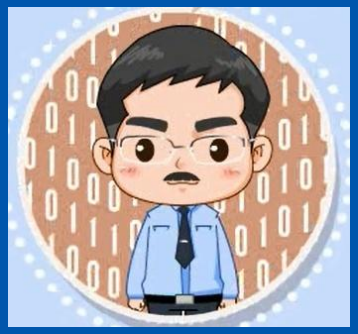


《数据采集与预处理》

教材官网： http://dblab.xmu.edu.cn/post/data-collection/

温馨提示：编辑幻灯片母版，可以修改每页PPT的厦大校徽和底部文字

第2章 大数据实验环境搭建

（PPT版本号：2022年1月版本）

林子雨 副教授

厦门大学计算机科学系

E-mail: ziyulin@xmu.edu.cn

主页：**http://dblab.xmu.edu.cn/linziyu**

提纲

2.1 Python的安装和使用
2.2 JDK的安装
2.3 MySQL数据库的安装和使用
2.4 Hadoop的安装和使用

本PPT是以下教材的配套讲义
林子雨编著《数据采集与预处理》
人民邮电出版社

教材官网：
http://dblab.xmu.edu.cn/post/data-collection

2.1 Python的安装和使用

2.1.1 Python简介
2.1.2 Python的安装
2.1.3 Python的基本使用方法
2.1.4 Python基础语法知识
2.1.5 Python第三方模块的安装

本PPT是以下教材的配套讲义
林子雨编著《数据采集与预处理》
人民邮电出版社

教材官网：
http://dblab.xmu.edu.cn/post/data-collection

2.1.1 Python简介

Python（发音[ˈpaɪθən]）是1989年由荷兰人Guido van Rossum发明的一种面向对象的解释型高级编程语言。Python的第一个公开发行版发行于1991年，自从2004年以后，Python的使用率呈线性增长。TIOBE在2019年1月发布的排行榜显示，Python获得“TIOBE最佳年度语言”称号，这是Python第3次获得“TIOBE最佳年度语言”，也是获奖次数最多的编程语言。发展到今天，Python已经成为最受欢迎的程序设计语言之一。

Python常被称为“胶水语言”，能够把用其它语言制作的各种模块（尤其是C/C++）很轻松地连接在一起。常见的一种应用情形是，使用Python快速生成程序的原型（有时甚至是程序的最终界面），然后对其中有特别要求的部分，用更合适的语言改写，比如3D游戏中的图形渲染模块，性能要求特别高，就可以用C/C++重写，而后封装为Python可以调用的扩展类库。

2.1.1 Python简介

Python的设计哲学是“优雅”、“明确”、“简单”。在设计Python语言时，如果面临多种选择，Python开发者一般会拒绝花哨的语法，而选择明确地没有或者很少有歧义的语法。总体来说，选择Python开发程序具有简单、开发速度快、节省时间和精力等特点，因此，在Python开发领域流传着这样一句话：“人生苦短，我用Python”。

Python作为一门高级编程语言，虽然诞生的时间并不长，但是发展速度很快，已经成为很多编程爱好者开展入门学习的第一门编程语言。总体而言，Python语言具有以下优点：

（1）语言简单。
（2）开源、免费。
（2）面向对象。
（3）跨平台。
（4）强大的生态系统。

2.1.1 Python简介

丰富的生态系统也给专业开发者带来了极大的便利。大量成熟的第三方库可以直接使用，专业开发者只需要使用很少的语法结构就可以编写出功能强大的代码，缩短了开发周期，提高了开发效率。常用的Python第三方库包括Matplotlib（数据可视化库）、NumPy（数值计算功能库）、SciPy（数学、科学、工程计算功能库）、pandas（数据分析高层次应用库）、Scrapy（网络爬虫功能库）、BeautifulSoup（HTML和XML的解析库）、Django（Web应用框架）、Flask（Web应用微框架）等。

2.1.2 Python的安装

Python自发布以来，主要经历了3个版本的变化，分别是1994年发布的1.0版本、2000年发布的2.0版本和2008年发布的3.0版本。其中，1.0版本已经过时，2.0和3.0版本都在保持持续更新。Python官方网站目前同时发行Python2.x和Python3.x两个不同系列的版本，并且彼此之间不兼容，除了输入输出方式有所不同，很多内置函数的实现和使用方式也有较大差别。本教程使用Python3.8.7版本，理由如下：

（1）Python2.x和Python3.x的思想是共通的。

（2）使用Python3.x是大势所趋。

2.1.2 Python的安装

Python可以用于多种平台，包括Windows、Linux和Mac OS等。本教程采用的操作系统是Windows7或以上版本，使用的Python版本是3.8.7。请到Python官网（https://www.python.org/）下载与自己计算机操作系统匹配的安装包，比如，64位Windows操作系统可以下载python-3.8.7-amd64.exe。在安装过程中，要注意选中“Add python 3.8 to PATH”，这样可以在安装过程中自动配置PATH环境变量，避免了手动配置的繁琐过程。

安装成功以后，需要检测是否安装成功。可以打开Windows系统的cmd命令界面，并在命令提示符后面输入“python”后回车，如果出现如图2-1所示信息，则说明Python已经安装成功。

2.1.3 Python的基本使用方法

假设在Windows系统的C盘根目录下已经存在一个代码文件hello.py，该文件里面只有如下一行代码：

```
print("Hello World")
```

现在我们要运行这个代码文件。可以打开Windows系统的cmd命令界面，并在命令提示符后面输入如下语句：

```
$ python C:\hello.py
```

运行结果如图2-2所示。

图2-2 在cmd命令界面中执行Python代码文件

2.1.3 Python的基本使用方法

Python安装成功以后，会自带一个集成式开发环境IDLE，它是一个Python Shell，程序开发人员可以利用Python Shell与Python交互。

在Windows系统的“开始”菜单中找到“IDLE(Python 3.8 64-bit)”，打开进入IDLE主窗口（如图2-3所示），窗口左侧会显示Python命令提示符“>>>”，可以在提示符后面输入Python代码，回车后就会立即执行并返回结果。

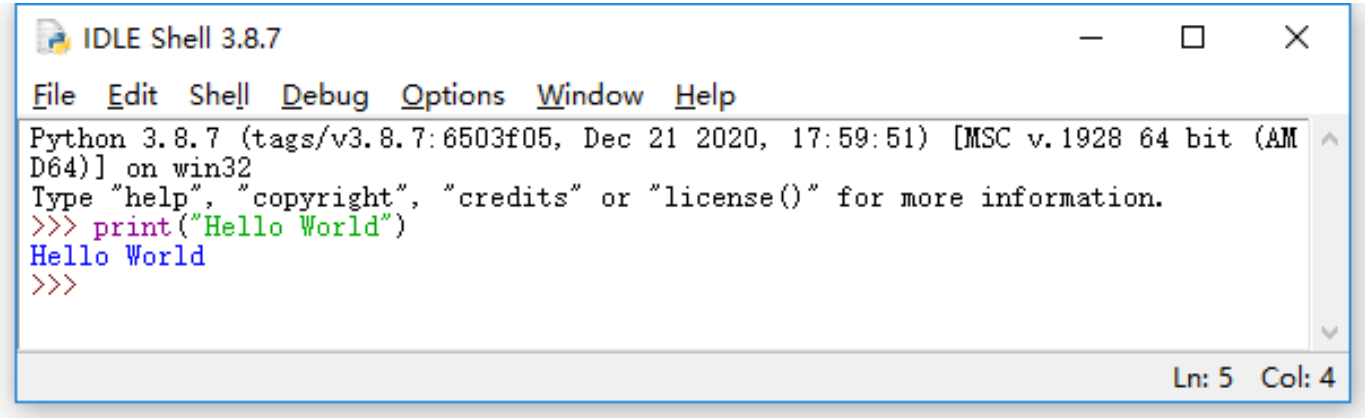


图2-3 IDLE主窗口

2.1.3 Python的基本使用方法

如果要创建一个代码文件，可以在IDLE主窗口的顶部菜单栏中选择“File->New File”，然后就会弹出如图2-4所示的文件窗口，可以在里面输入Python代码，然后，在顶部菜单栏中选择“File->Save As...”，把文件保存为hello.py。

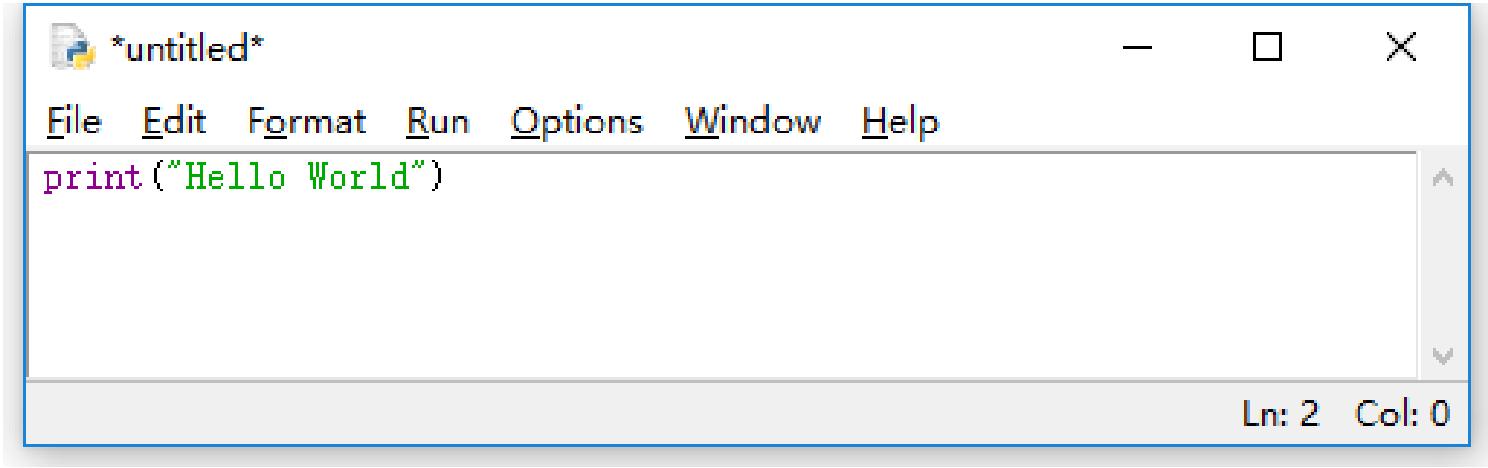
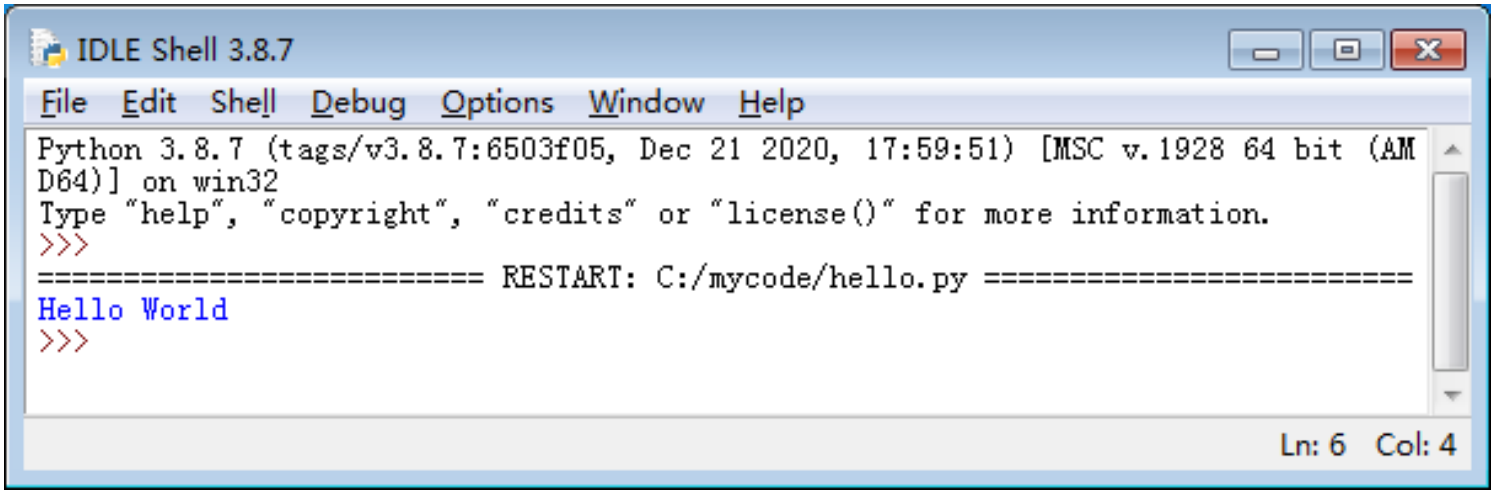


