大数据时代，R 语言已蓄势待发

作者：司马牵牛 编译 编辑： 李健 2012-06-25 12:40 来源：51CTO

【IT168 技术】上世纪60年代，大型机被引入学术领域与企业。从那时至今，统计分析一直存在。

然而，当今系统收集的遥测数据类型变得多种多样，并且为了深入理解，需要对数据进行过滤;同时，开源应用变得越来广受欢迎，这一切都在改变着 R 这一用于统计分析与可视化的语言。R 还有一个别名：统计领域的红帽子。

所有人都喜欢 R 语言，尤其是大数据产品销售商，比如数据仓库与 Hadoop 数据过滤器。部分原因在于，R 作为开源语言吸引了大量的统计学家与定量分析师，由这些聪明人构成的社区能够引领该语言开发。

字母语言的盛宴

对于美国赛仕研究所(SAS Institute)开发的专有工具和大型机时代肇始之初的 SPSS 统计软件，以及它们在分布式计算时代的后继产品，情况并非如此。

正如可将 Linux 视为 Unix 的开源式模仿，R 编程语言大量借鉴了 S 语言。S 语言由贝尔实验室的约翰·钱伯斯(John Chambers)于 1976 年创建，而在此十几年前出现的 SPSS 和 SAS 工具，令人尊敬但价格昂贵。S 语言的出现是对其作出的反击。在很大程度上，S 语言可以看作 VAX 与 Unix 小型计算机时代的产物，而 R 语言是 PC 与 Linxu 时代的果实。

1996 年，罗斯·艾卡(Ross Ihaka)和罗伯特·简特曼(Robert Gentleman)共同创建了R 语言。这两位来自新西兰奥克兰大学的统计学教授现在依然是 R 语言开发团队的核心成员。(顺便指出：S 语言的创建者钱伯斯也是该团队的核心成员。某些用于 S 语言的数据处理线程不做任何更改即可在 R 语言环境中运行，并非巧合。)

R 语言可视为 S 语言的现代化实现。S-PLUS 语言也是如此。一家名为 Insightful 的公司在 2004 年从Lucent Technologies 公司获得 S 语言授权，创建了 S-PLUS。Insightful 公司在 2008 年被 Tibco Software 公司收购。

革命来临

与 S 以及一定程度上的扩展 S-PLUS 不同，R 并非是在象牙塔里闭门造车而编写出了的代码。它是由统计学家与程序员构成的社区的产物，这一社区创建了 2500 多种插件，可处理各种各样的数据，并针对特定数据类型或行业进行相应的统计分析。

根据 Revolution Analytics 公司的评估，在世界各地有 200 多万定量分析师在使用 R 语言。该公司成立于 2007 年，提供了一种 R 语言的并行实现。从创始之初，该公司一直对 R 语言采取核心开源策略，为开源语言包提供支持，同时对 R 语言环境进行扩展，以便能够在计算机集群更好地运行并与 Hadoop 集群进行协作。

时至今日，尚未有人对 SPSS (2009 年 7 月被 IBM 收购)的开源对应物 PSPP 进行商业化，不过，毫无疑问，随着 PSSP 的成熟， 将会看到商业化的那一天到来。

Revolution Analytics 公司在 2008 年从 Intel Capital 获得了一些种子资金，并于 2009 年获得 900 万美元的风险投资，之后该公司开始在其 R Enterprise 产品中推广 R 专有扩展。该公司的这一策略并不仅仅是令 R 语言社区感到满意。从那时起，Revolution Analytics 开始对底层 R 统计引擎进行并行化处理，以便能够在多核/多线程处理与服务器集群上更好

的运行;增加 NoSQL 类格式 XDF，帮助对数据机进行并行化;同时增加对本地 SAS 文件格式以及转化为 XDF 的支持。

不久以前，该公司对其 R 实现进行调整，以便 Hadoop 集群的每个节点都可以对 Hadoop 集群上存储在 Hadoop 分布式文件系统中的数据进行本地 R 分析，并对这些计算的结果进行整合，类似 MapReduce 对非结构化数据的操作。

过去几年里，Revolution Analytics 公司从 R 社区里获得大量的营养。不过，其他公司也在做一些有趣的事情，将 R 工具集成至其自身的产品中，令从巨量数据中寻求答案的分析师的工作变得更加方便。

并行世界

Netezza 公司在2010 年 2 月开放 Netezza 软件栈，其目的是为了在数据仓储空间获得竞争对手所没有的某些优势。Netezza 是一家数据仓储应用制造商，其产品是基于高度定制及并行化的PostgreSQL 数据库版本，利用 FPGA(现场可编程门阵列)提升在 x86 集群上的运行性能。

Netezza 利用一组 API 开发其软件开发环境，这组 API 允许 SAS 和 R 算法在其仓储应用中并行运行。同样，它还为Java、C++、Fortran 和 Python 应用提供访问数据仓库的钩子(hook)，并利用 FPGA 而不是 SQL 数据库查询语言提取储存在仓库中的数据。

7 个月之后，当大数据将成为一个大市场这一趋势更加清晰可见时，IBM 以 17 亿美元的价格将 Netezza 收购。

2010 年 10 月，数据仓库制造商 Teradata 利用 TeradataR 软件包在其同名数据仓库中增加了自己的数据库内(in-database)分析。

这将 Teradata Warehouse Miner 工具转变为 R 控制台的一个插件，可在 Teradata 数据中执行 44 种不同的分析函数，同时任何在数据仓库中的存储流程都对 R 开发并可从 R 程序调用。另有 20 个函数可让 R 在 Teradata 环境中运行。

Oracle 的加入

甚至连 Oracle 也加入了 R 语言行动。2 月份，该公司推出Advanced Analytics 工具，作为 Oracle 数据库与 R 分析引擎之间的桥接。

Advanced Analytics 是 Oracle 在其 11g R2 数据库中部署的 Data Mining 附件。当 R 程序员需要运行统计例程时，他们可以在数据挖掘工具箱中调用等同的 SQL 函数，并在该数据库中运行。

如果没有这样的 SQL 函数，遍历数据库节点(如果为集群)的嵌入式 R 引擎将运行 R 例程，收集汇总数据并作为结果将其返回 R 控制台。

另外，Oracle 为其 Big Data Appliance 提供了一个名为 R Connector for Hadoop 的工具，这是一个在 Oracle Exa x86 集群上运行的 Cloudera CDH3 Hadoop 环境。该连接器可让 R 控制台与在 Big Data Appliance 上运行的Hadoop 分布式文件系统和NoSQL 数据库进行通信。

(原文地址：http://os.51cto.com/art/201206/340645.htm)