

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.

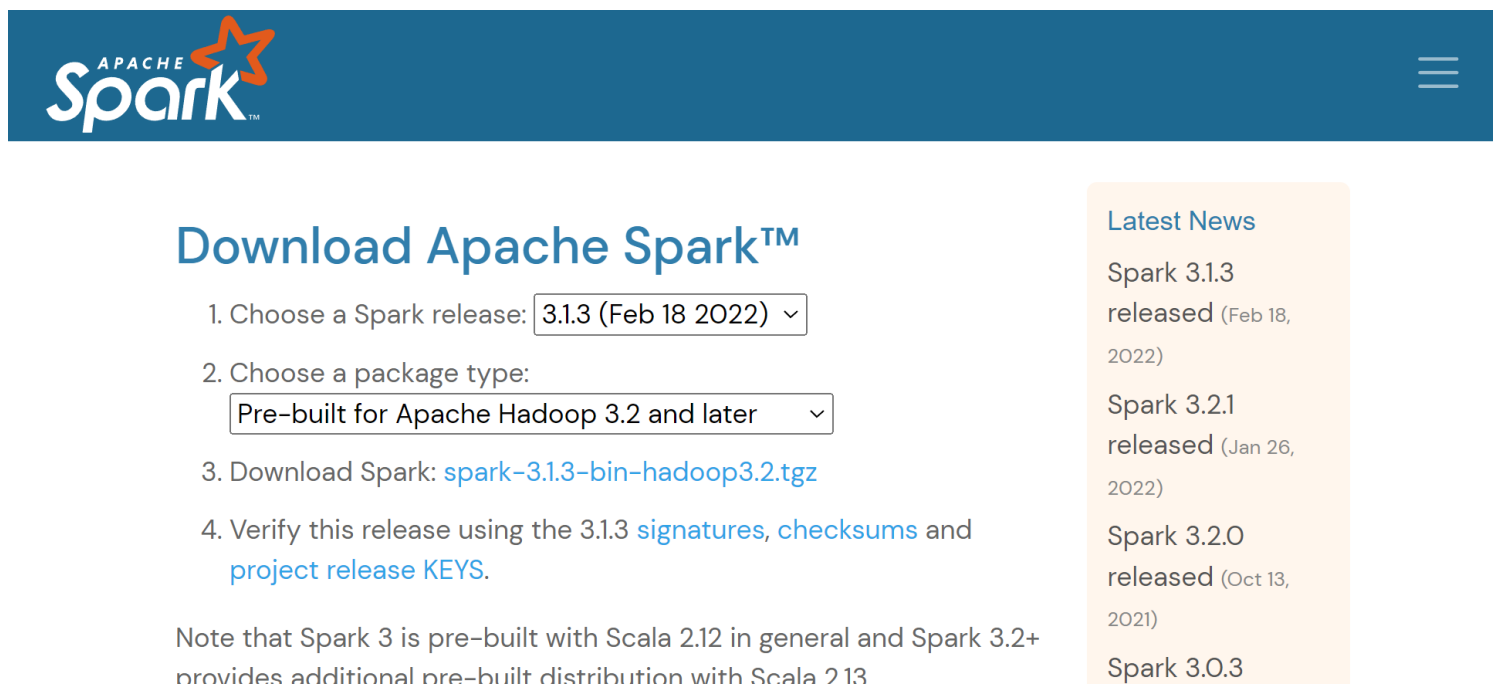
概要

主要分为如下5个步骤：

1. 下载并解压缩Spark （所用版本： spark-3.5.6-bin-hadoop3.3）
2. 安装java （java 21）
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。



The screenshot shows the Apache Spark download page. At the top is the Apache Spark logo. Below it, the heading "Download Apache Spark™" is followed by four steps: 1. Choose a Spark release (3.1.3 (Feb 18 2022) selected), 2. Choose a package type (Pre-built for Apache Hadoop 3.2 and later selected), 3. Download Spark (link to spark-3.1.3-bin-hadoop3.2.tgz), and 4. Verify this release using the 3.1.3 signatures, checksums and project release KEYS. A note mentions Spark 3 is pre-built with Scala 2.12. On the right, a "Latest News" sidebar lists Spark 3.1.3, 3.2.1, 3.2.0, and 3.0.3 releases with their dates.

Download Apache Spark™

1. Choose a Spark release:
2. Choose a package type:
3. Download Spark: [spark-3.1.3-bin-hadoop3.2.tgz](#)
4. Verify this release using the 3.1.3 [signatures](#), [checksums](#) and [project release KEYS](#).

Note that Spark 3 is pre-built with Scala 2.12 in general and Spark 3.2+ provides additional pre-built distribution with Scala 2.13

