数据处理的 9 大编程语言 - 文章

有关大数据的话题一直很火热。伴随着信息的爆炸式增长，大数据渗透到了各行各业，广泛应用于公司中，同时也使得传统的软件比如 Excel 看起来很笨拙。数据分析不再只是书呆子的事，同时其对高复杂性分析、实时处理的需求也比以往更加庞大。

那么筛选海量数据集最优的工具是什么呢？我们咨询了一些数据黑客关于他们在数据分析的核心工作中最喜欢的编程语言和工具包。

R 语言

这份名单如果不以 R 开头，那就是彻头彻尾的疏忽。自 1997 年起，作为一门免费的，可替代 Matlab 或 SAS 等昂贵统计软件的语言，R 被抛弃。

但是在过去的几年中，它却成了数据科学的宠儿—甚至成了统计学家、 华尔街交易员、生物学家和硅谷开发者必不可少的工具。 随着其商业价值的不断增长和传播，诸如谷歌、Facebook、 美国银行和纽约时代周刊都在使用。

R 简单易用。通过 R ，短短几行代码就可以筛选复杂的数据集，通过成熟的模型函数处理数据，制作精美的图表进行数据可视化。简直就是 Excel 的加强灵活版。

R 最大的价值就是围绕其开发的活跃的生态圈： R 社区在持续不断地向现存丰富的函数集增添新的包和特性。据估计 R 的使用者已经超过 200 万人，最近的一项调查也显示 R目前是数据科学领域最受欢迎的语言，大约 61% 的受访者使用 R（第二名是 Python，占比39%）。

在华尔街，R 的使用比例也在不断增长。美国银行副总裁Niall O’Connor 说：“以往，分析员通常是熬夜研究 Excel 文件，但是现在 R 正被逐渐地应用于金融建模，尤其是作为可视化工具。R 促使了表格化分析的出局。”

作为一门数据建模语言， R 正在走向成熟，尽管在公司需要大规模产品的时候 R 能力有限，也有些人说它已经被其他语言替代了。

Metamarkets 公司的 CEO Michael Driscoll 说：“ R 擅长的是勾画，而不是搭建，在 Google 的 page rank 算法和 Facebook 的好友推荐算法实现的核心中是不会有 R 的。工程师会用 R 进行原型设计，再用 Java 或者 Python将其实现。”

Paul Butler 在 2010 年用 R 构建了一个著名的 Facebook 世界地图，证明了 R 在数据可视化上的强大能力。然而他并不经常使用 R。

Butler 说：“由于在处理较大数据集时缓慢且笨拙，R 在行业中已经有些沦为明日黄花了 ”

那么使用什么作为它的替代呢？看下去。

Python

如果 R 是个有点神经质的可爱的极客，那么 Python 就是它容易相处的欢快的表弟。融合了 R 快速成熟的数据挖掘能力以及更实际的产品构建能力， Python 正迅速地获得主流的呼声。 Python 更直观，且比 R 更易学，近几年其整体的生态系统发展也成长得很快，使其在统计分析上的能力超越了之前的 R 语言。

Butler 说：“Python 是行业人员正在转换发展的方向。过去两年里，很明显存在由 R 向 Python 转化的趋势”

在数据处理中，通常存在规模和技巧的权衡，Python 作为一个折中出现了。 IPython notebook 和 NumPy 可以用于轻量工作的处理， 而 Python 则是中级规模数据处理的有力工具。丰富的数据交流社区也是 Python 的优势，它提供了大量的Python 工具包和特性。

美国银行利用 Python 开发新产品以及基础设施接口，同时也用于处理金融数据。O’Donnell 说：“Python 用途宽广且灵活，所以人们蜂拥而至”。

然而， Driscoll 也提到它并不是高性能的语言，偶尔才会用于装配驱动大规模的核心基础设施。

JULIA

最主流的数据科学处理语言包括 R、 Python、 Java、 Matlab和 SAS。但是这些语言仍然存在一些不足之处，而Julia 正是待以观察的新人。

对大规模商用来说， Julia 还是太晦涩了。但在谈到其取代 R 和 Python 领先地位的潜力的时候，数据极客们都会变得很激动。 Julia 是一门高级的，非常快的函数式语言。速度上比 R 快， 可能比 Python 的扩展性更高，且相对易学。

