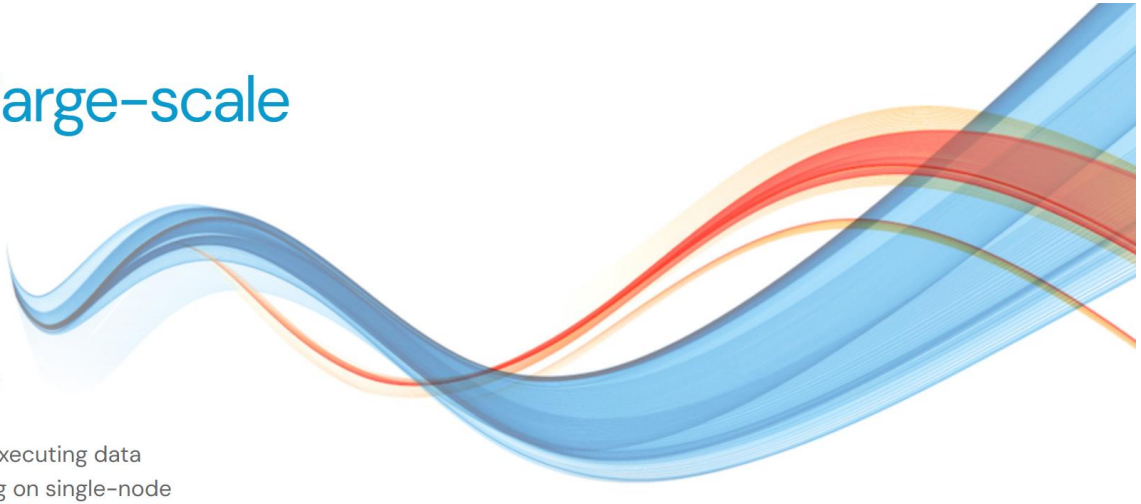
在Windows 下安装大数据平台Spark并使用python进行编程

Unified engine for large-scale data analytics

GET STARTED



What is Apache Spark™?

Apache Spark™ is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters.

Simple. Fast. Scalable. Unified.

Key features

Batch/streaming data

Unify the processing of your data in batches and real-time streaming, using your preferred language: Python, SQL, Scala, Java or R.

SQL analytics

Execute fast, distributed ANSI SQL queries for dashboarding and ad-hoc reporting. Runs faster than most data warehouses.

Data science at scale

Perform Exploratory Data Analysis (EDA) on petabyte-scale data without having to resort to downsampling

Machine learning

Train machine learning algorithms on a laptop and use the same code to scale to fault-tolerant clusters of thousands of machines.

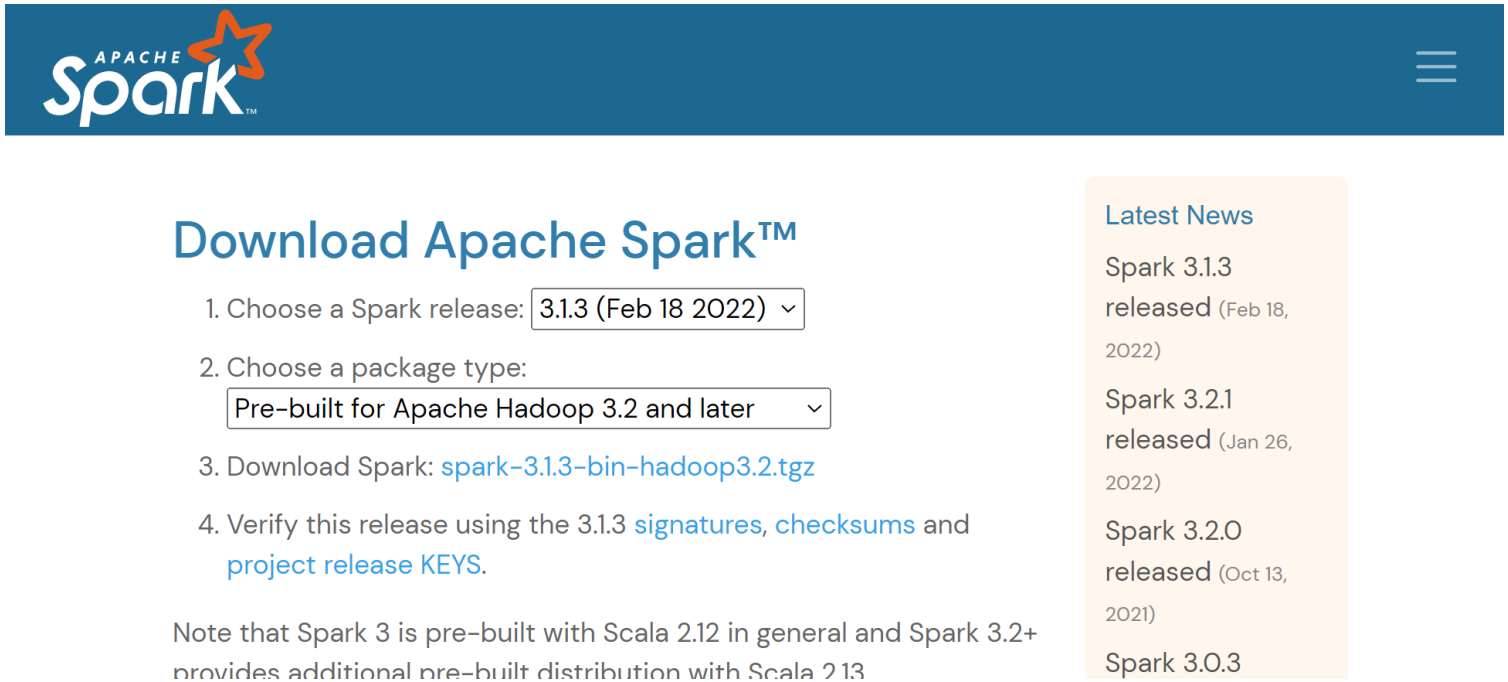
概要

主要分为如下5个步骤：

1. 下载并解压缩Spark （所用版本：spark-3.1.3-bin-hadoop3.2）
2. 安装java （java 8 或者java 11 都可以）
3. 下载winutils.exe
4. 设置环境变量
5. 调试
6. 在python环境中使用PySpark