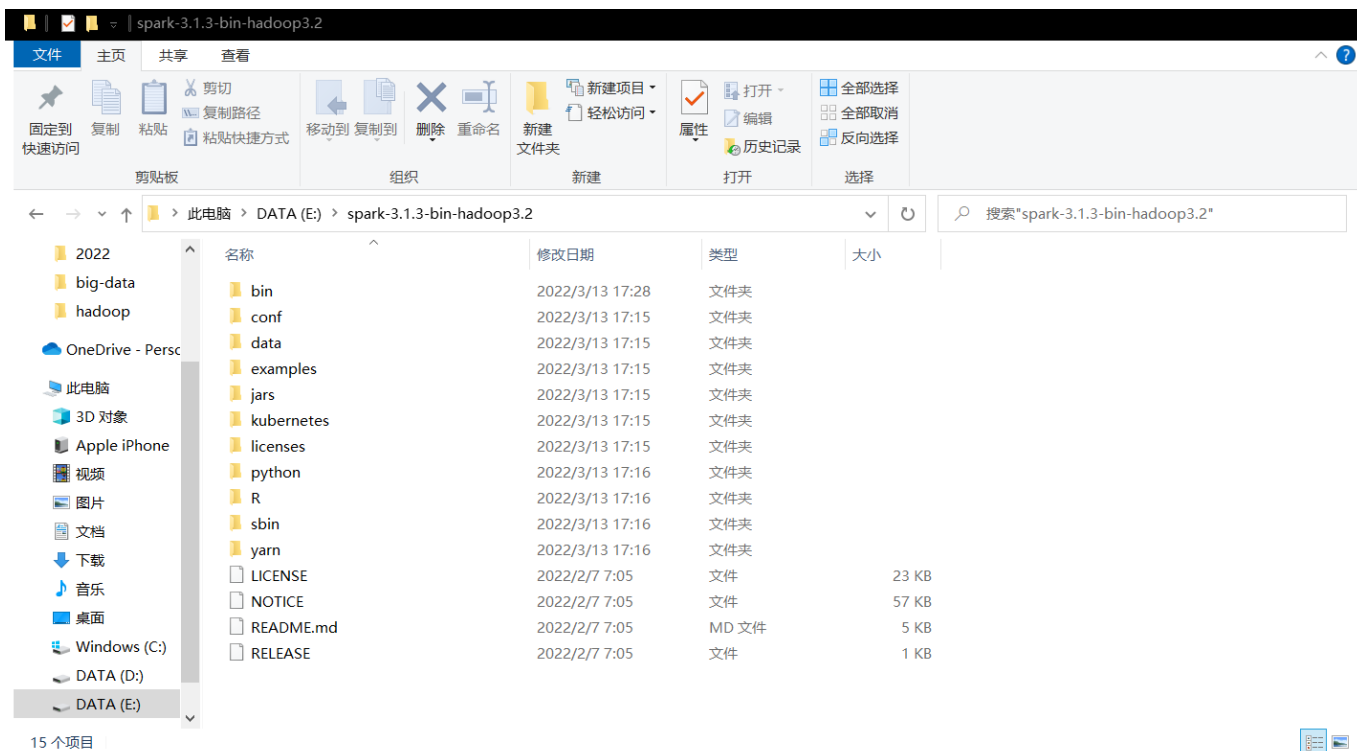
Latest News

- Spark 3.1.3 released (Feb 18, 2022)
- Spark 3.2.1 released (Jan 26, 2022)
- Spark 3.2.0 released (Oct 13, 2021)
- Spark 3.0.3

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

1. 下载spark

下载后得到spark-3.5.6-bin-hadoop3.3，将其解压缩，得到如下的文件夹



2. 下载安装java

安装java 21

<https://www.oracle.com/java/technologies/downloads>

ORACLE

Products Industries Resources Customers Partners Developers Events

View Accounts

Contact Sales

Java downloads Tools and resources Java archive

Java SE subscribers have more choices

Also available for development, personal use, and to run other licensed Oracle products.

Java 8 Java 11

Java SE Development Kit 8u321

Java SE subscribers will receive JDK 8 updates until at least December of 2030.

The Oracle JDK 8 license changed in April 2019

The [Oracle Technology Network License Agreement for Oracle Java SE](#) is substantially different from prior Oracle JDK 8 licenses. This license permits certain uses, such as personal use and development use, at no cost -- but other uses authorized under prior Oracle JDK licenses may no longer be available. Please review the terms carefully before downloading and using this product. FAQs are available [here](#).

Commercial license and support are available for a low cost with [Java SE Subscription](#).

JDK 8 software is licensed under the [Oracle Technology Network License Agreement for Oracle Java SE](#).

JDK 8u321 [checksum](#)

Linux macOS Solaris **Windows**

Product/file description	File size	Download
x86 Installer	157.99 MB	 jdk-8u321-windows-i586.exe
x64 Installer	171.09 MB	 jdk-8u321-windows-x64.exe