图2-4 IDLE的文件窗口

2.1.3 Python的基本使用方法

如果要运行代码文件hello.py，可以在IDLE的文件窗口的顶部菜单栏中，选择“Run->Run Module”，这时就会开始运行程序。程序运行结束后，会在IDLE Shell窗口显示执行结果（如图2-5所示）。



2.1.3 Python的基本使用方法

在实际开发中，可以通过使用IDLE提供的快捷键（如表2-1所示）来提高程序开发效率。

表2-1 IDLE常用快捷键

快捷键	功能说明
F1	打开Python帮助文档
Ctrl+]	缩进代码块
Ctrl+[	取消代码块缩进
Ctrl+F6	重新启动IDLE Shell
Ctrl+Z	撤销一步操作
Ctrl+Shift+Z	恢复上一次的撤销操作
Ctrl+S	保存文件
Alt+P	浏览历史命令（上一条）
Alt+N	浏览历史命令（下一条）
Alt+/	自动补全前面曾经出现过的单词，如果之前有多个单词具有相同前缀，可以连续按该快捷键，在多个单词中循环选择
Alt+3	注释代码块
Alt+4	取消代码块注释
Alt+g	转到某一行

2.1.4 Python基础语法知识

1.基本数据类型

Python3.x中有6个标准的数据类型，分别是数字、字符串、列表、元组、字典和集合。这6个标准的数据类型又可以进一步划分为基本数据类型和组合数据类型。其中，数字和字符串是基本数据类型；列表、元组、字典和集合是组合数据类型。

（1）数字

在Python中，数字类型包括整数（int）、浮点数（float）、布尔类型（bool）和复数（complex），而且，数字类型变量可以表示任意大的数值。

2.1.4 Python基础语法知识

①整数。整数类型用来存储整数数值。在Python中，整数包括正整数、负整数和0。按照进制的不同，整数类型还可以划分为十进制整数、八进制整数、十六进制整数和二进制整数。
②浮点数。浮点数也称为“小数”，由整数部分和小数部分构成，比如3.14、0.2、-1.648、5.87268492678 42等。浮点数也可以用科学计数法表示，比如1.3e4、-0.35e3、2.36e-3等。
③布尔类型。Python中的布尔类型主要用来表示“真”或“假”的值，每个对象天生具有布尔类型的True或False值。空对象、值为零的任何数字或者对象None的布尔值都是False。在Python3.x中，布尔值是作为整数的子类实现的，布尔值可以转换为数值，True的值为1，False的值为0，可以进行数值运算。
④复数。复数由实数部分和虚数部分构成，可以用a + bj或者complex(a,b)表示，复数的实部a和虚部b都是浮点型。例如，一个复数的实部为2.38，虚部为18.2j，则这个复数为2.38+18.2j。

2.1.4 Python基础语法知识

（2）字符串

字符串是 Python 中最常用的数据类型，它是连续的字符序列，一般使用单引号（' '）、双引号（" "）或三引号（''' '''或""" """）进行界定。其中，单引号和双引号中的字符序列必须在一行上，而三引号内的字符序列可以分布在连续的多行上，从而可以支持格式较为复杂的字符串。

例如，' xyz'、'123'、'厦门'、"hadoop"、"'spark"'、"""flink"""都是合法字符串，空字符串可以表示为''、" "或''' '''。

2.1.4 Python基础语法知识

2.序列

数据结构是通过某种方式组织在一起的数据元素的集合。序列是Python中最基本的数据结构，是指一块可存放多个值的连续内存空间，这些值按一定的顺序排列，可通过每个值所在位置的索引访问它们。在Python 中，序列类型包括字符串、列表、元组、字典和集合。

2.1.4 Python基础语法知识

（1）列表

列表是最常用的Python数据类型，列表的数据项不需要具有相同的类型。在形式上，只要把逗号分隔的不同的数据项使用方括号括起来，就可以构成一个列表，例如：

['hadoop', 'spark', 2021, 2010]
[1, 2, 3, 4, 5]
["a", "b", "c", "d"]
['Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Sunday']

同其他类型的Python变量一样，在创建列表时，也可以直接使用赋值运算符“=”将一个列表赋值给变量。例如，以下都是合法的列表定义：

student = ['小明', '男', 2010,10]
num = [1, 2, 3, 4, 5]
motto = ["自强不息","止于至善"]
list = ['hadoop', '年度畅销书',[2020,12000]]

可以看出，列表里面的元素仍然可以是列表。需要注意的是，尽管一个列表中可以放入不同类型的数据，但是，为了提高程序的可读性，一般建议在一个列表中只出现一种数据类型。

2.1.4 Python基础语法知识

（2）元组

Python中的列表适合存储在程序运行时变化的数据集。列表是可以修改的，这对要存储一些要变化的数据而言至关重要。但是，也不是任何数据都要在程序运行期间进行修改，有时候需要创建一组不可修改的元素，此时可以使用元组。

元组的创建和列表的创建很相似，不同之处在于，创建列表时使用的是方括号，而创建元组时则需要使用圆括号。元组的创建方法很简单，只需要在圆括号中添加元素，并使用逗号隔开即可，具体实例如下：

```
>>> tuple1 = ('hadoop','spark',2008,2009)
>>> tuple2 = (1,2,3,4,5)
>>> tuple3 = ('hadoop',2008,("大数据","分布式计算
"),["spark","flink","storm"])
```

2.1.4 Python基础语法知识

（3）字典

字典也是Python提供的一种常用的数据结构，它用于存放具有映射关系的数据。比如有一份学生成绩表数据，语文67分，数学91分，英语78分，如果使用列表保存这些数据，则需要两个列表，即["语文","数学","英语"]和[67,91,78]。但是，使用两个列表来保存这组数据以后，就无法记录两组数据之间的关联关系。为了保存这种具有映射关系的数据，Python 提供了字典，字典相当于保存了两组数据，其中一组数据是关键数据，被称为“键”（key）；另一组数据可通过键来访问，被称为“值”（value）。

2.1.4 Python基础语法知识

字典具有如下特性：

•字典的元素是“键值对”，由于字典中的键是非常关键的数据，而且程序需要通过键来访问值，因此字典中的键不允许重复，必须是唯一值，而且键必须不可变；

•字典不支持索引和切片，但可以通过“键”查询“值”；

•字典是无序的对象集合，列表是有序的对象集合，两者之间的区别在于，字典当中的元素是通过键来存取的，而不是通过偏移量存取；

•字典是可变的，并且可以任意嵌套。

字典用大括号{}标识。在使用大括号语法创建字典时，大括号中应包含多个“键值对”，键与值之间用英文冒号隔开，多个键值对之间用英文逗号隔开。具体实例如下：

```
>>> grade = {"语文":67, "数学":91, "英语":78}  #键是字符串
>>> grade
{'语文': 67, '数学': 91, '英语': 78}
```

2.1.4 Python基础语法知识

（4）集合

集合（set）是一个无序的不重复元素序列。集合中的元素必须是不可变类型。在形式上，集合的所有元素都放在一对大括号“{}”中，两个相邻的元素之间使用逗号分隔。

可以直接使用大括号{}创建集合，实例如下：

```
>>> dayset = {'Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday',
'Saturday', 'Sunday'}
>>> dayset
{'Tuesday', 'Monday', 'Wednesday', 'Saturday', 'Thursday', 'Sunday', 'Friday'}
```

在创建集合时，如果存在重复元素，Python只会自动保留一个，实例如下：

```
>>> numset = {2,5,7,8,5,9}
>>> numset
{2, 5, 7, 8, 9}
```

2.1.4 Python基础语法知识

3.控制结构

（1）选择语句

选择语句也称为“条件语句”，就是对语句中不同条件的值进行判断，从而根据不同的条件执行不同的语句。

选择语句可以分为以下3种形式：

简单的if语句；

if…else语句；

if…elif…else多分支语句。

2.1.4 Python基础语法知识

【例2-1】使用if语句求出两个数的较小值。

```
1. # two_number.py
2. a,b,c = 4,5,0
3. if a>b:
4.     c = b
5. if a<b:
6.     c = a
7. print("两个数的较小值是： ",c)
```

2.1.4 Python基础语法知识

【例2-2】判断一个数是奇数还是偶数。

```
1.  # odd_even.py
2.  a = 5
3.  if a % 2 == 0:
4.      print("这是一个偶数。")
5.  else:
6.      print("这是一个奇数。")
```

2.1.4 Python基础语法知识

【例2-3】判断每天上课的内容。

```
1.  # lesson.py
2.  day = int(input("请输入第几天课程： "))
3.  if day == 1:
4.      print("第1天上数学课")
5.  elif day == 2:
6.      print("第2天上语文课")
7.  else:
8.      print("其他时间上计算机课")
```

2.1.4 Python基础语法知识

（2）循环语句

循环语句就是重复执行某段程序代码，直到满足特定条件为止。在Python语言中，循环语句有以下两种形式：

（1）while循环语句；

（2）for循环语句。

2.1.4 Python基础语法知识

【例2-4】用while循环实现计算1~99的整数和。

```
1. # int_sum.py
2. n = 1
3. sum = 0
4. while(n <= 99):
5.     sum += n
6.     n += 1
7. print("1~99的整数和是：",sum)
```

2.1.4 Python基础语法知识

【例2-5】用for循环实现计算1~99的整数和。

```
1. # int_sum_for.py
2. sum=0
3. for n in range(1,100):  #range(1,100)用于生成
   1到100（不包括100）的整数
4.     sum+=n
5. print("1到99的整数和是： ",sum)
```