1. 下载spark

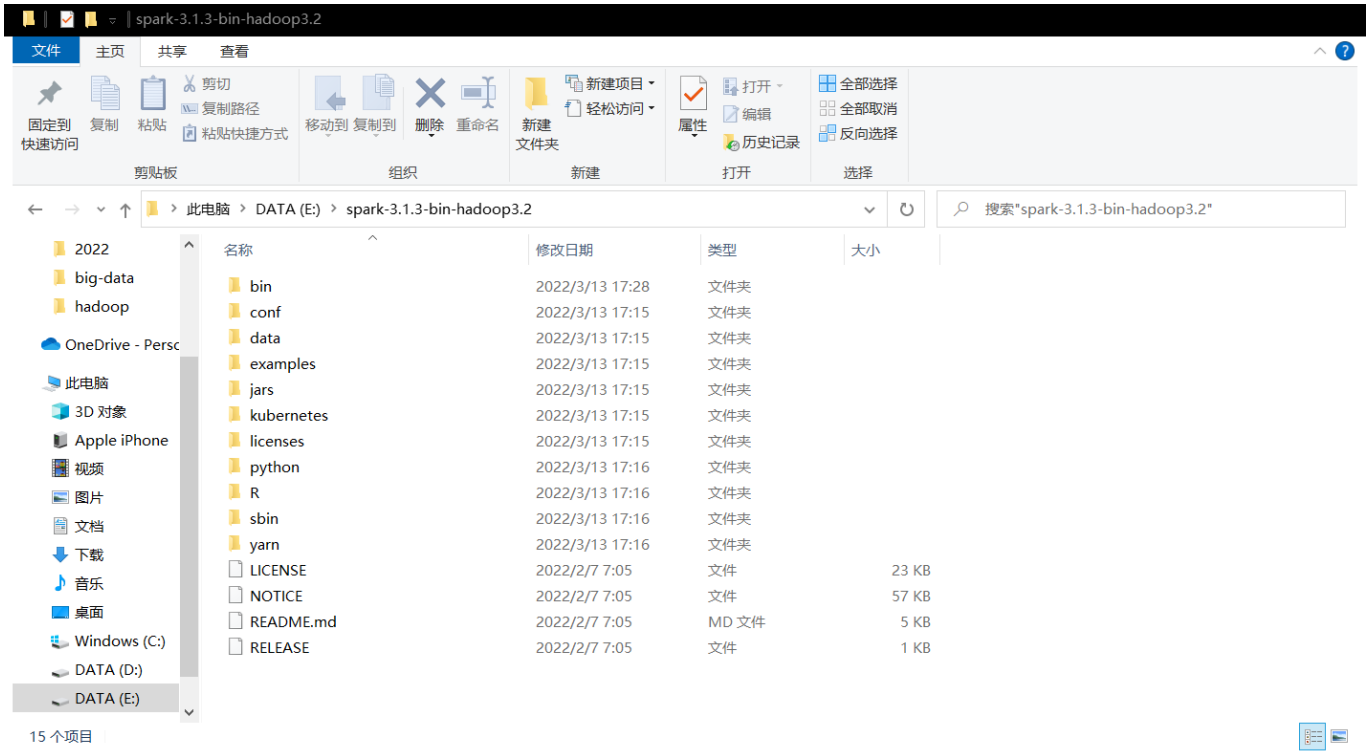
Hadoop 的文件系统（HDFS）是一种广泛使用的分布式存储系统，常被用作 Spark 数据处理的底层存储层。HDFS 提供了高吞吐量的数据访问，非常适合于大规模数据集的存储和处理。



下载地址：https://spark.apache.org/downloads.html

1. 下载spark

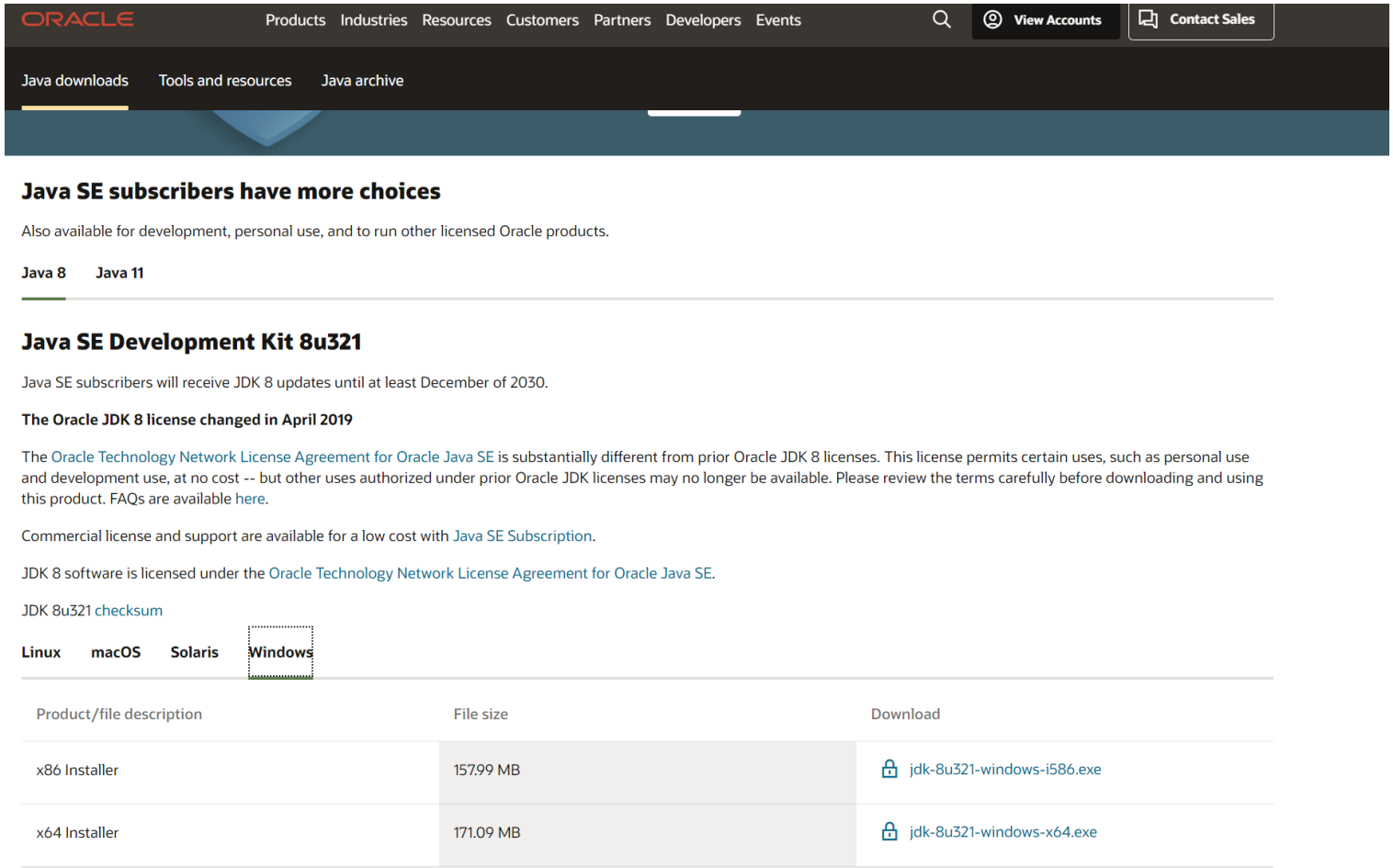
下载后得到spark-3.1.3-bin-hadoop3.2.tgz，将其解压缩，得到如下的文件夹