2. 下载安装java

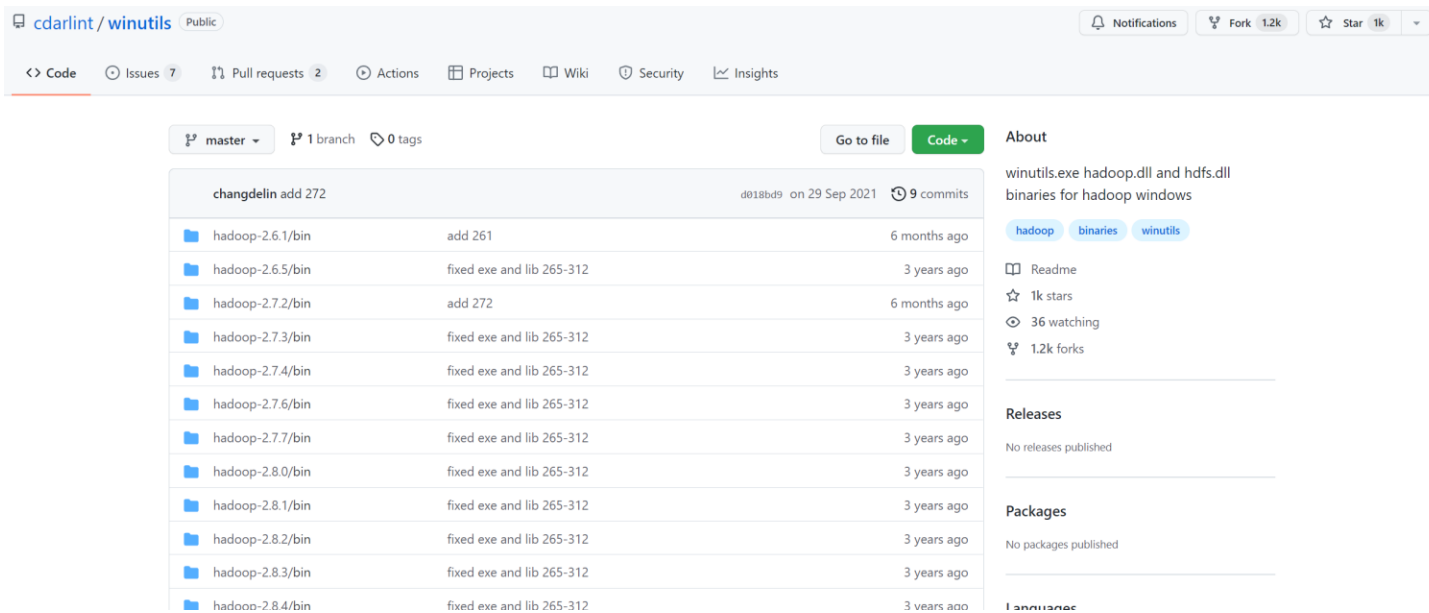
安装过程中会出现如下界面，单击“更改”，以将java安装在一个不带空格的路径(默认路径里 Program Files的中间带有空格)，比如安装到 D:\Java\jdk-21



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>



cdarlint/winutils Public

Notifications Fork 1.2k Star 1k

<> Code Issues 7 Pull requests 2 Actions Projects Wiki Security Insights

master 1 branch 0 tags

Go to file Code

About

winutils.exe hadoop.dll and hdfs.dll binaries for hadoop windows

hadoop binaries winutils

Readme

1k stars

36 watching

1.2k forks

Releases

No releases published

Packages

No packages published

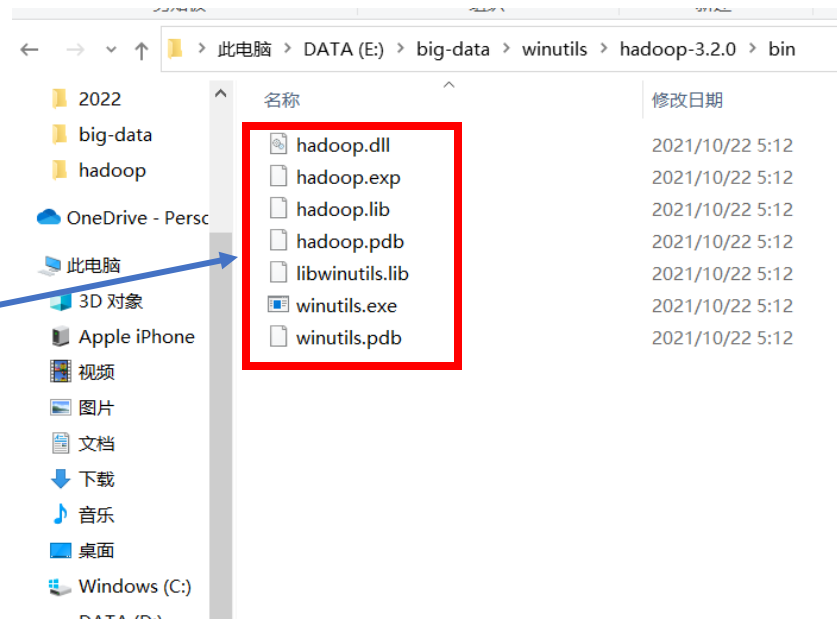
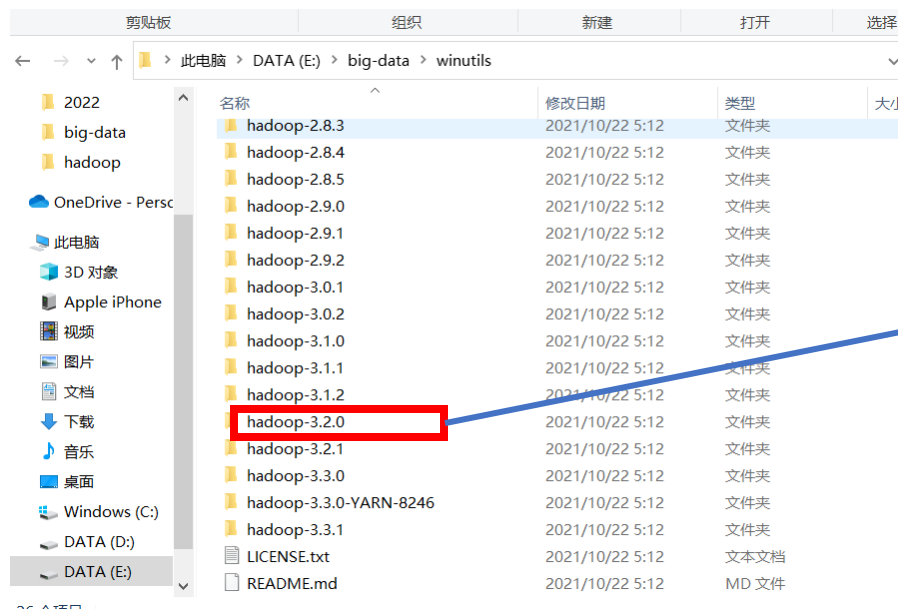
Language