2.1.4 Python基础语法知识

4.函数

函数是可以重复使用的用于实现某种功能的代码块。与其他语言类似，在Python中，函数的优点也是提高程序的模块性和代码复用性。

【例2-6】定义一个带有参数的函数。

```
01  # i_like.py
02  # 定义带有参数的函数
03  def like(language):
04      '''打印喜欢的编程语言！'''
05      print("我喜欢{}语言！".format(language))
06      return
07  # 调用函数
08  like("C")
09  like("C#")
10  like("Python")
```

上面代码的执行结果如下：

我喜欢C语言！

我喜欢C#语言！

我喜欢Python语言！

2.1.5 Python第三方模块的安装

Python的强大之处在于它拥有非常丰富的第三方模块（或第三方库），可以帮我们方便、快捷地实现网络爬虫、数据清洗、数据可视化和科学计算等功能。为了便于用于安装和管理第三方库和软件，Python提供了一个扩展模块（或扩展库）管理工具pip，Python3.8.7在安装的时候会默认安装pip。

pip之所以能够成为最流行的扩展模块管理工具，并不是因为它被Python官方作为默认的扩展模块管理器，而是因为它自身有很多优点，主要包括：
pip提供了丰富的功能，包括扩展模块的安装和卸载，以及显示已经安装的扩展模块；
- pip能够很好地支持虚拟环境；
- pip可以集中管理依赖；
- pip能够处理二进制格式；
- pip是先下载后安装，如果安装失败，也会清理干净，不会留下一个中间状态。

2.1.5 Python第三方模块的安装

pip提供的命令不多，但是都很实用。表2-2给出了常用pip命令的使用方法。

表2-2 常用pip命令的使用方法

pip命令	说明
pip install SomePackage	安装SomePackage模块
pip list	列出当前已经安装的所有模块
pip install --upgrade SomePackage	升级SomePackage模块
pip uninstall SomePackage	卸载SomePackage模块

例如，Matplotlib是最著名的Python绘图库，它提供了一整套和Matlab相似的API，十分适合交互式地进行制图，可以使用如下命令安装Matplotlib：

```
$ pip install matplotlib
```

安装成功以后，使用如下命令就可以看到安装的Matplotlib：

```
$ pip list
```

2.2 JDK的安装

访问Oracle官网（https://www.oracle.com/java/technologies/javase/javase-jdk8-downloads.html）下载JDK安装包并完成安装。安装完成后需要设置Path环境变量，右键点击“我的电脑”->“高级系统设置”->“环境变量”，然后，在用户变量Path中加入类似如下的信息：

C:\Program Files\Java\jdk1.8.0_111\bin

这个新添加的值和此前已经存在的值之间用英文分号隔开（如图2-6所示）。上面的“jdk1.8.0_111”是刚才已经安装的JDK的版本号。

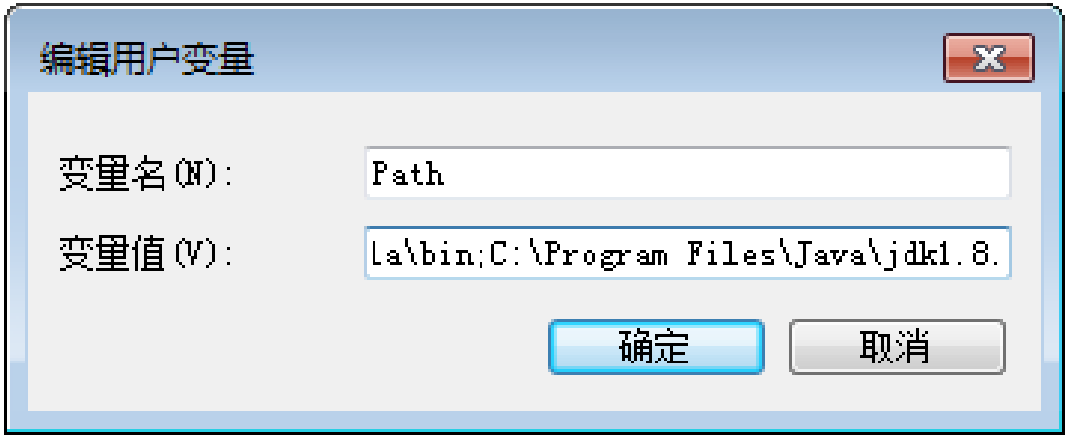


图2-6 编辑用户变量

2.2 JDK的安装

然后，再新建一个环境变量JAVA_HOME，把它的值设置为如下内容（如图2-7所示）：
C:\Program Files\Java\jdk1.8.0_111

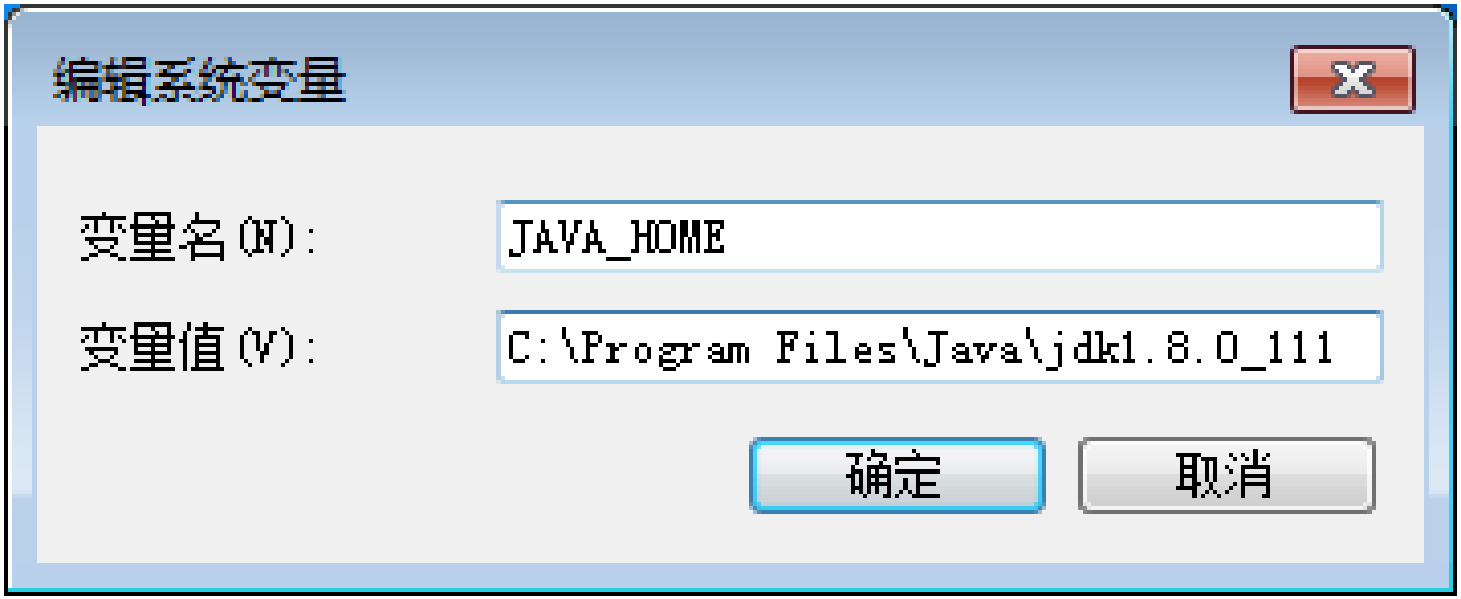


图2-7 编辑系统变量

2.2 JDK的安装

打开cmd窗口，输入“java -version”命令测试是否安装成功，如果安装成功，则会返回如图2-8所示信息。

图2-8 java –version命令执行结果

2.3 MySQL数据库的安装和使用

2.3.1 关系数据库
2.3.2 关系数据库标准语言SQL
2.3.3 安装MySQL
2.3.4 MySQL数据库的使用方法

2.3.1 关系数据库

数据库是一种主流的数据存储和管理技术。数据库指的是以一定方式储存在一起、能为多个用户共享、具有尽可能小的冗余度、与应用程序彼此独立的数据集合。

目前比较主流的数据库是关系数据库，它采用了关系数据模型来组织和管理数据。一个关系数据库可以看成是许多关系表的集合，每个关系表可以看成一张二维表格，如表2-3所示的学生信息表。

表2-3 学生信息表

学号	姓名	性别	年龄	考试成绩
95001	张三	男	21	88
95002	李四	男	22	95
95003	王梅	女	22	73
95004	林莉	女	21	96

2.3.2关系数据库标准语言SQL

结构化查询语言（Structured Query Language，SQL）是关系数据库的标准语言，也是一个通用的功能极强的关系数据库语言，其功能不仅仅是查询，而是包括数据库创建、数据库数据的插入与修改、数据库安全性完整性定义等一系列功能。

SQL的主要特点如下：

（1）综合统一。

（2）高度非过程化。

（3）面向集合的操作方式。

（4）以同一种语法结构提供多种使用方式。

（5）语言简洁，易学易用。

2.3.2关系数据库标准语言SQL

下面介绍一些常用的SQL语句。

1.创建数据库

在使用数据库之前，需要创建数据库，具体语法如下：
CREATE DATABASE 数据库名称;

每条SQL语句的末尾用英文分号结束。

可以使用如下语句查看已经创建的所有数据库：
SHOW DATABASES;

创建好数据库以后，可以使用如下语句打开数据库：
USE 数据库名称;

2.3.2关系数据库标准语言SQL

2.创建表

在一个数据库中，会包含多个表。创建一个表的语法如下：

```
CREATE TABLE 表名称
(
列名称1 数据类型,
列名称2 数据类型,
列名称3 数据类型,
....
);
```

可以使用如下SQL语句查看所有已经创建的表：

```
SHOW TABLES;
```

2.3.2关系数据库标准语言SQL

3.插入数据

可以使用INSERT INTO语句向表中插入新的记录，其语法形式如下：
INSERT INTO 表名称 VALUES (值1, 值2,....);

也可以指定所要插入数据的列：
INSERT INTO表名称(列1, 列2,...) VALUES (值1, 值2,....);

2.3.2关系数据库标准语言SQL

4.查询数据

可以使用SELECT语句从数据库中查询数据，其语法形式如下：

SELECT 列名称 FROM 表名称;

5.修改数据

可以使用UPDATE语句修改表中的数据，其语法形式如下：

UPDATE 表名称 SET 列名称 = 新值 WHERE 列名称 = 某值;

6.删除数据

DELETE FROM 表名称 WHERE列名称 = 某值;

7.删除表

可以使用DROP TABLE语句从数据库中删除一个表，其语法形式如下：

DROP TABLE 表名称;

8.删除数据库

可以使用DROP DATABASE语句删除一个数据库，其语法形式如下：

DROP DATABASE 数据库名称;