2. 下载安装java

安装java 8 或者 java 11 都可以（建议java 11）

https://www.oracle.com/java/technologies/downloads/#java8-windows



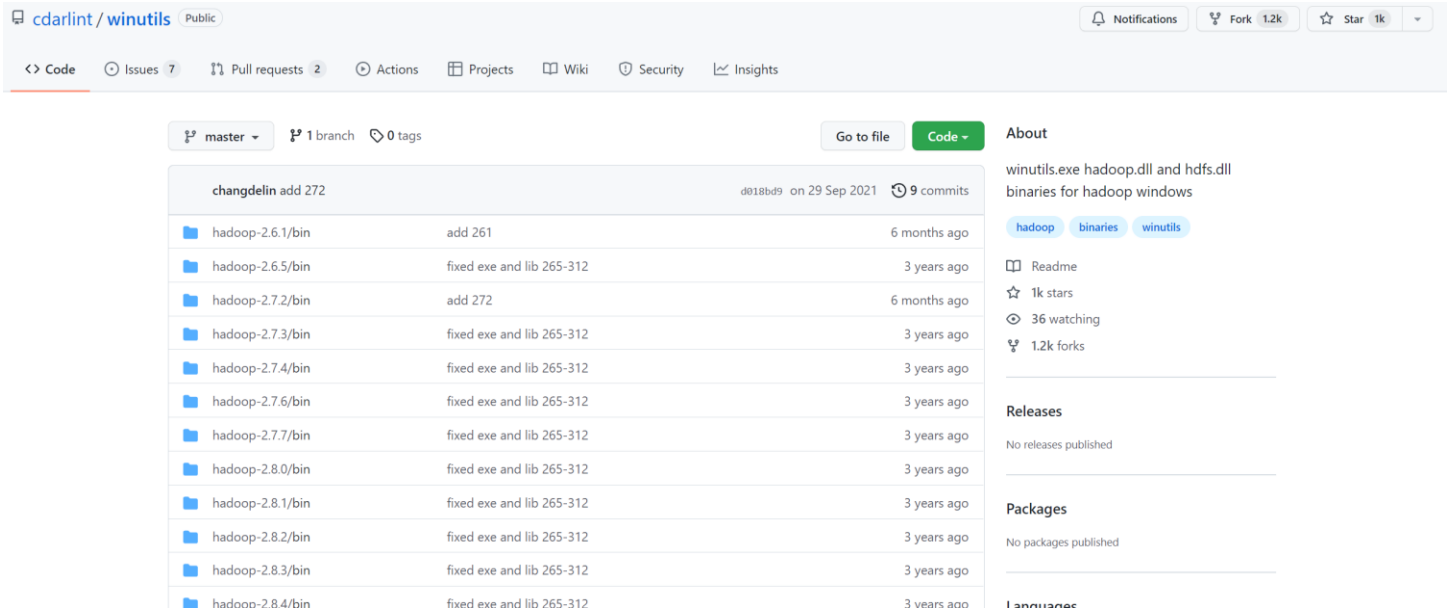
2. 下载安装java

安装过程中会出现如下界面，单击“更改”，以将java安装在一个不带空格的路径(默认路径里 Program Files的中间带有空格)，比如安装到D:\Java\jdk-11.0.14

JVM（Java虚拟机）依赖：Spark 是用 Scala 编写的，Scala 是一种运行在 Java 虚拟机（JVM）上的编程语言。因此，Spark 需要 JVM 来执行。这就是为什么安装 Spark 之前需要安装 Java。