commit message	commit hash	date	commit count
changdelin add 272	d018bd9	on 29 Sep 2021	9 commits
hadoop-2.6.1/bin add 261		6 months ago	
hadoop-2.6.5/bin fixed exe and lib 265-312		3 years ago	
hadoop-2.7.2/bin add 272		6 months ago	
hadoop-2.7.3/bin fixed exe and lib 265-312		3 years ago	
hadoop-2.7.4/bin fixed exe and lib 265-312		3 years ago	
hadoop-2.7.6/bin fixed exe and lib 265-312		3 years ago	
hadoop-2.7.7/bin fixed exe and lib 265-312		3 years ago	
hadoop-2.8.0/bin fixed exe and lib 265-312		3 years ago	
hadoop-2.8.1/bin fixed exe and lib 265-312		3 years ago	
hadoop-2.8.2/bin fixed exe and lib 265-312		3 years ago	
hadoop-2.8.3/bin fixed exe and lib 265-312		3 years ago	
hadoop-2.8.4/bin fixed exe and lib 265-312		3 years ago	

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

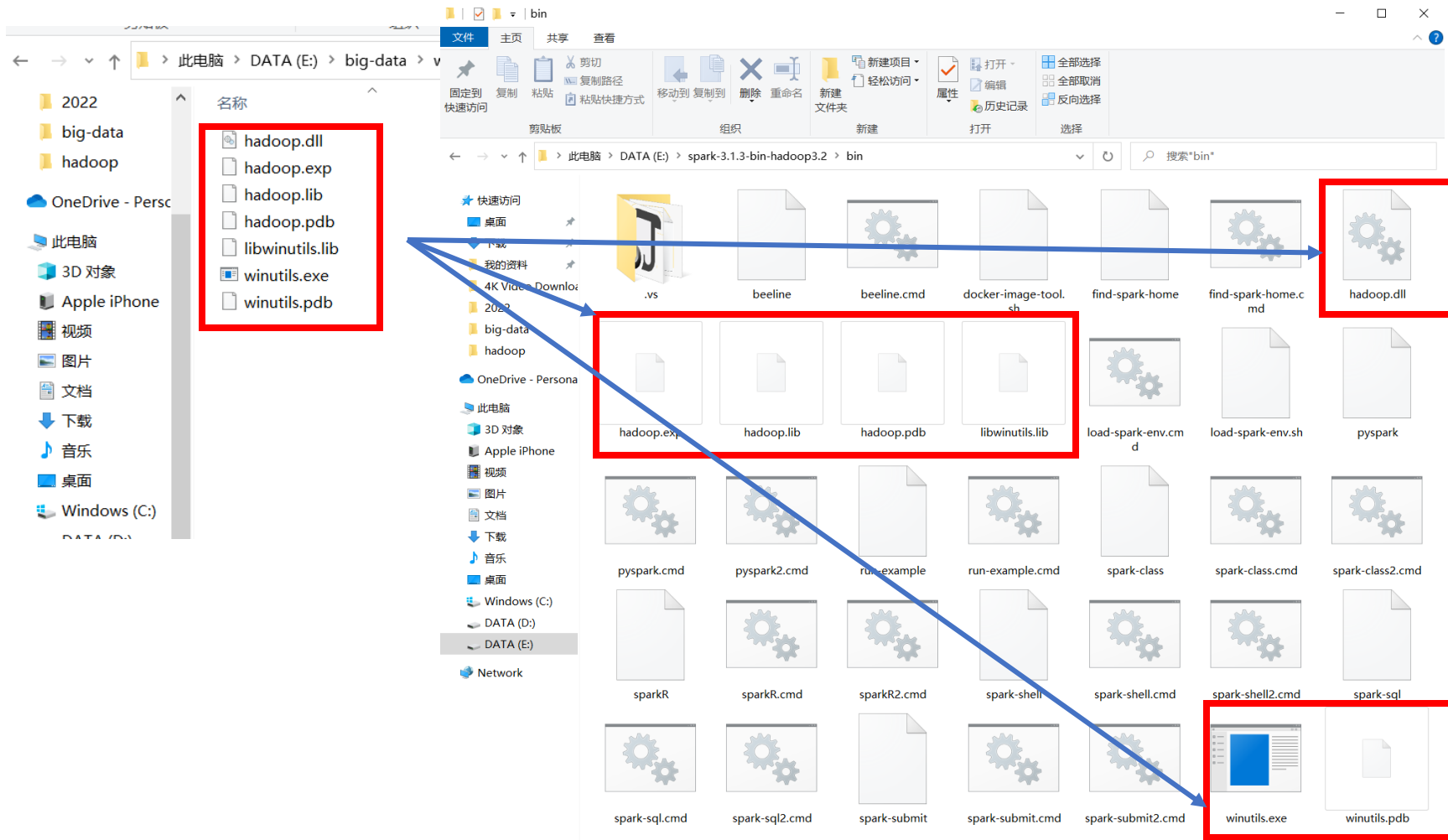
3. 下载winutils.exe

- 因为我们在第一步下载的是Hadoop 3.3，所以我们提取hadoop-3.3/bin下的文件，将其复制到spark-3.5.6-bin-hadoop3.3 /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