2.3.3安装MySQL

访问如下MySQL官网地址下载安装包：
https://dev.mysql.com/downloads/windows/installer/8.0.html

在MySQL下载页面中（如图2-9所示），选择“mysql-installer-community-8.0.23.0.msi”下载。

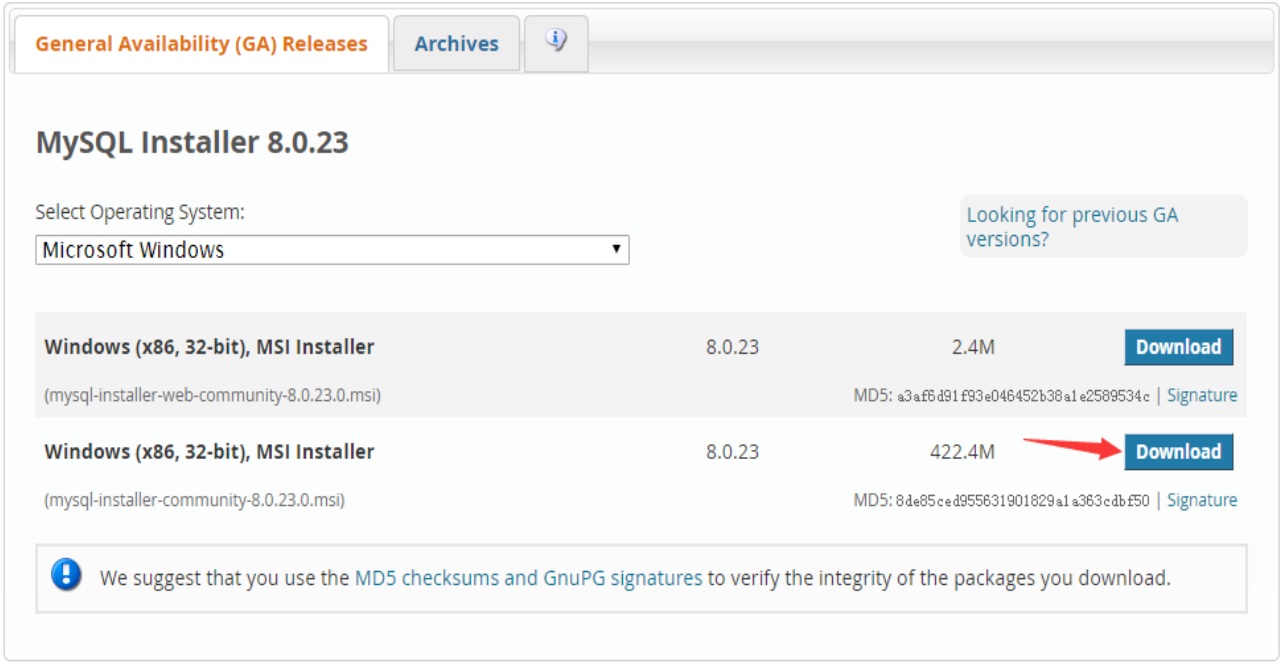


图2-9 MySQL下载网页

2.3.3安装MySQL

使用安装包mysql-installer-community-8.0.23.0.msi开始安装，如果在安装过程中提示需要安装“.NET Framework 4.5.2”，则需要到如下网址下载.NET Framework 4.5.2的安装文件NDP452-KB2901907-x86-x64-AllOS-ENU.exe并安装：

https://www.microsoft.com/zh-CN/download/confirmation.aspx?id=42642

2.3.3安装MySQL

在安装MySQL过程中，当出现“Choosing a Setup Type”界面时，需要选择“Server only”（如图2-10所示）。

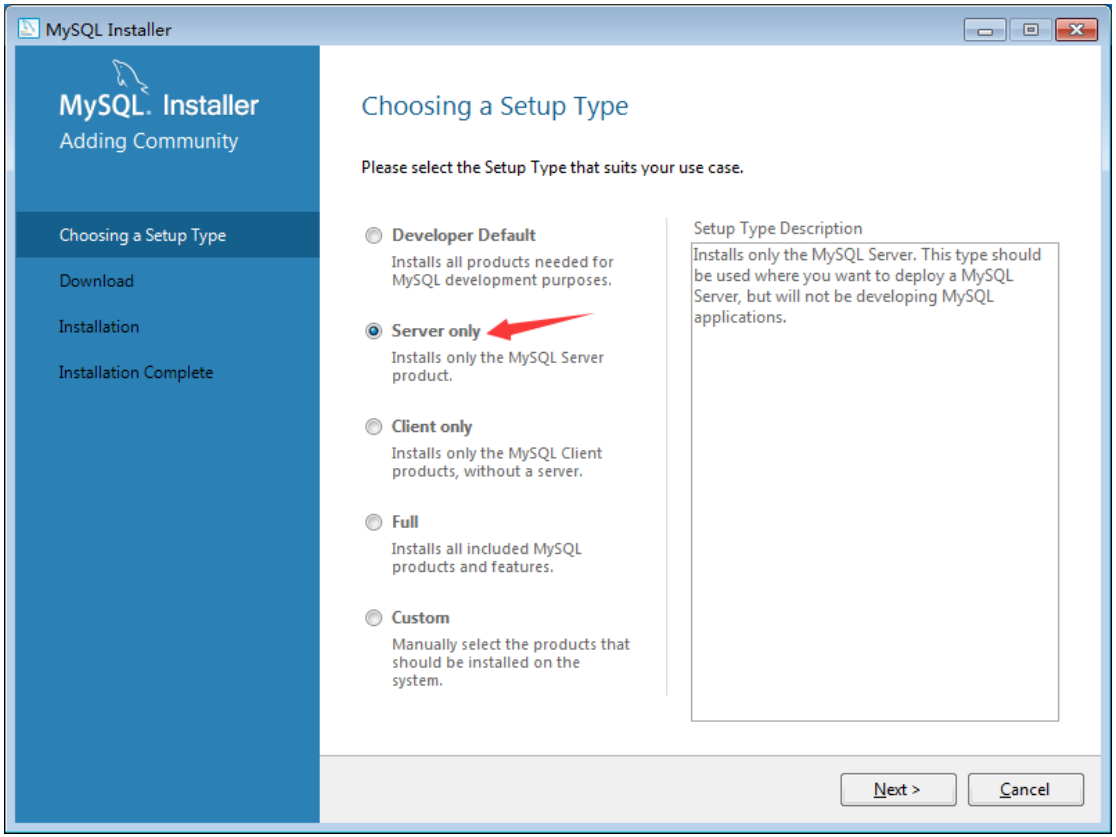


图2-10 选择安装类型界面

2.3.3安装MySQL

在安装MySQL过程中，如果提示需要安装“Microsoft Visual C++ 2015-2019 Redistributable (x64) – 14.28.29325”时，选择同意安装即可（如图2-11所示）。

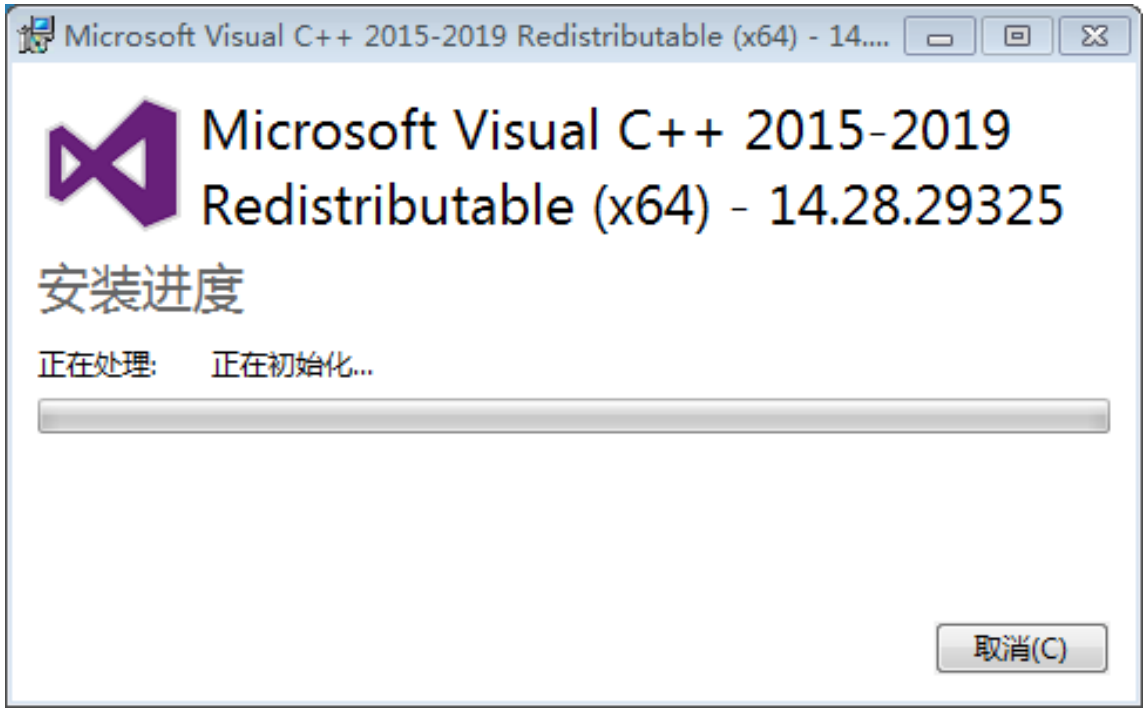


图2-11 安装过程中显示的界面

2.3.3安装MySQL

安装完成以后，MySQL数据库的后台服务进程已经被自动启动，这时就需要使用一个客户端工具来操作MySQL数据库，我们可以使用MySQL安装时自带的命令行界面作为客户端工具来操作数据库。具体方法是，在Windows7的开始菜单中点击“MySQL 8.0 Command Line Client”图标，然后输入数据库密码（这个密码是在安装MySQL的过程中用户自己设置的），就会出现如图2-12所示界面。可以在命令提示符“mysql>”后面输入SQL语句来执行数据库的各种操作。

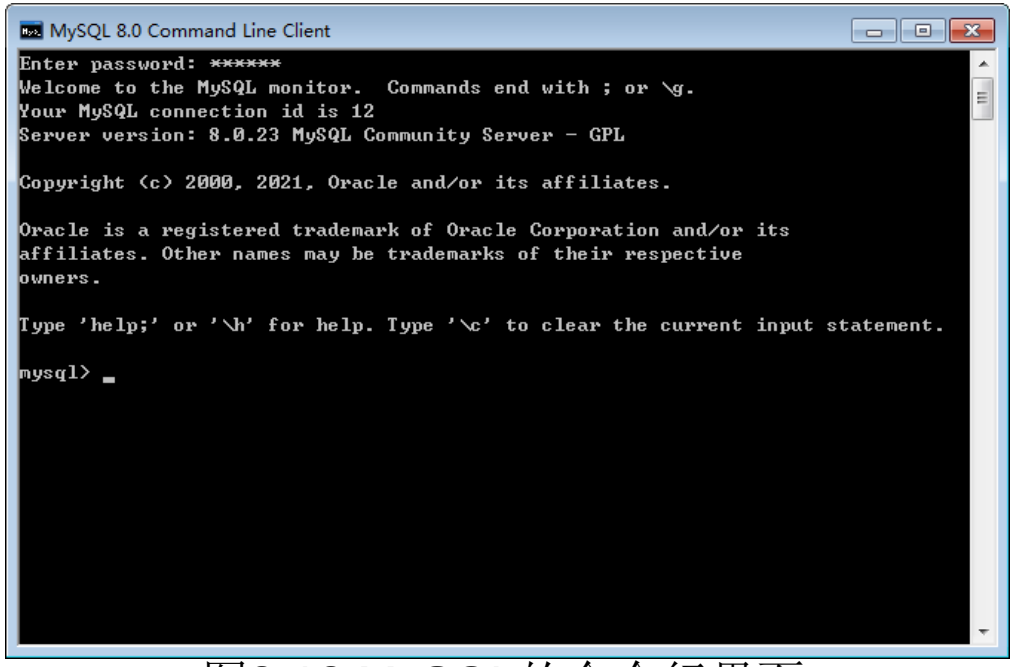


图2-12 MySQL的命令行界面

2.3.3安装MySQL

需要说明的是，MySQL数据库服务进程启动以后，会占用一定的系统资源。实际上，我们平时在计算机上很少使用MySQL数据库，因此，为了减少对系统资源的占用，没有必要每次开机都自动启动MySQL数据库后台服务进程，可以设置为“手动”启动服务进程，这样，只有当需要用到MySQL数据库时，再去手动启动即可。这里以Windows7操作系统为例介绍如何把MySQL数据库服务设置为“手动”启动。