3.下载winutils.exe

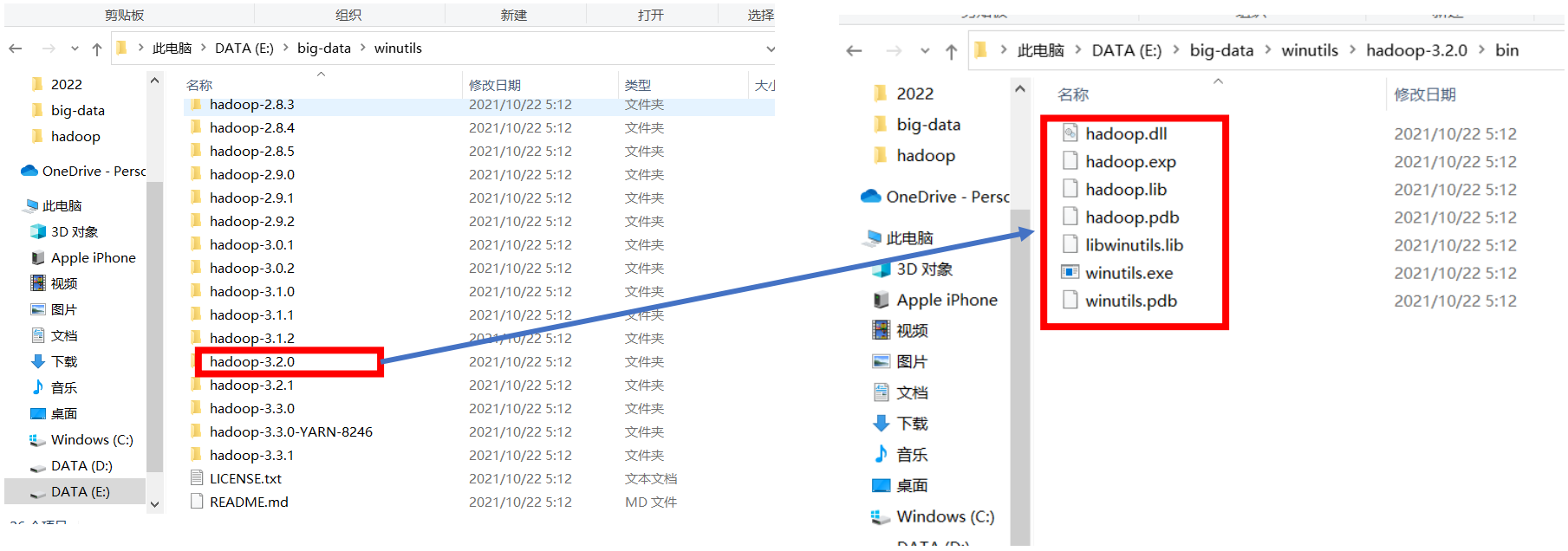
- Windows binaries for Hadoop versions
- 下载地址： https://github.com/cdarlint/winutils



注意这里有很多版本，使用的winutils.exe应与第一步hadoop的版本相对应

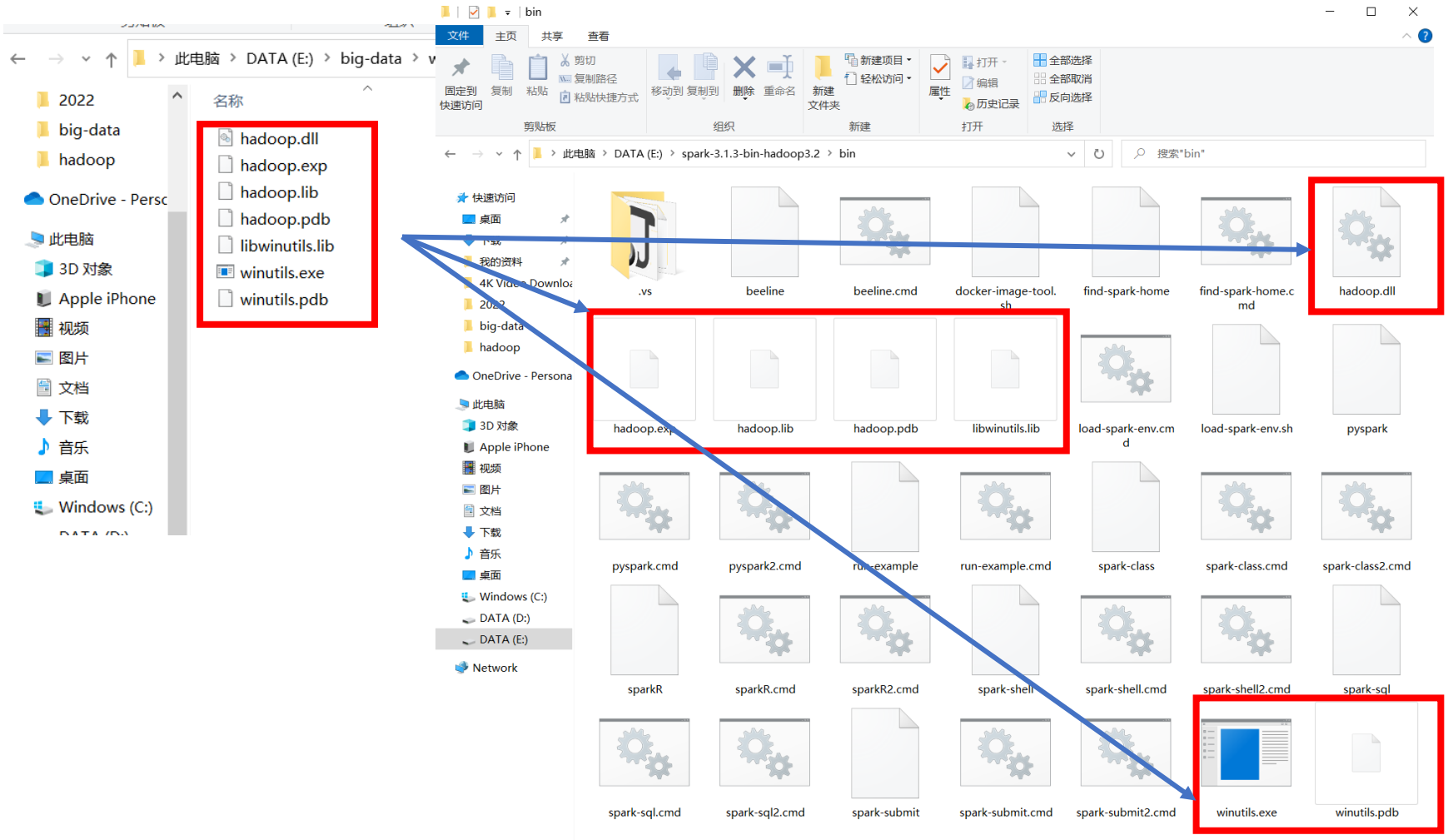
3.下载winutils.exe

- 因为我们在第一步下载的是Hadoop 3.2，所以我们提取hadoop-3.2.0/bin下的文件，将其复制到spark-3.1.3-bin-hadoop3.2/bin的路径下



winutils 是一个在Hadoop环境中使用的工具，尤其是在Windows操作系统上。Hadoop主要是为Linux和Unix系统设计的，但是winutils提供了一种方式在Windows上模拟类似环境。