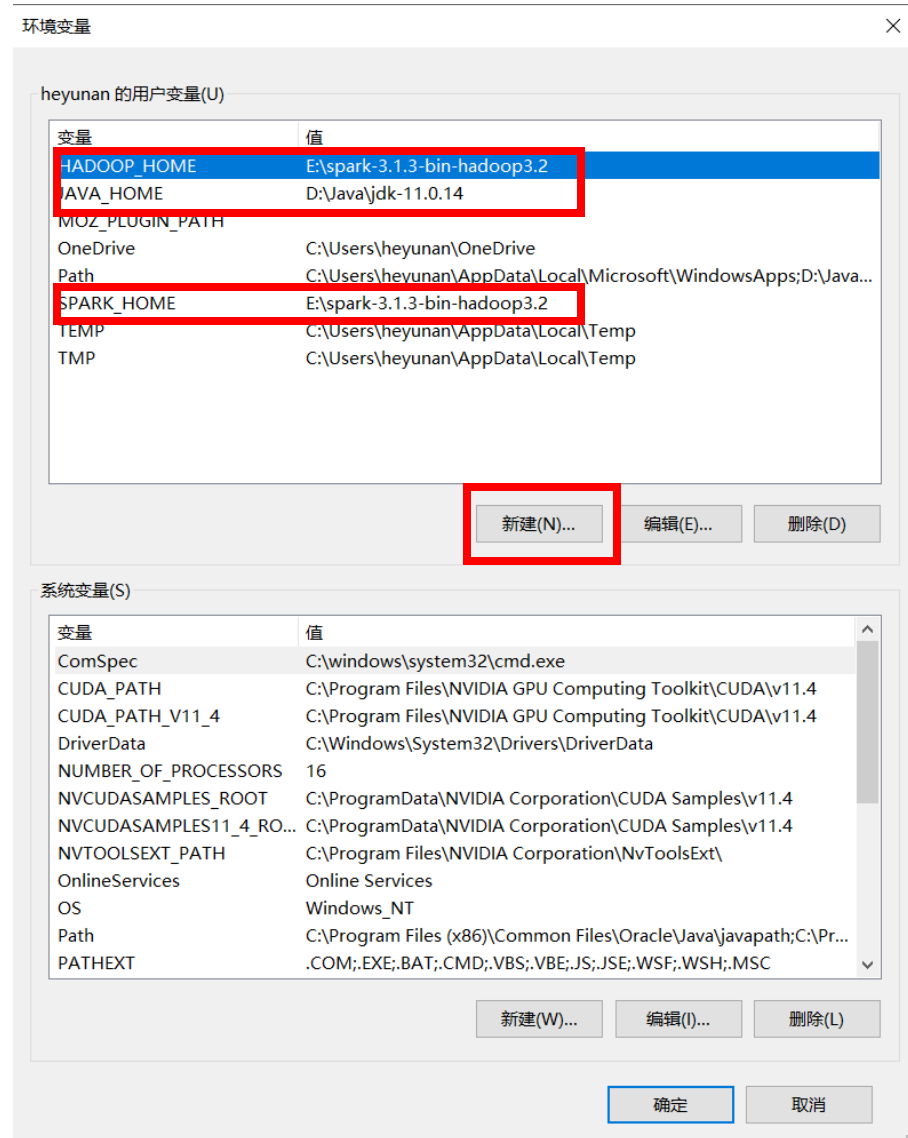
4.编辑环境变量

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。



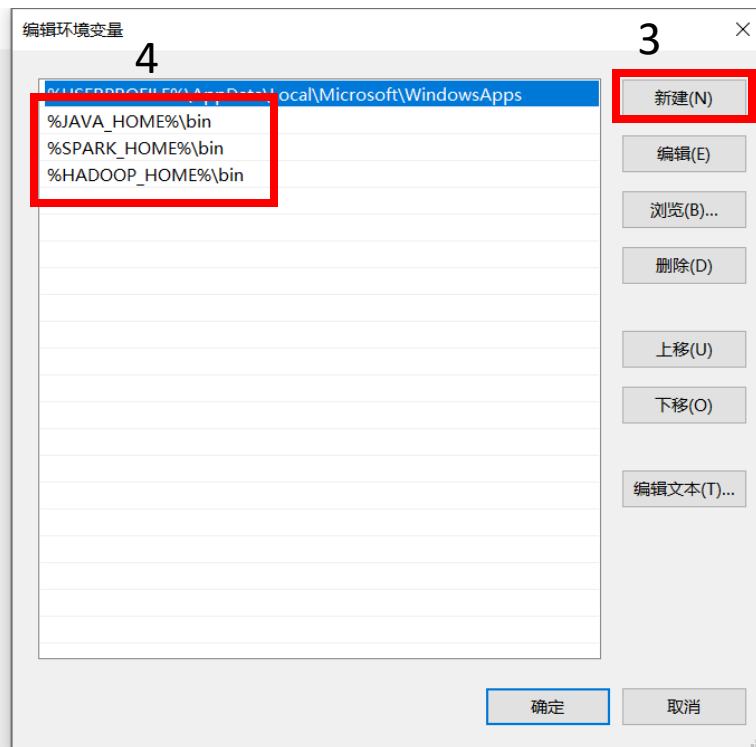
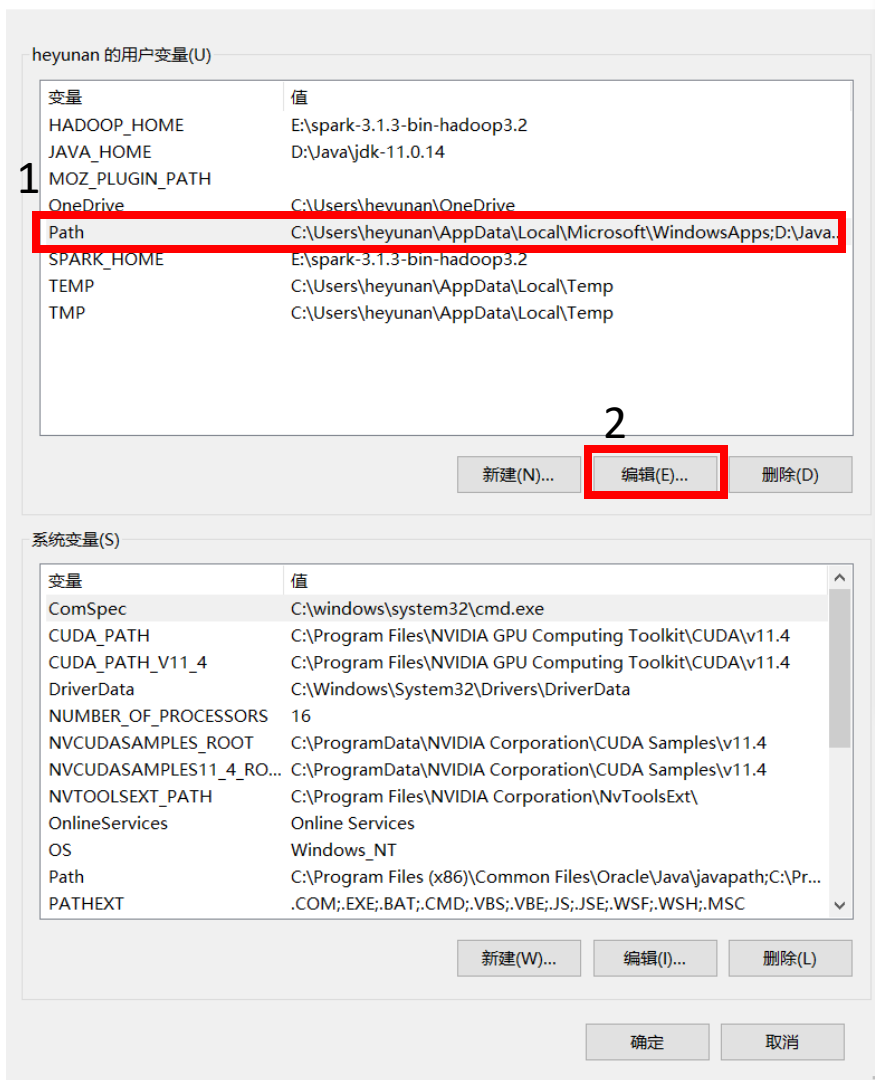
4.编辑环境变量

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-3.5.6-bin-hadoop3.3 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

