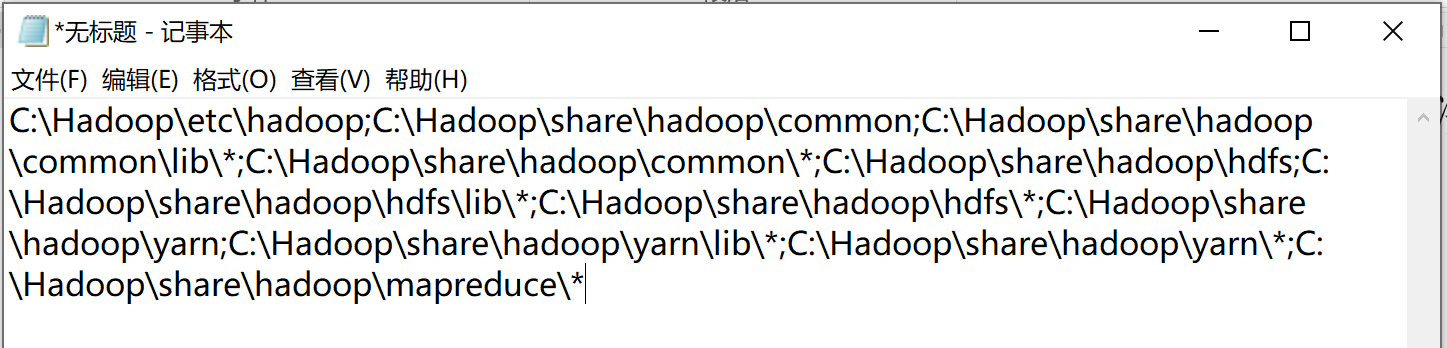
1. 将下载的安装包解压到自己选定的Spark安装目录。在本人的电脑上，我把安装包解压到了C:\Spark目录，所以看起来是这个样子：



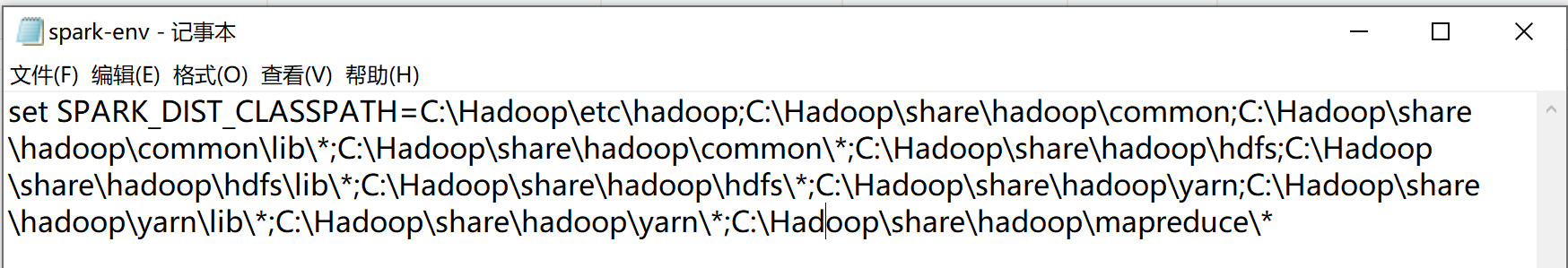
1. 把C:\Spark\bin添加到环境变量Path中。
2. 获得本机安装的Hadoop的类库引用路径，方法是：在命令行方式下运行“hadoop classpath”命令，然后把运行结果拷贝到记事本中，该结果将被用在第5步中。在本人电脑上的运行结果如下：