3.下载winutils.exe

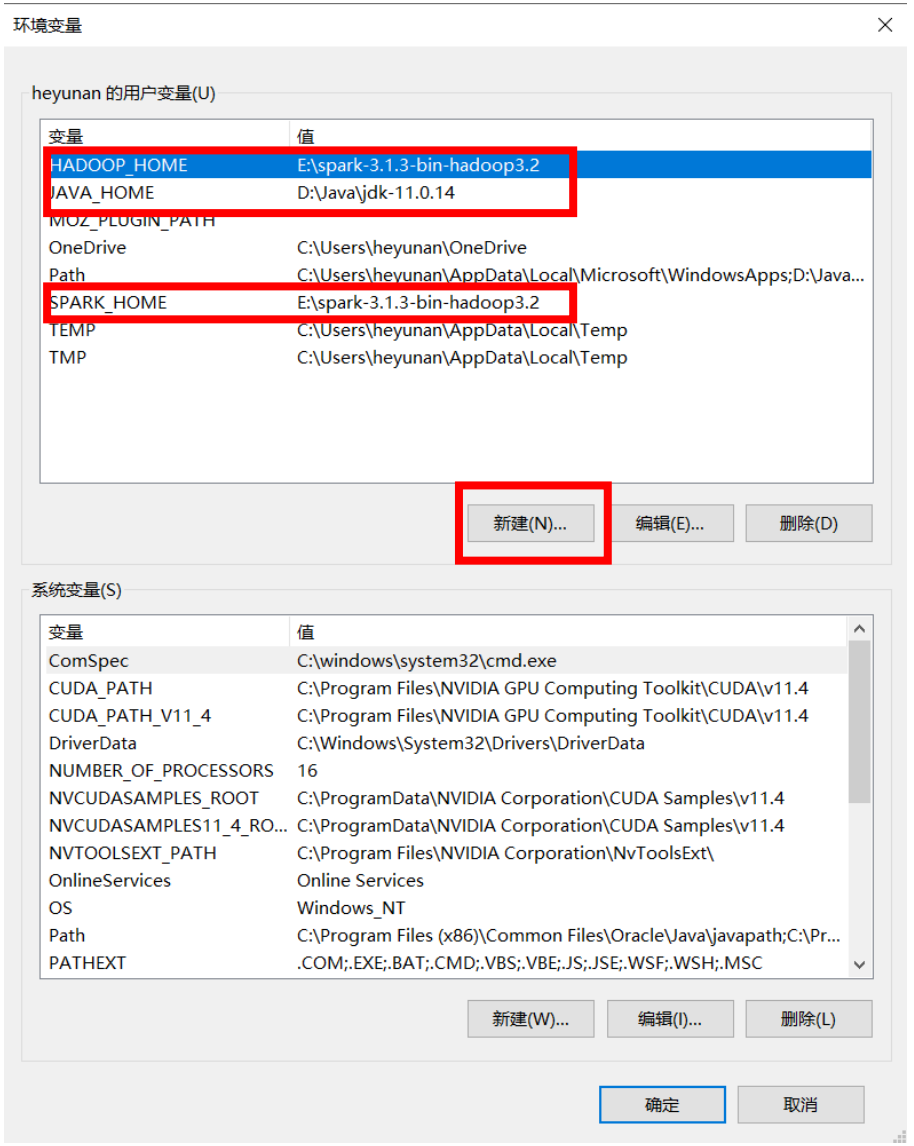


4.编辑环境变量

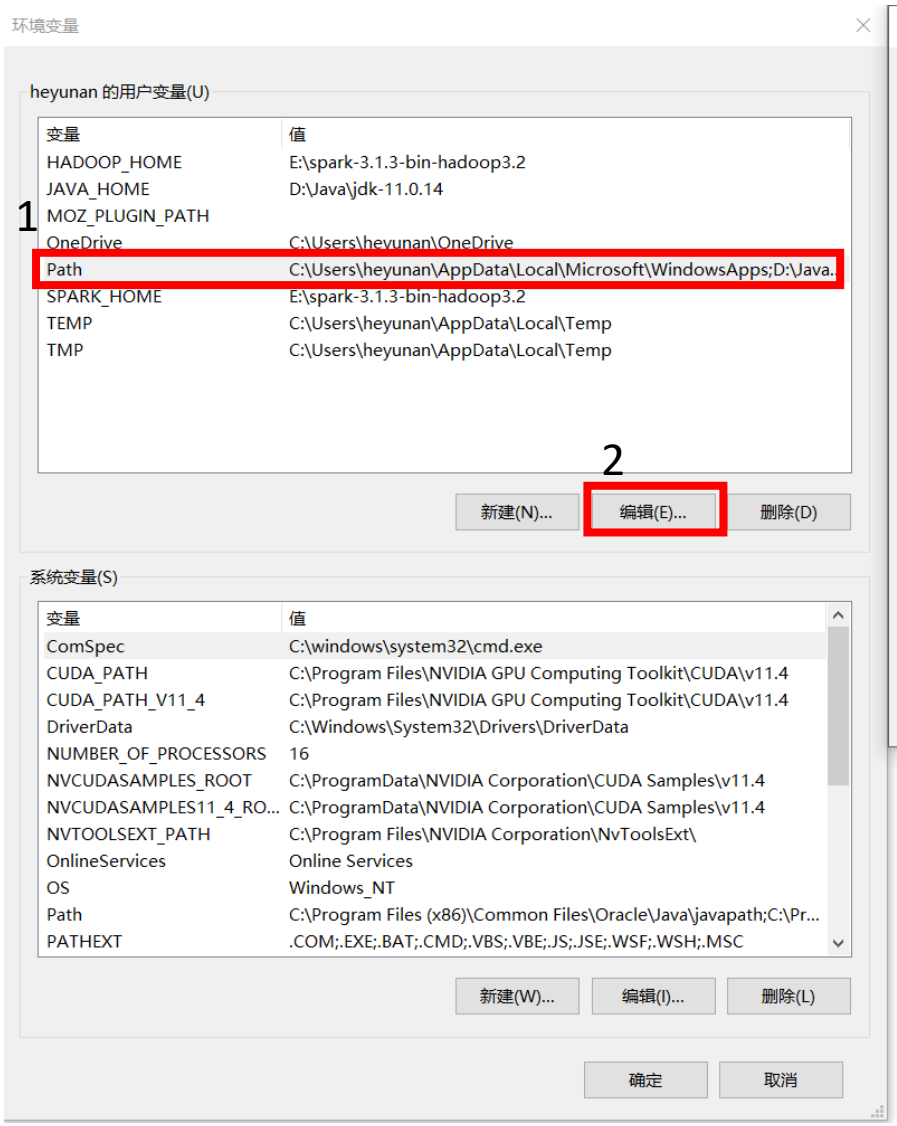
环境变量在计算机的“系统->高级系统设置”当中找到，也可以在搜索框输入“environment”的关键次，就可以找到如下界面。