环境变量

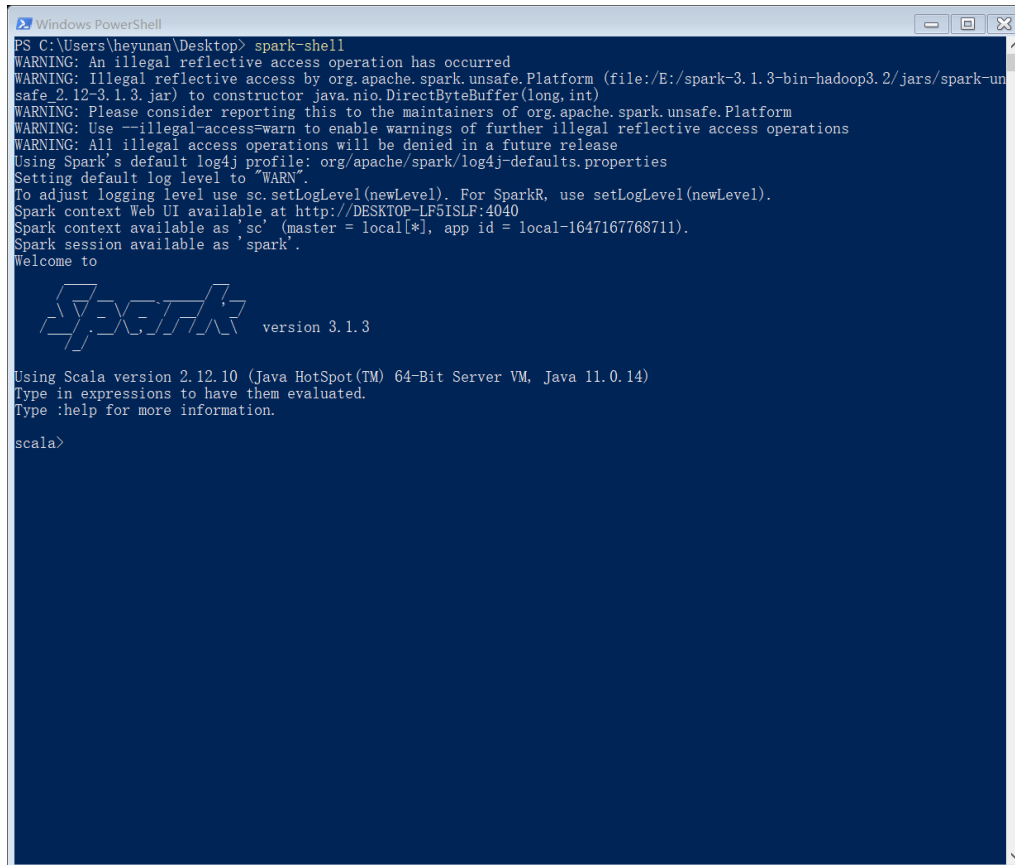


编辑"Path", 并设置如下路径到"Path"