因此，我在记事本中拷贝的内容是：



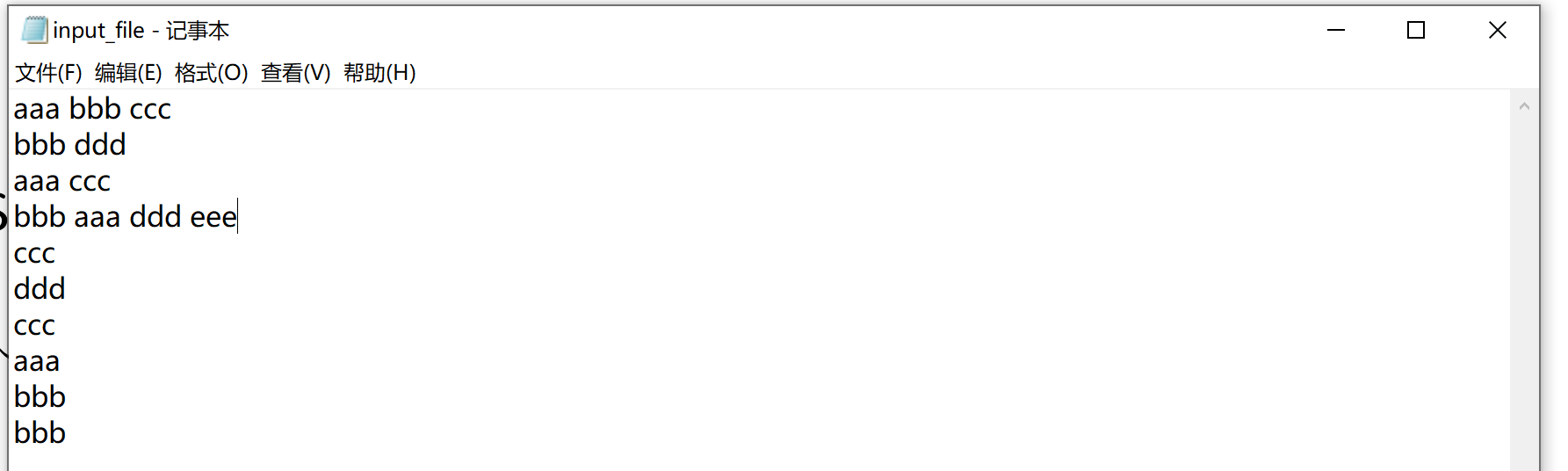
1. 在C:\Spark\conf目录下新建一个文件，命名为“spark-env.cmd”，内容如下：



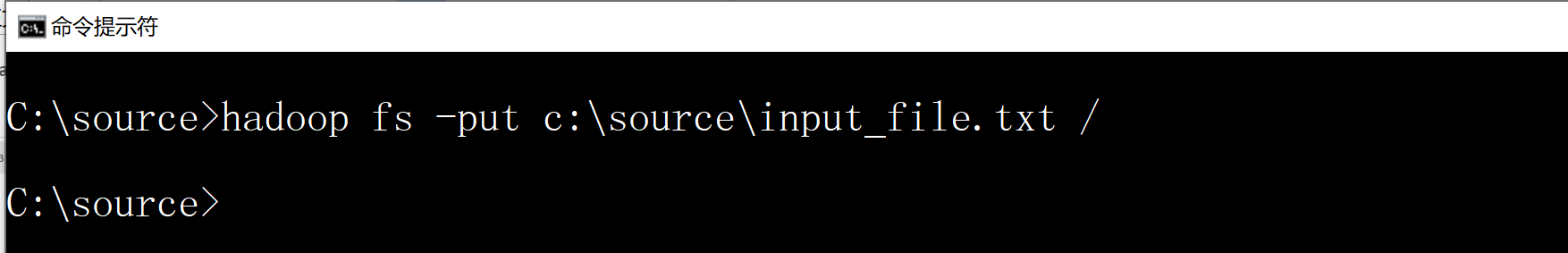
即在该文件中设置SPARK\_DIST\_CLASSPATH变量为第4步中获得的结果。

