1.7.数据科学的常用工具

▲ 1.6.数据科学的人才类型

▼ 1.8.数据科学的相关应用

1 数据科学家的常用工具

Python、R、Scala、Clojure、Haskell 等数据科学语言工具

HBase、MongoDB、Couchbase、Cassandra 等 NoSQL 工具

Web Scraper、Flume Avro、Sqoop、Hume 等支持数据采集、聚合或传递的工具

SQL、RDMS、DW、OLAP 等传统数据库和数据仓库工具

Weka、KNIME、RapidMiner、SciPy、Pandas 等支持数据挖掘的工具

Hadoop、HDFS、MapReduce、Spark、Storm 等支持大数据计算的工具

ggplot2、Tableau、D3.js、Shiny、Flare、Gephi 等支持数据可视化的工具

HBase、Pig、Hive、Impala、Cascalog 等支持大数据管理、存储和查询的工具

SAS、SPSS、Matlab 等数据统计分析工具

（1）Python 有关的知识

一种解释性、交互式、动态类型的语言
“优雅”“明确”“简单”

基于 Python 进行数据科学研究的优点

- 用 Python 编写的源代码的代码量少，且易于编写、阅读、理解和维护。
- Python 中可用于数据科学的第三方扩展包的数量多、功能强，据 Python 第三方扩展包官网介绍，Python 第三方扩展包项目已超过 20 万。
- Python 是一种解释性语言，因此能较好地支持数据科学中的交互式分析任务。

Python 在数据科学应用中的缺点

- Python 是一种解释型语言，所以运行速度慢。
- Python 代码不能加密，因此安全性较低。

（2）Python 和 R 的对比分析

表 1-1 Python 与 R 的主要区别与联系

	Python	R
设计者	计算机科学家吉多·范·罗瑟（Guido Van Rossum）	统计学家罗斯·艾卡（Ross Ihaka）和 罗伯特·金特尔曼（Robert Gentleman
设计目的	提高软件开发的效率与源代码的可读性	方便统计处理、数据分析及图形化显示
设计哲学	（源代码层次上）优雅、明确、简单	（功能层次上）简单、有效、完善
发行年	1991	1995
前身	ABC 语言、C 语言和 Modula-3	S 语言
主要维护者	Python Software Foundation（Python 软件基金会）	The R-Core Team（R-核心团队） The R Foundation（R 基金会）
主要用户群	软件工程师/程序员	学术/科学研究/统计学家

（2）Python 和 R 的对比分析

表 1-1 Python 与 R 的主要区别与联系

	Python	R
可用性	源代码的语法更规范，便于编码与调试	可以用简单几行代码实现复杂的数据统计、机器学习和数据可视化功能
学习成本曲线	入门相对容易，入门后学习难度随着学习内容逐步提高	入门难，入门后相对容易
第三方提供的功能	以“包”的形式存在 可从 PyPI 下载	以“库”的形式存在可从CRAN下载
常用包/库	数据处理：pandas 科学计算：SciPy、NumPy 可视化：matplotlib 统计建模：statsmodels 机器学习：sckikit-learn、TensorFlow、Theano	数据科学工具集：tidyverse 数据处理：dplyr、plyr、data.table、stringr 可视化：ggplot2、ggvis、lattice 机器学习：RWeka、caret

（2）Python 和 R 的对比分析

表 1-1 Python 与 R 的主要区别与联系

	Python	R
常用 IDE（集成开发环境）	Jupyter Notebook（iPython Notebook/Spyder/Rodeo/Eclipse/PyCharm	RStudio、RGui
R 与 Python 之间的相互调用	在 Python 中，可以通过库 RPy2 调用R 代码	在 R 中，可以通过包 rPython 调用 Python代码

（3）Python 和 R 在数据科学中广泛应用的原因

第一层原因—程序员的设计目的

数据科学家如果还是用Java、 C 语言等语言完成数据科学任务，主要精力将消耗在流程控制、数据结构的定义和算法设计上，而难以集中精力去处理数据问题。

第二层原因—第三方扩展包/模块

可以通过 Python 或 R 调用面向数据科学任务的专业级服务—Python 或 R 的第三方扩展包/模块，如数据可视化可以采用 Python 第三方扩展包 Seaborn 或 R 第三方扩展包 ggplot2 可以轻松实现。

第三层原因—主流第三方扩展包/模块的开发者的智慧

主流第三方扩展包、模块的 开发者都是统计学、机器学习等数据科学领域的顶级人才。