%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：



```
Windows PowerShell
PS C:\Users\heyunan\Desktop> spark-shell
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/E:/spark-3.1.3-bin-hadoop3.2/jars/spark-un
safe_2.12-3.1.3.jar) to constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://DESKTOP-LF5ISLF:4040
Spark context available as 'sc' (master = local[*], app id = local-l647167768711).
Spark session available as 'spark'.
Welcome to

  SPARK  version 3.1.3

Using Scala version 2.12.10 (Java HotSpot(TM) 64-Bit Server VM, Java 11.0.14)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

5.调试

打开 <http://localhost:4040/>，会显示如下界面



Executors

[Show Additional Metrics](#)

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(1)	0	0.0 B / 434.4 MiB	0.0 B	16	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(1)	0	0.0 B / 434.4 MiB	0.0 B	16	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0

Executors

Show 20 entries

Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Thread Dump
driver	DESKTOP-LF5ISLF:53232	Active	0	0.0 B / 434.4 MiB	0.0 B	16	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	Thread Dump

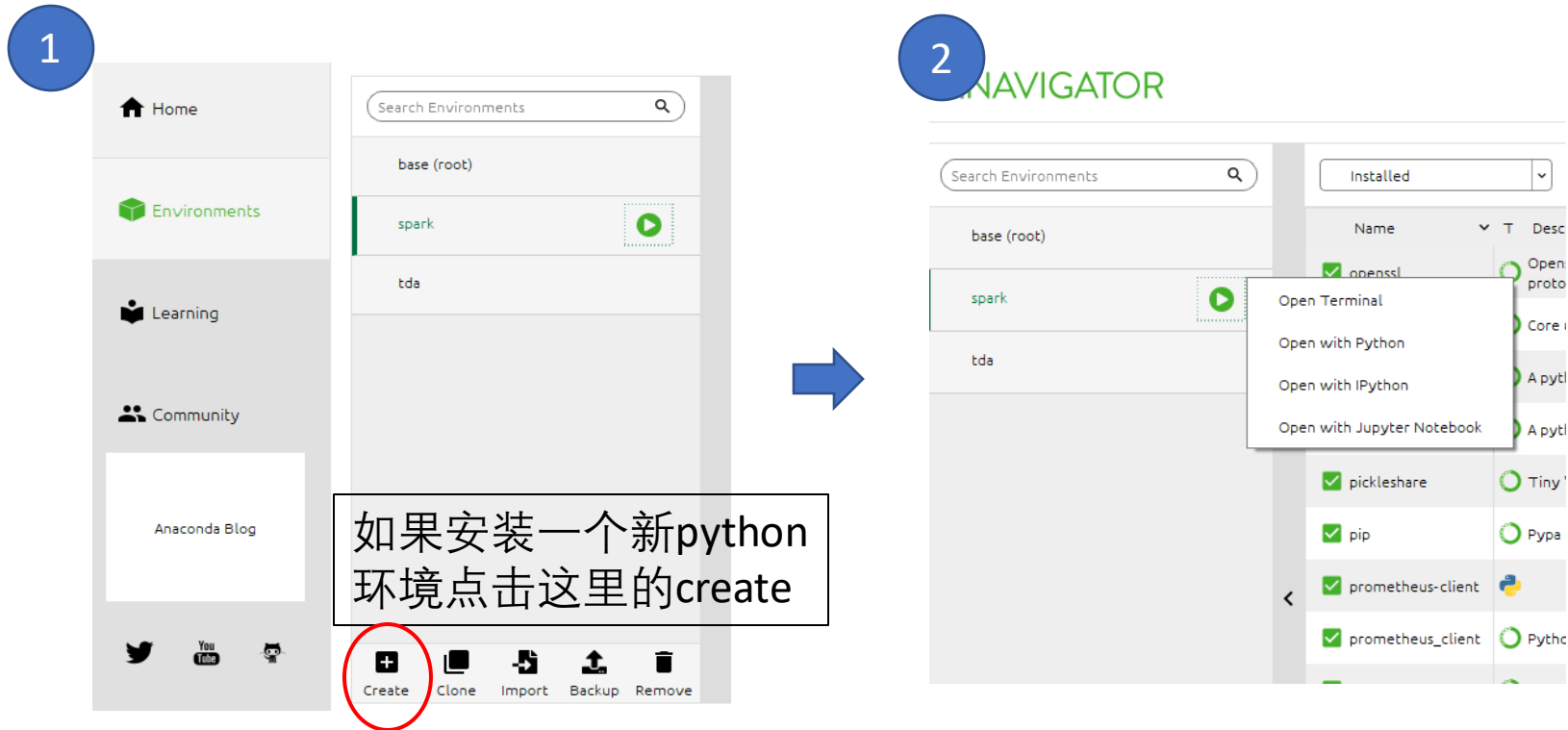
Showing 1 to 1 of 1 entries

Previous 1 Next

6. 在python环境中使用PySpark

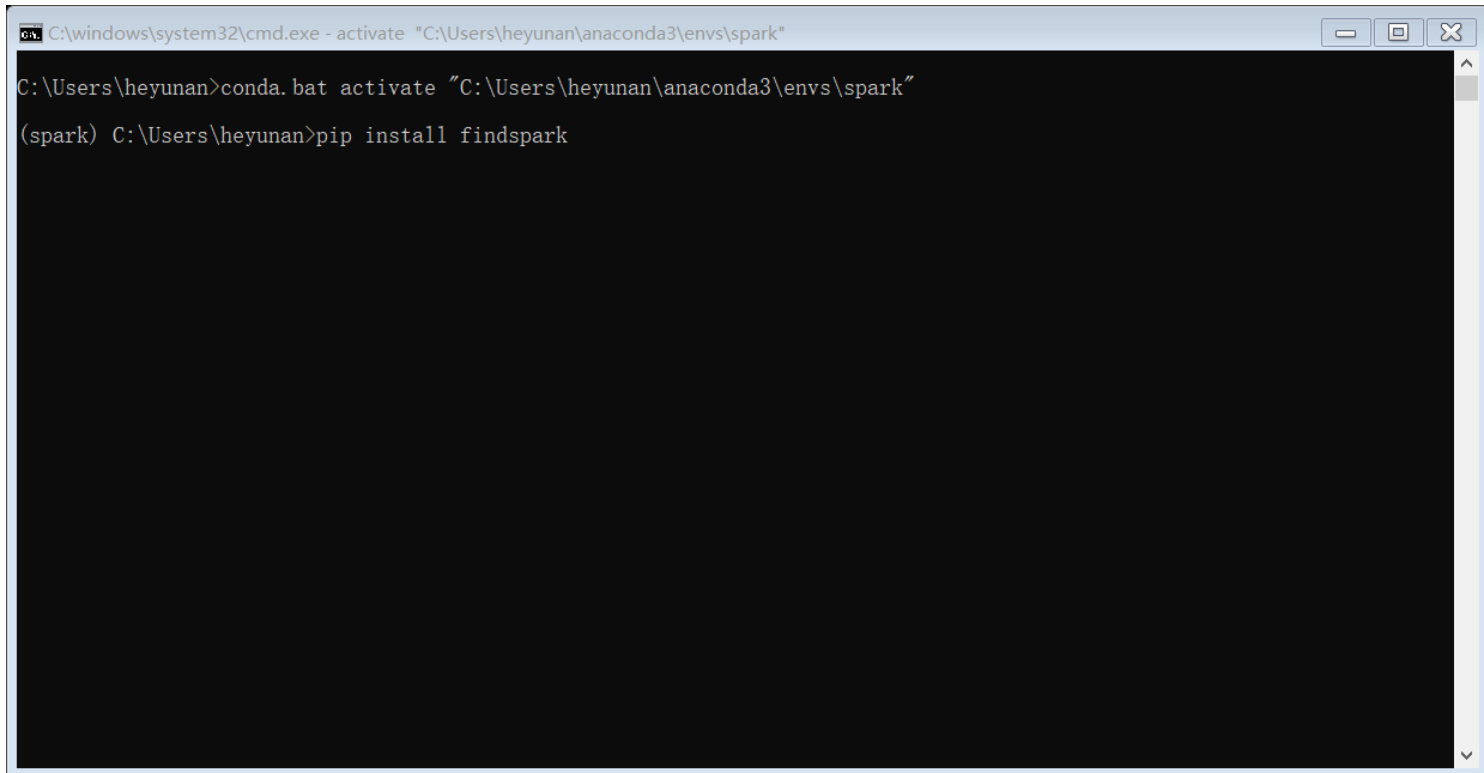
进入Anaconda的navigator中，选择现有python环境的terminal或者新建一个python环境（Python版本要大于或者等于3.7）并打开其terminal（截图中我新建一个python环境并命名为spark）。

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



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。



```
Ca C:\windows\system32\cmd.exe - activate "C:\Users\heyunan\anaconda3\envs\spark"
C:\Users\heyunan>conda.bat activate "C:\Users\heyunan\anaconda3\envs\spark"
(spark) C:\Users\heyunan>pip install findspark
```

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()
```

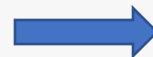


通过findspark找到系统中spark的路径, 每次编程前都需要用findspark导入pyspark。

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

Hello, Spark

```
In [24]: from pyspark.sql import SparkSession  
  
spark = SparkSession.builder.getOrCreate()  
df = spark.sql("select 'spark' as hello ")  
  
df.show()
```



Hello, spark示例程序

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