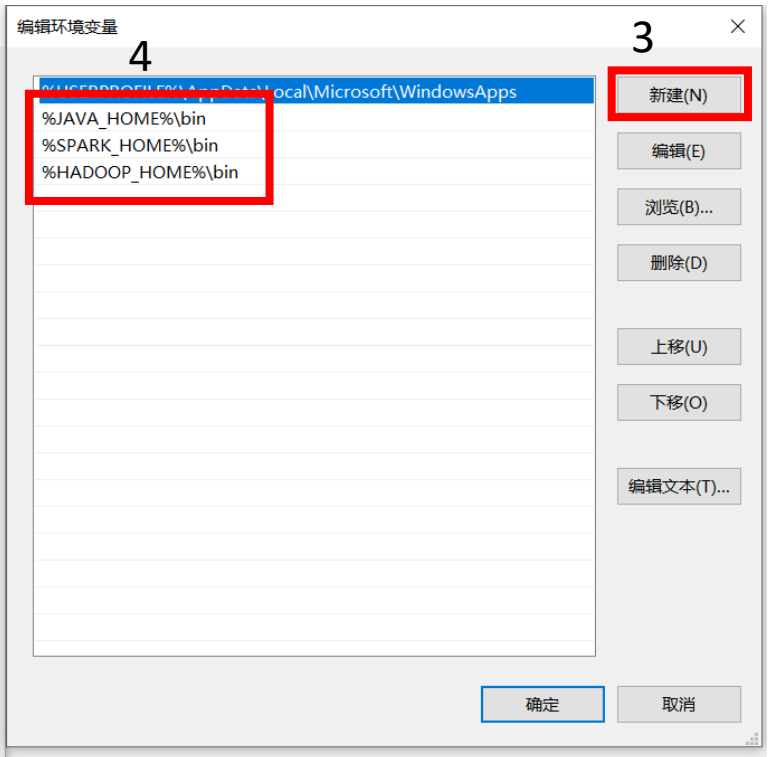
4.编辑环境变量

通过点击“新建”按钮，分别新建三个变量，分别为 HADOOP_HOME, JAVA_HOME, SPARK_HOME，并分别给出HADOOP，JAVA，SPARK的路径。（注意HADOOP 和spark的路径是在第一步中你存放spark-3.1.3-bin-hadoop3.2的位置，JAVA的路径是在第二步中，你指定的JAVA安装位置）



4.编辑环境变量



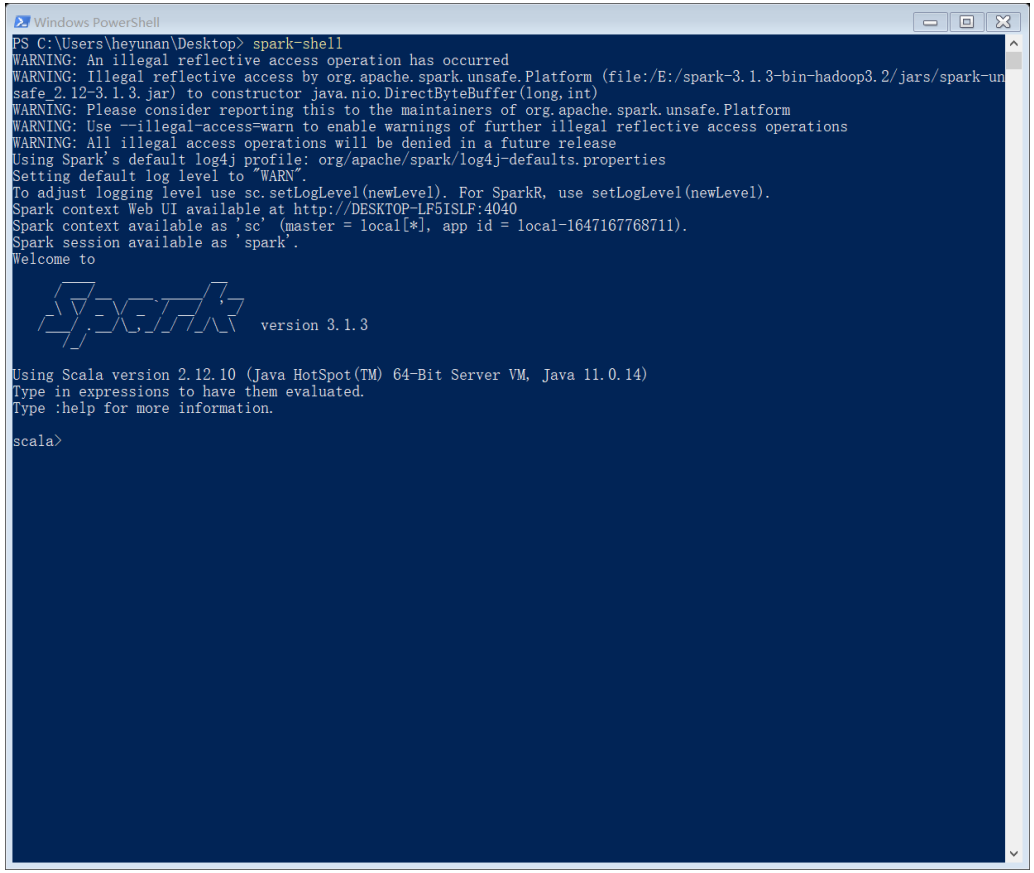


编辑”Path”，并设置如下路径到”Path”