2.3.3安装MySQL

在Windows系统桌面上的“计算机”图标上单击鼠标右键，在弹出的菜单中点击“管理”，在出现的计算机管理界面（如图2-13所示）中，在左侧栏目中点击“服务”，在右侧栏目中会出现很多服务进程，其中，就可以找到名称为“MySQL80”的服务进程，可以看到，该服务进程的状态为“已启动”，启动类型为“自动”。

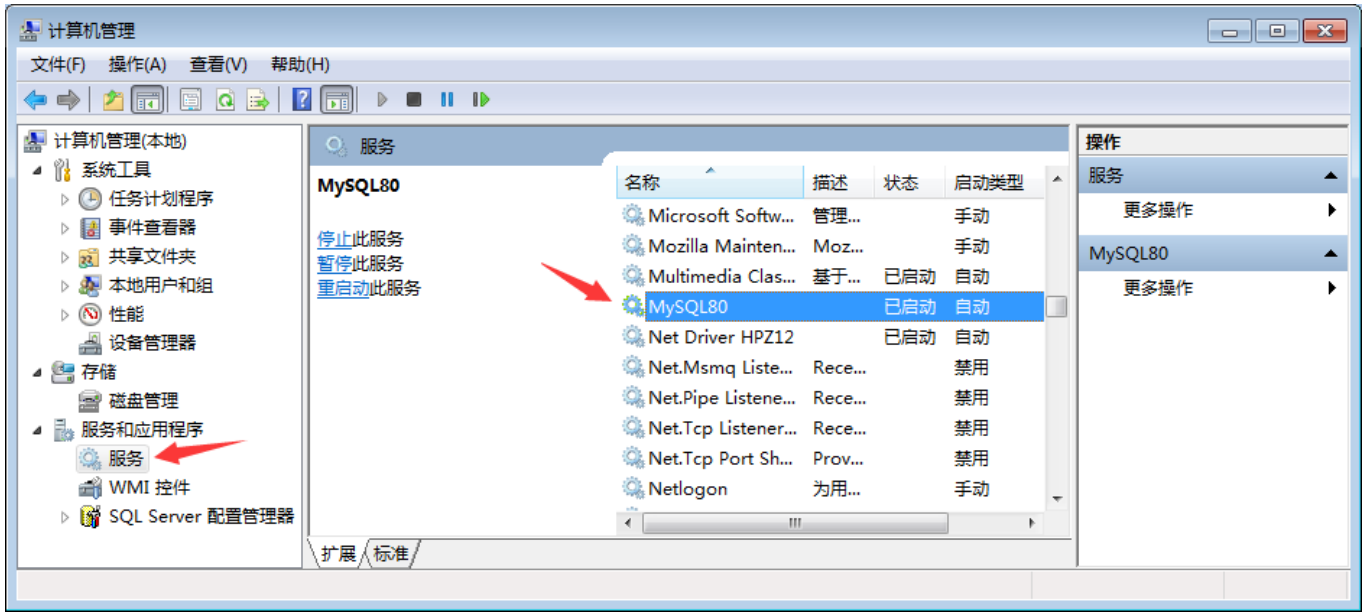


图2-13 计算机管理界面

2.3.3安装MySQL

在“MySQL80”这一行上单击鼠标右键，在弹出的菜单中点击“属性”，会弹出如图2-14所示的界面，在这个界面中，在“服务状态”下面有三个按钮，即“启动”、“停止”和“暂停”，分别用来启动、停止和暂停MySQL服务进程。为了修改启动类型，可以在“启动类型”右侧的下拉列表中选择“手动”，最后单击“确定”按钮即可。

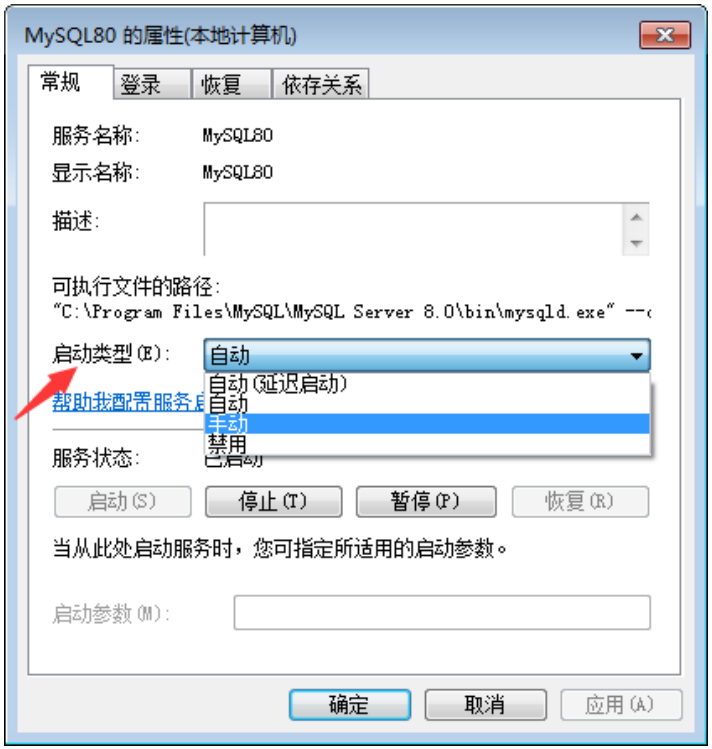


图2-14 MySQL启动类型设置界面

2.3.4 MySQL数据库的使用方法

这里给出一个综合实例来演示MySQL数据库的用法。需要创建一个管理学生信息的数据库，并把表2-5中的数据填充到数据库中，完成相关的数据库操作。

表2-5 学生表

学号	姓名	性别	年龄
95001	王小明	男	21
95002	张梅梅	女	20

2.3.4 MySQL数据库的使用方法

打开MySQL数据库的命令行界面，输入如下SQL语句创建数据库school：
mysql> CREATE DATABASE school;

需要注意的是，SQL语句中可以不用区分字母大小写。
可以使用如下SQL语句查看已经创建的所有数据库：
mysql> SHOW DATABASES;

创建好数据库school以后，可以使用如下SQL语句打开数据库：
mysql> USE school;

2.3.4 MySQL数据库的使用方法

使用如下SQL语句创建一个表student：

```
mysql>CREATE TABLE student(
    -> sno char(5),
    -> sname char(10),
    -> ssex char(2),
    -> sage int);
```

使用如下SQL语句查看已经创建的表：

```
mysql> SHOW TABLES;
```

2.3.4 MySQL数据库的使用方法

使用如下SQL语句向student表中插入两条记录：
mysql> INSERT INTO student VALUES('95001','王小明','男',21);
mysql> INSERT INTO student VALUES('95002','张梅梅','女',20);

使用如下SQL语句查询student表中的记录：
mysql> SELECT * FROM student

使用如下SQL语句修改表中的数据：
mysql> UPDATE student SET age =21 WHERE sno='95001';

使用如下SQL语句删除student表：
mysql> DROP TABLE student;

使用如下SQL语句查询数据库中还存在哪些表：
mysql> SHOW TABLES;
使用如下SQL语句删除数据库school：
mysql> DROP DATABASE school;
使用如下SQL语句查询系统中还存在哪些数据库：
mysql> SHOW DATABASES;

2.3.5使用Python操作MySQL数据库

使用Python操作MySQL数据库之前，需要安装PyMySQL，它是Python中操作MySQL的模块。在Windows操作系统的cmd中运行如下命令安装PyMySQL：

```
> pip install PyMySQL
```

1. 连接数据库

首先打开MySQL数据库的命令行界面，在MySQL数据库中创建一个名称为school的数据库（如果已经存在该数据库，则需要先删除再创建），然后，编写如下代码发起对数据库的连接：

```
1.  # mysql1.py
2.  import pymysql.cursors
3.  # 连接数据库
4.  connect = pymysql.Connect(
5.      host='localhost', # 主机名
6.      port=3306, # 端口号
7.      user='root', # 数据库用户名
8.      passwd='123456', # 密码
9.      db='school', # 数据库名称
10.     charset='utf8' #编码格式
11. )
12. # 获取游标
13. cursor = connect.cursor()
14. # 执行SQL查询
15. cursor.execute("SELECT VERSION()")
16. # 获取单条数据
17. version = cursor.fetchone()
18. # 打印输出
19. print("MySQL数据库版本是：%s" % version)
20. # 关闭数据库连接
21. connect.close()
```

上面代码的执行结果如下：
MySQL数据库版本是：8.0.23

2.创建表

在school数据库中创建一个表student，具体代码如下：

```
1.  # mysql2.py
2.  import pymysql.cursors
3.  # 连接数据库
4.  connect = pymysql.Connect(
5.      host='localhost',
6.      port=3306,
7.      user='root',
8.      passwd='123456'
9.      db='school',
10.     charset='utf8'
11. )
12. # 获取游标
13. cursor = connect.cursor()
14. # 如果表存在，则先删除
15. cursor.execute("DROP TABLE IF EXISTS student")
16. # 设定SQL语句
17. sql = """
18. CREATE TABLE student(
19.     sno char(5),
20.     sname char(10),
21.     ssex char(2),
22.     sage int);
23. """
24. # 执行SQL语句
25. cursor.execute(sql)
26. # 关闭数据库连接
27. connect.close()
```

3. 插入数据

把表2-5中的两条数据插入到student表中，具体代码如下：

```
1.  # mysql3.py
2.  import pymysql.cursors
3.  # 连接数据库
4.  connect = pymysql.Connect(
5.      host='localhost',
6.      port=3306,
7.      user='root',
8.      passwd='123456',
9.      db='school',
10.     charset='utf8'
11. )
12. # 获取游标
13. cursor = connect.cursor()
14. # 插入数据
15. sql = "INSERT INTO student(sno,sname,ssex,sage) VALUES ('%s', '%s', '%s', %d)"
16. data1 = ('95001','王小明','男',21)
17. data2 = ('95002','张梅梅','女',20)
18. cursor.execute(sql % data1)
19. cursor.execute(sql % data2)
20. connect.commit()
21. print('成功插入数据')
22. # 关闭数据库连接
23. connect.close()
```