Butler 说：“Julia 正在快速上升。最终将可以用 Julia 完成任何 R 和 Python 可以完成的事”。

如今的问题是 Julia 太“年轻”了。 其数据交流社区仍处在早期发展阶段，在没有足够的包和工具之前是不足以与 R 和 Python 竞争的。

Driscoll 说：“Julia 很年轻，但正在积攒力量而且未来很可观”。

JAVA

在硅谷最大的科技公司里，Java 和基于 Java 的框架构成了其底层的技术骨架。Driscoll 说：“如果深入观察Twitter，Linkedin 或者 Facebook，你会发现 Java 是他们公司数据引擎架构的基础语言”。

Java 并没有 R 和 Python 那样的数据可视化的能力， 同时也不是最好的用于统计模型的语言。但是如果需要进行原型的基础开发和构建大规模系统， Java 往往是最好的选择。

HADOOP 和 HIVE

为了满足数据处理的巨大需求，基于 Java 的工具群涌而现。 作为基于 Java 的框架，Hadoop 在批处理领域成为热点。Hadoop 比其他处理工具速度要慢，但是它非常精确且被广泛的应用于后台分析，它很好的融合了 Hive， 一个运行在 Hadoop 上的基于查询的框架。

SCALA

Scala 是另一个基于 Java的语言，和 Java 很相似，它正在逐渐成长为大规模机器学习或高级算法的工具。它是函数式语言，也能够构建健壮的系统。

Driscoll 说：“Java 就像是直接用钢筋进行搭建， Scala 则像是在处理黏土原材料，可以将其放进窑中烧制成钢筋”。

KAFKA 和 STORM

当需要快速、实时分析时怎么办？Kafka 可以帮助你。它已经发展了大概五年时间，但最近才成为一个流处理的流行框架。

Kafka 诞生于 Linkedin 公司的内部项目，是一个快速查询系统。至于 Kafka 的缺点呢？ 它太快了，实时的操作也导致了自身的错误，且偶尔还会遗失信息。

Driscoll 说：“在精度和速度之间总需要做权衡，所以硅谷所有的大公司一般都双管齐下： 用 kafka 和 Storm 进行实时处理，用 Hadoop 做批处理系统，虽然会慢一点但却十分精确”。

Storm 是另一个用 Scala 写的框架，且它在硅谷以擅长流处理而受到极大的关注。毫无疑问，Twitter， 一个对快速消息处理有着巨大兴趣的公司会收购了 Storm。

荣幸的提到：

MATLAB

MATLAB 已经存在很长时间了，尽管价格昂贵，但它仍在某些特定领域被广泛使用： 机器学习研究、信号处理、图像识别等领域。

OCTAVE

Octave 与 Matlab 非常相似，只不过它是免费的。然而除了信号处理的学术圈之外很少见到使用。

GO

GO 是另外一个获得关注的新手。它由 Google 开发，与 C 有一定渊源，且在构建稳定系统方面与 Java 和 Python 展开了竞争。

> 打赏支持译者翻译更多好文章，谢谢！

Email：bd@Jobbole.com
QQ： 2302462408 （加好友请注明来意）

更多频道

小组 - 好的话题、有启发的回复、值得信赖的圈子
头条 - 分享和发现有价值的内容与观点
相亲 - 为IT单身男女服务的征婚传播平台
资源 - 优秀的工具资源导航
翻译 - 翻译传播优秀的外文文章
文章 - 国内外的精选文章
设计 - UI,网页，交互和用户体验
iOS - 专注iOS技术分享
安卓 - 专注Android技术分享
前端 - JavaScript, HTML5, CSS
Java - 专注Java技术分享
Python - 专注Python技术分享