%JAVA_HOME%\bin
%SPARK_HOME%\bin
%HADOOP_HOME%\bin

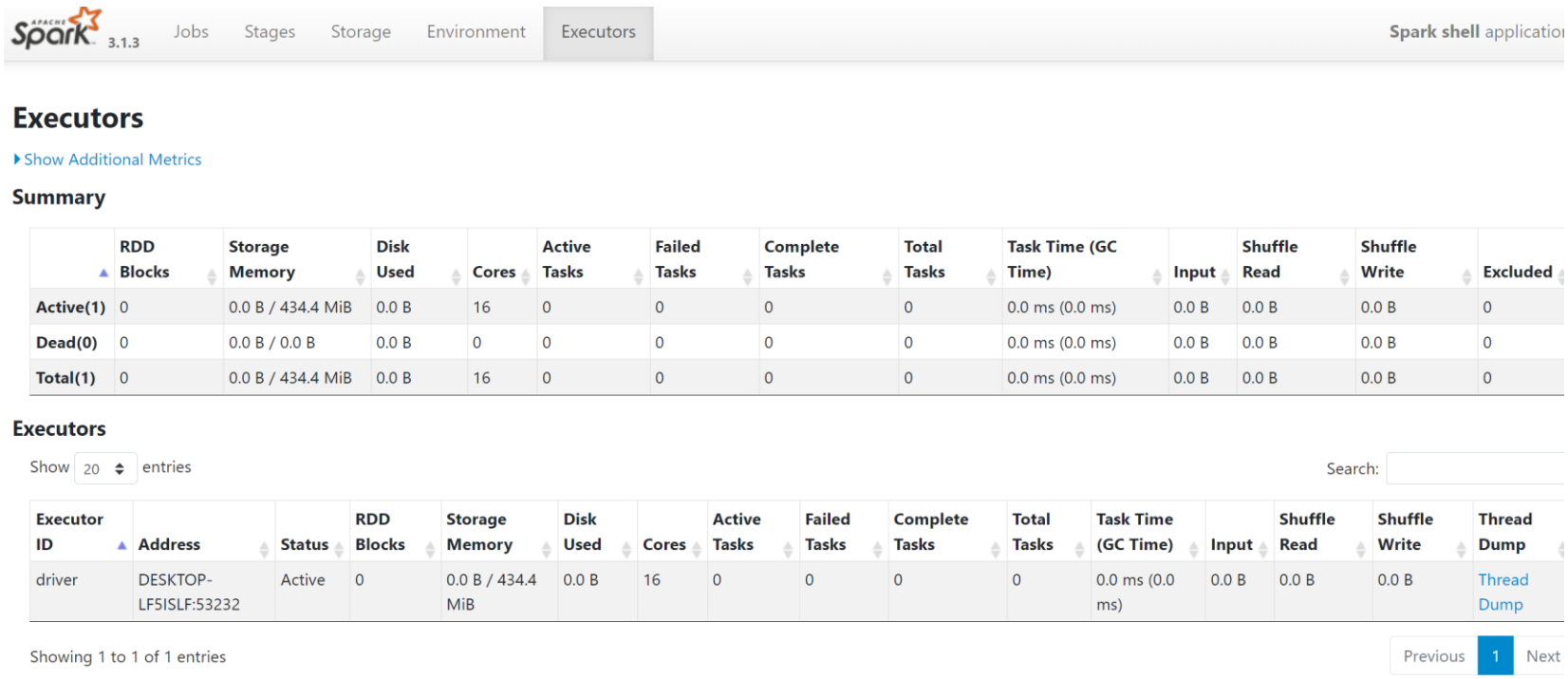
5.调试

打开PowerShell（可以在任意文件夹的空白处按住shift键的同时，单击鼠标右键，会有“在此处打开PowerShell窗口”的选项），并在窗口中输入spark-shell，回车后有如下界面：