4. 修改数据

把学号为“95002”的学生的年龄修改为21岁，具体代码如下：

```
# mysql4.py
import pymysql.cursors
# 连接数据库
connect = pymysql.Connect(
    host='localhost',
    port=3306,
    user='root',
    passwd='123456',
    db='school',
    charset='utf8'
)
# 获取游标
cursor = connect.cursor()
# 修改数据
sql = "UPDATE student SET sage = %d WHERE sno = '%s' "
data = (21, '95002')
cursor.execute(sql % data)
connect.commit()
print('成功修改数据')
# 关闭数据库连接
connect.close()
```

5. 查询数据

找出学号为“95001”的学生的具体信息，具体代码如下：

```
1.  # mysql5.py
2.  import pymysql.cursors
3.  # 连接数据库
4.  connect = pymysql.Connect(
5.      host='localhost',
6.      port=3306,
7.      user='root',
8.      passwd='123456',
9.      db='school',
10.     charset='utf8'
11. )
12. # 获取游标
13. cursor = connect.cursor()
14. # 查询数据
15. sql = "SELECT sno,sname,ssex,sage FROM student WHERE sno = '%s'
    "
16. data = ('95001',)   #元组中只有一个元素的时候需要加一个逗号
17. cursor.execute(sql % data)
18. for row in cursor.fetchall():
19.     print("学号:%s\t姓名:%s\t性别:%s\t年龄:%d" % row)
20. print('共查找出', cursor.rowcount, '条数据')
21. # 关闭数据库连接
22. connect.close()
```

6. 删除数据

删除学号为“95002”的学生记录，具体代码如下：

```
# mysql6.py
import pymysql.cursors
# 连接数据库
connect = pymysql.Connect(
    host='localhost',
    port=3306,
    user='root',
    passwd='123456',
    db='school',
    charset='utf8'
)
# 获取游标
cursor = connect.cursor()
# 删除数据
sql = "DELETE FROM student WHERE sno = '%s'"
data = ('95002',)  #元组中只有一个元素的时候需要加一个逗号
cursor.execute(sql % data)
connect.commit()
print('成功删除', cursor.rowcount, '条数据')
# 关闭数据库连接
connect.close()
```

2.4 Hadoop的安装和使用

2.4.1 Hadoop简介
2.4.2 分布式文件系统HDFS
2.4.3 Hadoop的安装
2.4.4 HDFS的基本使用方法

2.4.1 Hadoop简介

Hadoop是一个能够对大量数据进行分布式处理的软件框架，并且是以一种可靠、高效、可伸缩的方式进行处理的，它具有以下几个方面的特性。

- 高可靠性。
- 高效性。
- 高可扩展性。
- 高容错性。
- 成本低。
- 运行在Linux平台上。
- 支持多种编程语言。

2.4.2 分布式文件系统HDFS

1. HDFS简介

Hadoop分布式文件系统（Hadoop Distributed File System，HDFS）是Hadoop项目的两大核心之一，是针对谷歌文件系统（Google File System，GFS）的开源实现。
总体而言，HDFS要实现以下目标：
•兼容廉价的硬件设备。
•流数据读写。
•大数据集。
•简单的文件模型。
•强大的跨平台兼容性。

2.4.2 分布式文件系统HDFS

2.HDFS体系结构

HDFS采用了主从（Master/Slave）结构模型，一个HDFS集群包括一个名称节点和若干个数据节点（见图2-15）。

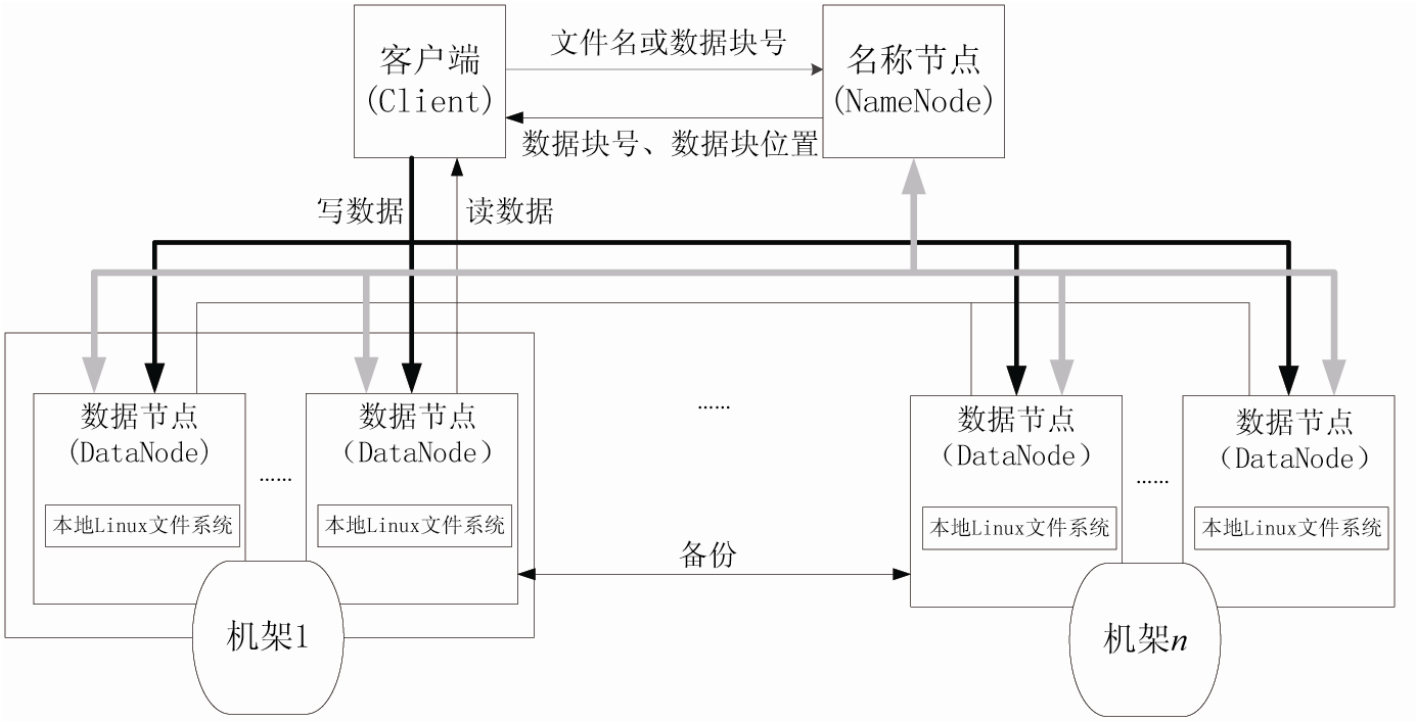


图2-15 HDFS的体系结构

2.4.3 Hadoop的安装

Hadoop包含了HDFS和MapReduce两大核心组件，本教程主要使用HDFS，没有使用MapReduce，但是，仍然要完整地安装Hadoop。这里采用的Apache Hadoop版本是3.1.3。

Hadoop包括三种安装模式：

- 单机模式：只在一台机器上运行，存储是采用本地文件系统，没有采用分布式文件系统HDFS；
- 伪分布式模式：存储采用分布式文件系统HDFS，但是，HDFS的名称节点和数据节点都在同一台机器上；
- 分布式模式：存储采用分布式文件系统HDFS，而且，HDFS的名称节点和数据节点位于不同机器上。

这里介绍Hadoop伪分布式模式的安装方法。

2.4.3 Hadoop的安装

到Hadoop官网（https://archive.apache.org/dist/hadoop/common/hadoop-3.1.3/）下载Hadoop3.1.3安装文件hadoop-3.1.3.tar.gz。

由于Hadoop不直接支持Windows系统，因此，需要使用工具集winutils进行支持。到github.com网站（https://github.com/s911415/apache-hadoop-3.1.3-winutils）下载与Hadoop3.1.3配套的winutils。进入下载页面后，如图2-16所示，点击“Code”按钮，然后在弹出的菜单中点击“Download ZIP”即可下载得到压缩文件apache-hadoop-3.1.3-winutils-master.zip，再将该压缩文件进行解压缩。

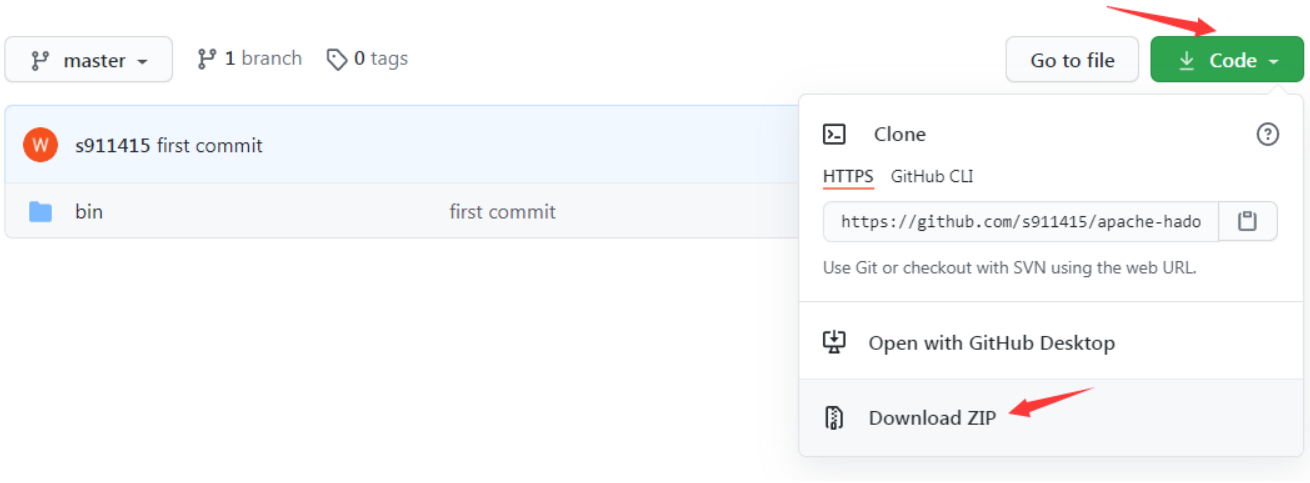


图2-16 winutils的下载页面

2.4.3 Hadoop的安装

把Hadoop3.1.3安装文件hadoop-3.1.3.tar.gz解压缩到“C:\”（或者其他目录），使用winutils中的bin目录整个替换Hadoop中的bin目录。

在“C:\ hadoop-3.1.3”目录下新建tmp目录，再在tmp目录下新建两个子目录，分别是datanode和namenode。

2.4.3 Hadoop的安装

安装完成后需要设置Path环境变量，点击“我的电脑”->“高级系统设置”->“环境变量”，然后，在系统变量中新增一个名称为“HADOOP_HOME”的变量，设置其值为“C:\hadoop-3.1.3”（如图2-17所示）。

图2-17 新建系统变量HADOOP_HOME

2.4.3 Hadoop的安装

然后，再在用户变量Path中加入如下信息：
%HADOOP_HOME%\bin
注意，新增加的路径和Path中原来已有的路径之间要用英文分号隔开（如图2-18所示）。