## 三．运行Spark

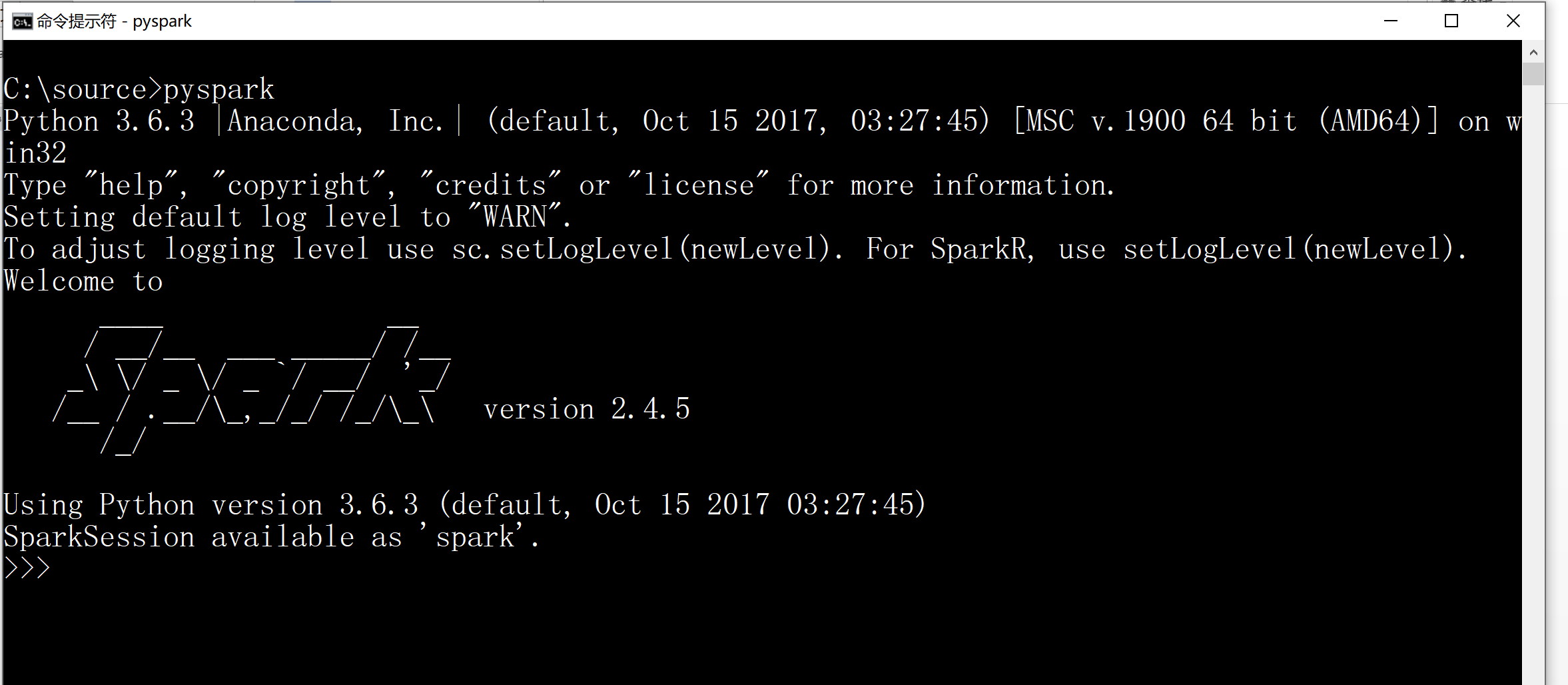
1. 准备好一个用于测试的文本文件input\_file.txt，内容如下：



1. 在命令行方式下运行“start-dfs”命令启动HDFS。HDFS启动成功之后用“hadoop fs –put ……”命令把input\_file.txt上传到HDFS文件系统的根目录。



1. 在命令行方式下运行“pyspark”命令启动spark的基于python交互式客户端，如果spark安装成功，可以看到如下结果：



1. 在pyspark中输入下面4个python语句，测试是否能正常执行：

file="/input\_file.txt"