5.调试

打开 http://localhost:4040/，会显示如下界面

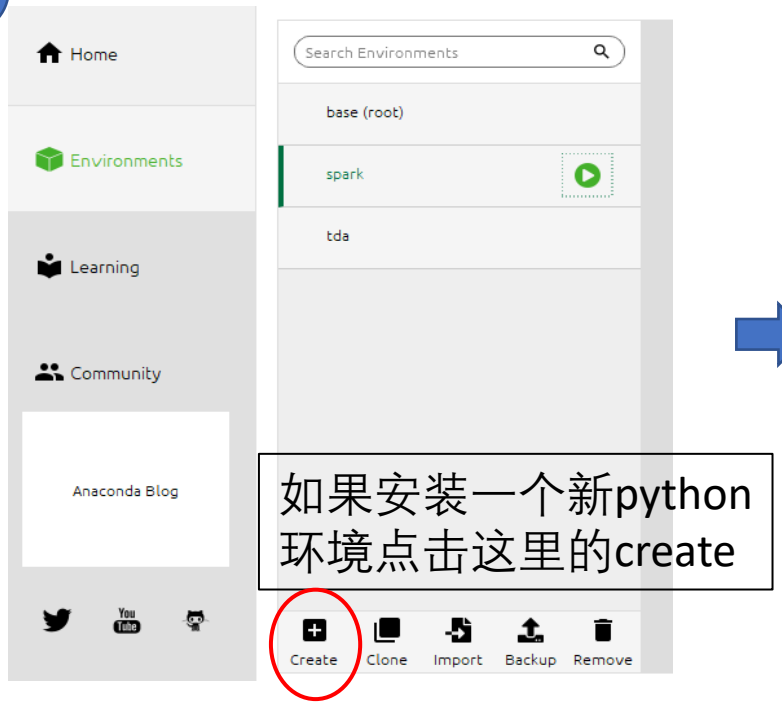


6. 在python环境中使用PySpark

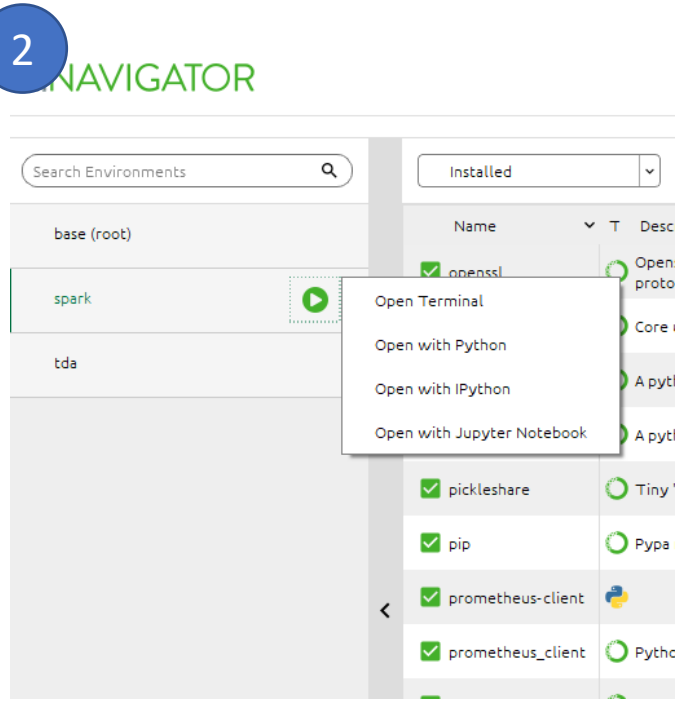
进入Anaconda的navigator中，选择现有python环境的terminal或者新建一个python环境（Python版本要大于或者等于3.7）并打开其terminal（截图中我新建一个python环境并命名为spark）。

注意：一定要保证当前python环境中没有通过pip或者conda安装过pySpark

1

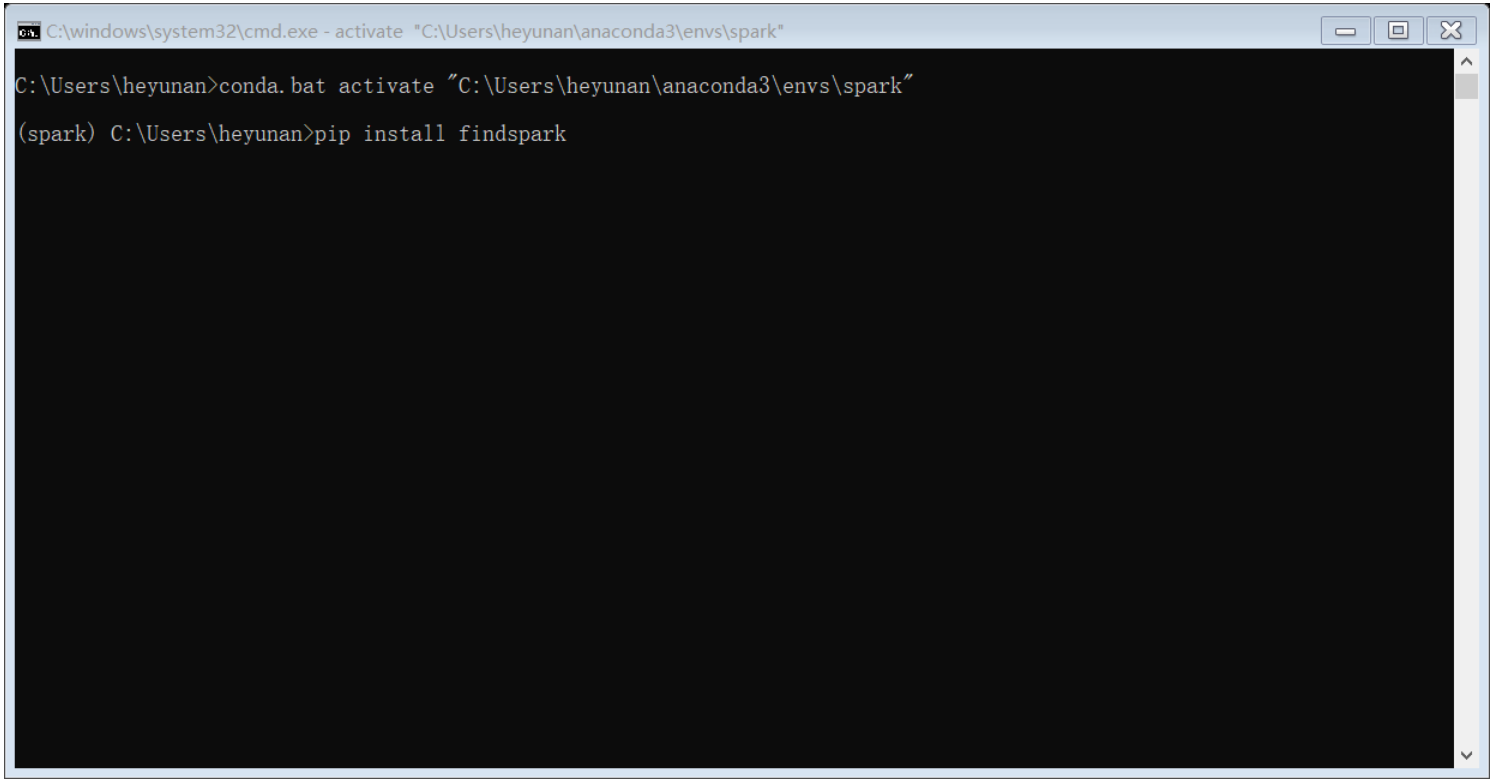


2



6. 在python环境中使用PySpark

- 打开terminal后，安装findspark包.输入： pip install findspark
- 安装findsaprk的目的是找到系统中的spark的位置，并将其pyspark模块导入到python环境中。



6. 在python环境中使用PySpark

在当前Python环境中安装Spyder，jupyter notebook等开发环境.
使用这些开发环境时，首先导入findspark包（见下图），找到系统中pyspark模块后，就可以使用pySpark了！

```
In [1]: import findspark
        findspark.init()
        findspark.find()

Out[1]: 'E:\\spark-3.1.3-bin-hadoop3.2'
```

通过findspark找到系统中spark的路径，每次编程前都需要用findspark导入pyspark。

Hello, Spark

```
In [24]: from pyspark.sql import SparkSession

         spark = SparkSession.builder.getOrCreate()
         df = spark.sql("select 'spark' as hello ")

         df.show()
```

```
+-----+
|hello|
+-----+
|spark|
+-----+
```

Hello，spark示例程序