图2-18 编辑用户变量Path

2.4.3 Hadoop的安装

对“C:\ hadoop-3.1.3\etc\hadoop”下面的3个配置文件进行修改。

把core-site.xml文件的配置修改为如下：

```
<configuration>
    <property>
        <name>fs.default.name</name>
        <value>hdfs://localhost:9000</value>
    </property>
</configuration>
```

2.4.3 Hadoop的安装

把hdfs-site.xml文件的配置修改为如下：

```
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>
    <property>
     <name>dfs.permissions</name>
     <value>false</value>
    </property>
   <property>
        <name>dfs.namenode.name.dir</name>
        <value>/C:/hadoop-3.1.3/tmp/namenode</value>
   </property>
   <property>
        <name>dfs.datanode.data.dir</name>
        <value>/C:/hadoop-3.1.3/tmp/datanode</value>
   </property>
</configuration>
```

2.4.3 Hadoop的安装

修改hadoop-env.cmd文件，找到如下一行：
set JAVA_HOME=%JAVA_HOME%

把%JAVA_HOME%替换成JDK的绝对路径，比如：
set JAVA_HOME=C:\ Java\jdk1.8.0_111

需要注意的是，如果JDK路径中包含了空格，如果直接使用如下设置后面步骤会报错：
set JAVA_HOME= C:\Program Files\Java\jdk1.8.0_111

如果采用这种带有空格的路径，后面运行“hdfs namenode -format”命令时就会报错，因为Program Files中存在空格。为了解决这个问题，可以使用下面两种方式之一进行处理：
（1）只需要用PROGRA~1 代替Program Files，即改为
C:\PROGRA~1\Java\jdk1.8.0_111
（2）或是使用双引号，即改为 “C:\Program Files”\Java\jdk1.8.0_111

2.4.3 Hadoop的安装

然后，在Windows系统中打开一个cmd窗口，执行如下命令对Hadoop系统进行格式化：

> cd c:\hadoop-3.1.3\bin

> hdfs namenode -format

上述命令执行以后，如果返回类似如下的信息则表示格式化成功：

\hadoop-3.1.3\tmp\namenode has been successfully formatted.

执行如下命令启动

> cd c:\hadoop-3.1.3\sbin

> start-dfs.cmd

执行该命令以后，会同时弹出另外2个cmd窗口，这2个新弹出的cmd窗口不要关闭，然后，在刚才执行start-dfs.cmd命令的cmd窗口内，继续执行JDK自带的命令jps查看Hadoop已经启动的进程：

> jps

2.4.3 Hadoop的安装

需要注意的是，这里在使用jps命令的时候，没有带上绝对路径，是因为已经把JDK添加到了Path环境变量中。
执行jps命令以后，如果能够看到“DataNode”和“NameNode”这两个进程，就说明Hadoop启动成功。
需要关闭Hadoop时，可以执行如下命令：

```
> cd c:\hadoop-3.1.3\sbin
> stop-dfs.cmd
```

2.4.4 HDFS的基本使用方法

1.使用WEB管理页面操作HDFS

首先启动Hadoop，然后可以在浏览器中输入“http://localhost:9870”，就可以访问Hadoop的WEB管理页面（如图2-19所示）。

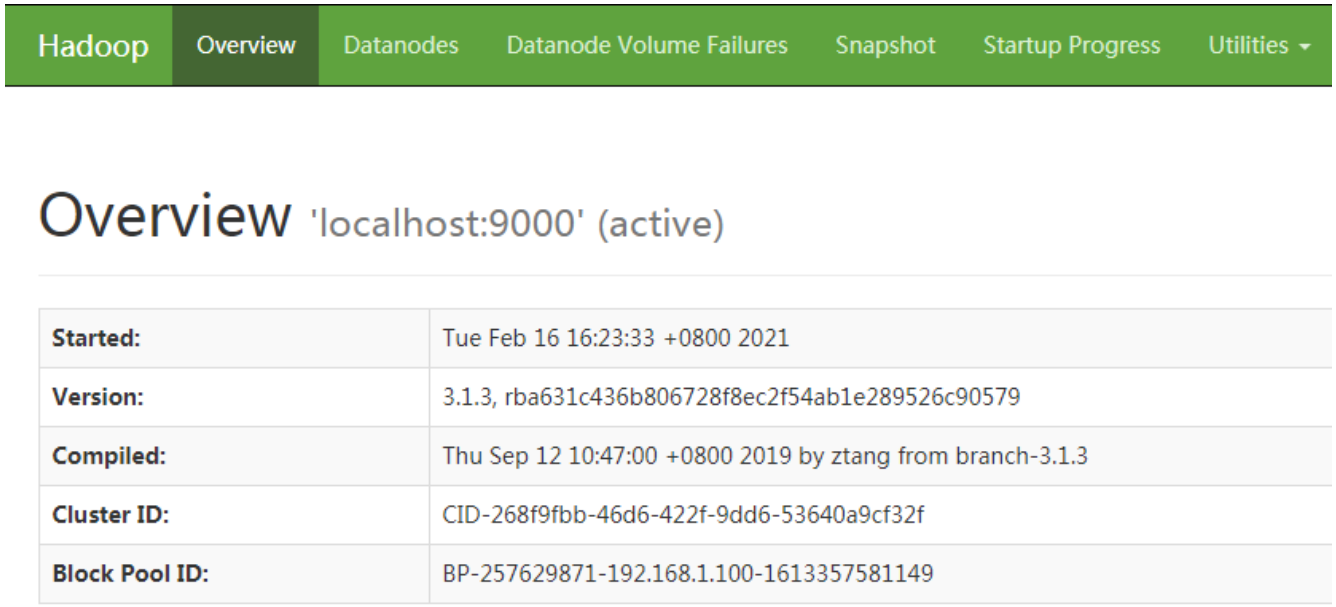


图2-19 Hadoop的WEB管理页面

2.4.4 HDFS的基本使用方法

在WEB管理页面中，点击顶部右侧的菜单选项“Utilities”，在弹出的子菜单中点击“Browse the file system”，会出现如图2-20所示的HDFS文件系统操作页面，在这个页面中可以创建、查看、删除目录和文件。

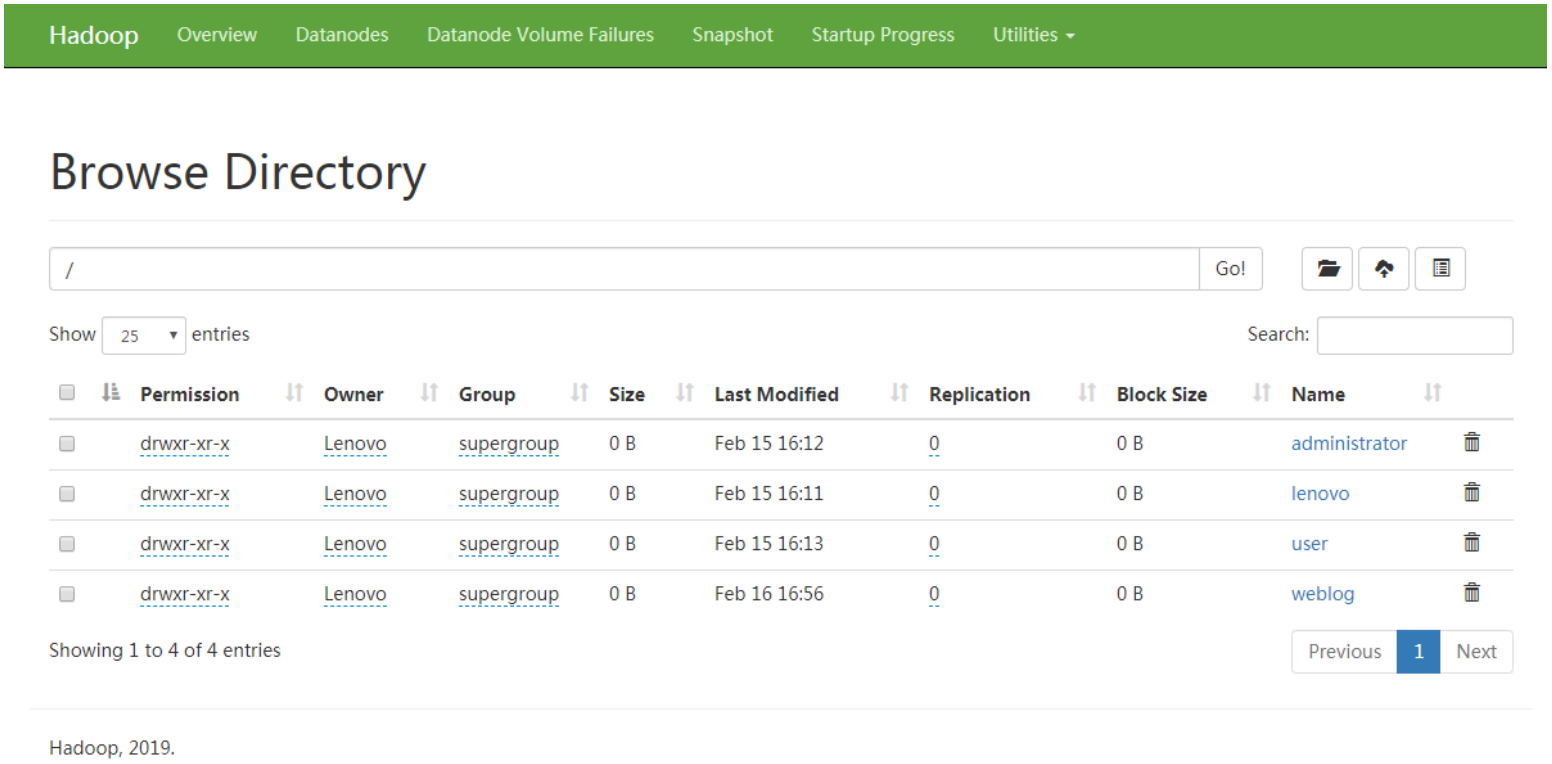


图2-20 HDFS文件系统操作页面

2.4.4 HDFS的基本使用方法

2.使用命令操作HDFS

除了在浏览器中通过WEB方式操作HDFS以外，还可以在cmd窗口中使用命令对HDFS进行操作。
首先，创建一个名称为“user”的目录，命令如下：
> cd c:\hadoop-3.1.3\bin
> hadoop fs -mkdir hdfs://localhost:9000/user/
➢hadoop fs -mkdir hdfs://localhost:9000/user/xiaoming

然后，在“C:\”下创建一个文件test.txt，里面输入一行语句“I love hadoop”，使用如下命令把该文件上传到HDFS中：
> hadoop fs -put C:\test.txt hdfs://localhost:9000/user/xiaoming