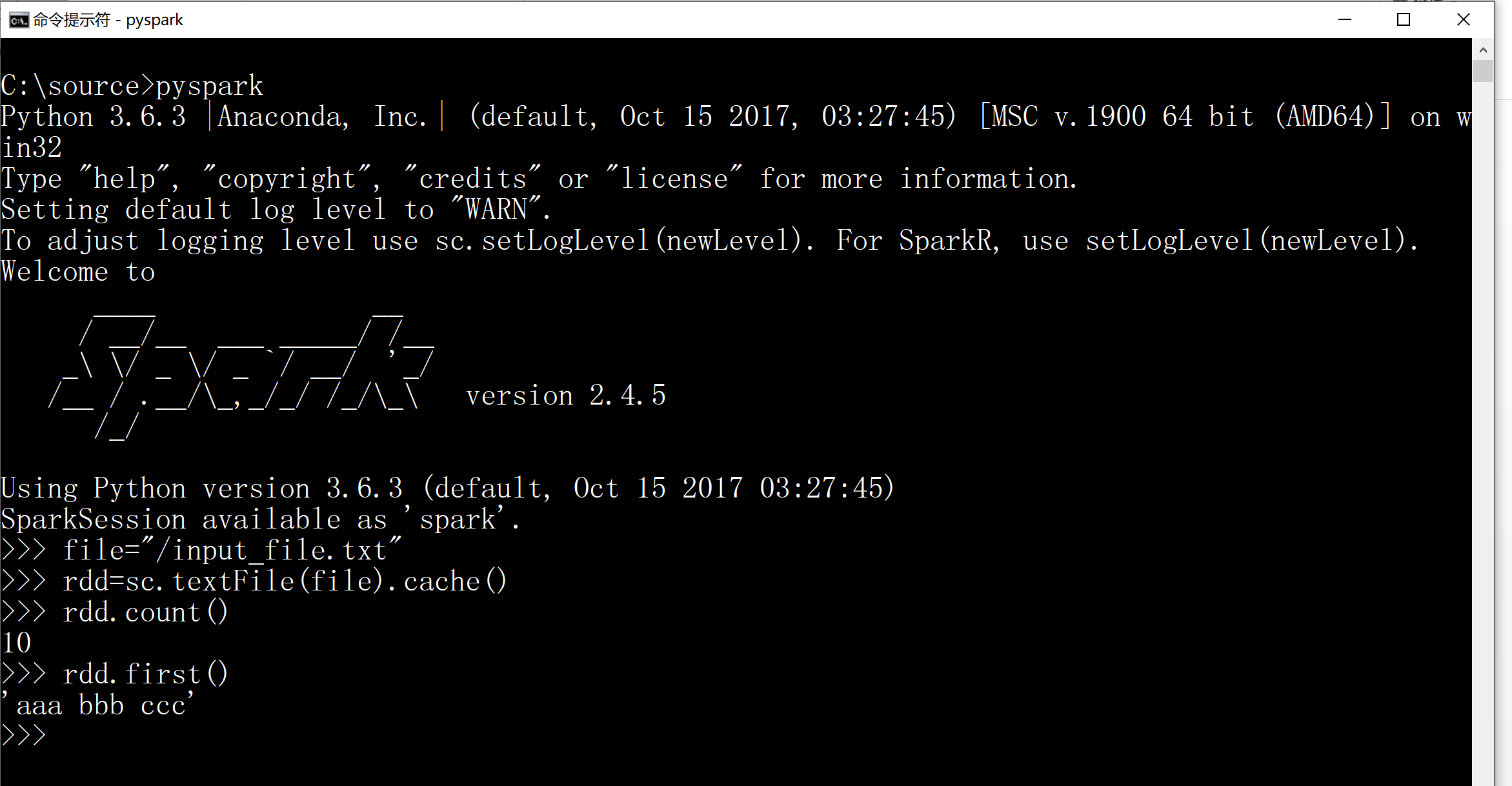
rdd=sc.textFile(file).cache()

# 获得rdd中的元素个数：

rdd.count()

# 获得rdd的第一个元素

rdd.first()



1. 用“quit()”命令从pyspark退出。
2. 将下面的python程序保存在wordcount.py文件：

|  |
| --- |
| from pyspark import SparkConf,SparkContext  conf = SparkConf().setMaster("local").setAppName("wordcount")  sc = SparkContext(conf=conf)  textData = sc.textFile("/input\_file.txt")  splitData = textData.flatMap(lambda line:line.split(" "))  flagData = splitData.map(lambda word:(word,1))  countData = flagData.reduceByKey(lambda x,y:x+y)  countData.saveAsTextFile("/result") |

1. 在命令行下运行“**spark-submit wordcount.py**”命令，运行“单词计数”任务。如果运行成功，结果将被写入到了HDFS的“/result”子目录下。可以用如下方式查看统计结果：