2.4.4 HDFS的基本使用方法

使用如下命令查看HDFS中的目录和文件：

➢hadoop fs -ls hdfs://localhost:9000/user/xiaoming

使用如下命令把HDFS中的文件内容显示到本地屏幕上：

➢hadoop fs -cat hdfs://localhost:9000/user/xiaoming/test.txt

把上面的HDFS中的文件test.txt下载到本地文件系统，并重命名为test1.txt：

➢hadoop fs -get hdfs://localhost:9000/user/xiaoming/test.txt C:\test1.txt

使用如下命令删除HDFS中的一个文件：

> hadoop fs -rm hdfs://localhost:9000/user/xiaoming/test.txt

使用如下命令删除HDFS中的一个目录及其下面的文件：

> hadoop fs -rm -r hdfs://localhost:9000/user/xiaoming

2.5 本章小结

本章内容介绍了大数据实验环境（包括Python、JDK、MySQL、Hadoop等）的搭建方法，这是在后面章节中开展实践操作的基础。这里需要说明的是，对于实际的企业大数据应用场景而言，一般都是采用Linux系统，为了让教学环境更加贴近企业实际生产环境，教学环节按道理应当首选Linux系统。但是，由于各种大数据软件在Linux系统下安装会比较繁琐，这就会给学习者带来很大障碍，直接导致很多实验无法顺利开展。因此，为了让后续章节的实验环节能够顺利开展，本教程使用Windows系统来搭建大数据实验环境。

附录A：主讲教师林子雨简介

主讲教师：林子雨

单位：厦门大学计算机科学与技术系
E-mail: ziyulin@xmu.edu.cn
个人网页：http://dblab.xmu.edu.cn/linziyu
数据库实验室网站：http://dblab.xmu.edu.cn

扫一扫访问个人主页

林子雨，男，1978年出生，博士（毕业于北京大学），全国高校知名大数据教师，现为厦门大学计算机科学系副教授，厦门大学信息学院实验教学中心主任，曾任厦门大学信息科学与技术学院院长助理、晋江市发展和改革局副局长。中国计算机学会数据库专业委员会委员，中国计算机学会信息系统专业委员会委员。国内高校首个“数字教师”提出者和建设者，厦门大学数据库实验室负责人，厦门大学云计算与大数据研究中心主要建设者和骨干成员，2013年度、2017年度和2020年度厦门大学教学类奖教金获得者，荣获2019年福建省精品在线开放课程、2018年厦门大学高等教育成果特等奖、2018年福建省高等教育教学成果二等奖、2018年国家精品在线开放课程。主要研究方向为数据库、数据仓库、数据挖掘、大数据、云计算和物联网，并以第一作者身份在《软件学报》《计算机学报》和《计算机研究与发展》等国家重点期刊以及国际学术会议上发表多篇学术论文。作为项目负责人主持的科研项目包括1项国家自然科学青年基金项目(No.61303004)、1项福建省自然科学青年基金项目(No.2013J05099)和1项中央高校基本科研业务费项目(No.2011121049)，主持的教改课题包括1项2016年福建省教改课题和1项2016年教育部产学协作育人项目，同时，作为课题负责人完成了国家发改委城市信息化重大课题、国家物联网重大应用示范工程区域试点泉州市工作方案、2015泉州市互联网经济调研等课题。中国高校首个“数字教师”提出者和建设者，2009年至今，“数字教师”大平台累计向网络免费发布超过1000万字高价值的研究和教学资料，累计网络访问量超过1000万次。打造了中国高校大数据教学知名品牌，编著出版了中国高校第一本系统介绍大数据知识的专业教材《大数据技术原理与应用》，并成为京东、当当网等网店畅销书籍；建设了国内高校首个大数据课程公共服务平台，为教师教学和学生学习大数据课程提供全方位、一站式服务，年访问量超过400万次，累计访问量超过1500万次。

附录B：大数据学习路线图

大数据学习路线图访问地址：http://dblab.xmu.edu.cn/post/10164/

附录C：林子雨大数据系列教材

了解全部教材信息：http://dblab.xmu.edu.cn/post/bigdatabook/

附录D：《大数据导论（通识课版）》教材

开设全校公共选修课的优质教材

本课程旨在实现以下几个培养目标：

- 引导学生步入大数据时代，积极投身大数据的变革浪潮之中
- 了解大数据概念，培养大数据思维，养成数据安全意识
- 认识大数据伦理，努力使自己的行为符合大数据伦理规范要求
- 熟悉大数据应用，探寻大数据与自己专业的应用结合点
- 激发学生基于大数据的创新创业热情

高等教育出版社 ISBN:978-7-04-053577-8 定价：32元 版次：2020年2月第1版
教材官网：http://dblab.xmu.edu.cn/post/bigdataintroduction/

附录E：《大数据导论》教材

- 林子雨 编著《大数据导论》
- 人民邮电出版社，2020年9月第1版
- ISBN:978-7-115-54446-9 定价：49.80元

教材官网：http://dblab.xmu.edu.cn/post/bigdata-introduction/

开设大数据专业导论课的优质教材

扫一扫访问教材官网

附录F：《大数据技术原理与应用（第3版）》教材

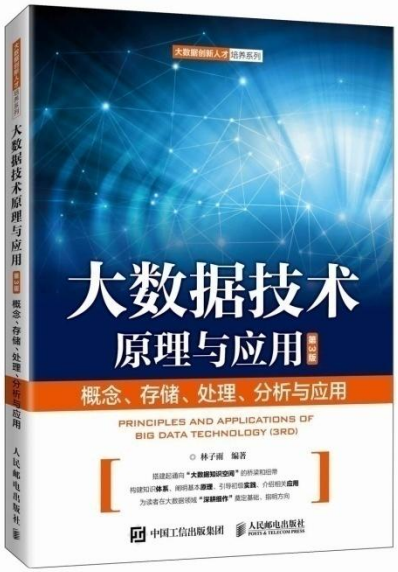
《大数据技术原理与应用——概念、存储、处理、分析与应用（第3版）》，由厦门大学计算机科学系林子雨博士编著，是国内高校第一本系统介绍大数据知识的专业教材。人民邮电出版社　ISBN:978-7-115-54405-6　定价：59.80元

全书共有17章，系统地论述了大数据的基本概念、大数据处理架构Hadoop、分布式文件系统HDFS、分布式数据 库HBase、NoSQL数据库、云数据库、分布式并行编程模型MapReduce、Spark、流计算、Flink、图计算、数据可视化以及大数据在互联网、生物医学和物流等各个领域的应用。在Hadoop、HDFS、HBase、MapReduce、Spark和Flink等重要章节，安排了入门级的实践操作，让读者更好地学习和掌握大数据关键技术。

本书可以作为高等院校计算机专业、信息管理等相关专业的大数据课程教材，也可供相关技术人员参考、学习、培训之用。

欢迎访问《大数据技术原理与应用——概念、存储、处理、分析与应用》教材官方网站：
http://dblab.xmu.edu.cn/post/bigdata3

扫一扫访问教材官网



附录G：《大数据基础编程、实验和案例教程（第2版）》

本书是与《大数据技术原理与应用（第3版）》教材配套的唯一指定实验指导书

大数据教材

1+1黄金组合

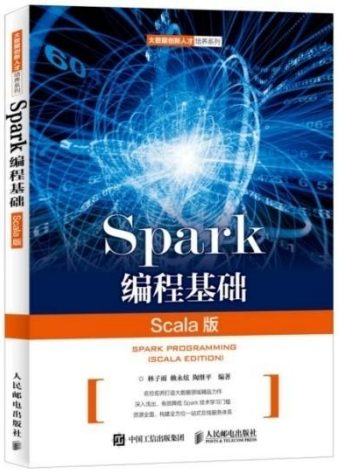
厦门大学林子雨编著

配套实验指导书

- 步步引导，循序渐进，详尽的安装指南为顺利搭建大数据实验环境铺平道路
- 深入浅出，去粗取精，丰富的代码实例帮助快速掌握大数据基础编程方法
- 精心设计，巧妙融合，八套大数据实验题目促进理论与编程知识的消化和吸收
- 结合理论，联系实际，大数据课程综合实验案例精彩呈现大数据分析全流程

林子雨编著《大数据基础编程、实验和案例教程（第2版）》
清华大学出版社 ISBN:978-7-302-55977-1 定价：69元 2020年10月第2版

附录H：《Spark编程基础（Scala版）》



《Spark编程基础（Scala版）》

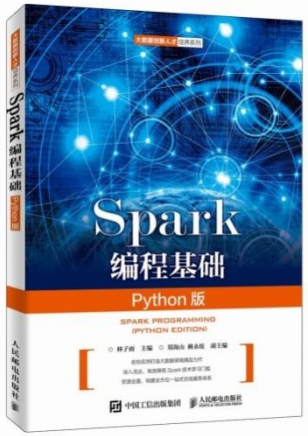
厦门大学 林子雨，赖永炫，陶继平 编著

披荆斩棘，在大数据丛林中开辟学习捷径
填沟削坎，为快速学习Spark技术铺平道路
深入浅出，有效降低Spark技术学习门槛
资源全面，构建全方位一站式在线服务体系

人民邮电出版社出版发行，ISBN:978-7-115-48816-9
教材官网：http://dblab.xmu.edu.cn/post/spark/

本书以Scala作为开发Spark应用程序的编程语言，系统介绍了Spark编程的基础知识。全书共8章，内容包括大数据技术概述、Scala语言基础、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、授课视频、上机实验指南等。

附录I：《Spark编程基础（Python版）》



《Spark编程基础（Python版）》

厦门大学 林子雨，郑海山，赖永炫 编著

披荆斩棘，在大数据丛林中开辟学习捷径
填沟削坎，为快速学习Spark技术铺平道路
深入浅出，有效降低Spark技术学习门槛
资源全面，构建全方位一站式在线服务体系

人民邮电出版社出版发行，ISBN:978-7-115-52439-3

教材官网： http://dblab.xmu.edu.cn/post/spark-python/

本书以Python作为开发Spark应用程序的编程语言，系统介绍了Spark编程的基础知识。全书共8章，内容包括大数据技术概述、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Structured Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、上机实验指南等。

附录J：高校大数据课程公共服务平台

高校大数据课程

公 共 服 务 平 台

http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/

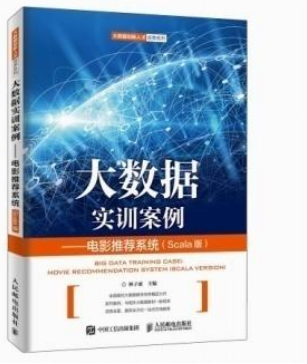
扫一扫访问平台主页

扫一扫观看3分钟FLASH动画宣传片

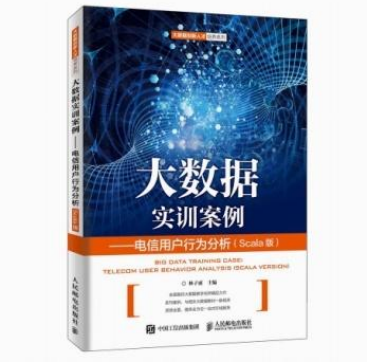
附录K：高校大数据实训课程系列案例教材

为了更好满足高校开设大数据实训课程的教材需求，厦门大学数据库实验室林子雨老师团队联合企业共同开发了《高校大数据实训课程系列案例》，目前已经完成开发的系列案例包括：

《电影推荐系统》（已经于2019年5月出版）
《电信用户行为分析》（已经于2019年5月出版）
《实时日志流处理分析》
《微博用户情感分析》
《互联网广告预测分析》
《网站日志处理分析》



系列案例教材将于2019年陆续出版发行，教材相关信息，敬请关注网页后续更新！
http://dblab.xmu.edu.cn/post/shixunkecheng/



扫一扫访问大数据实训课程系列案例教材主页


Thank You!