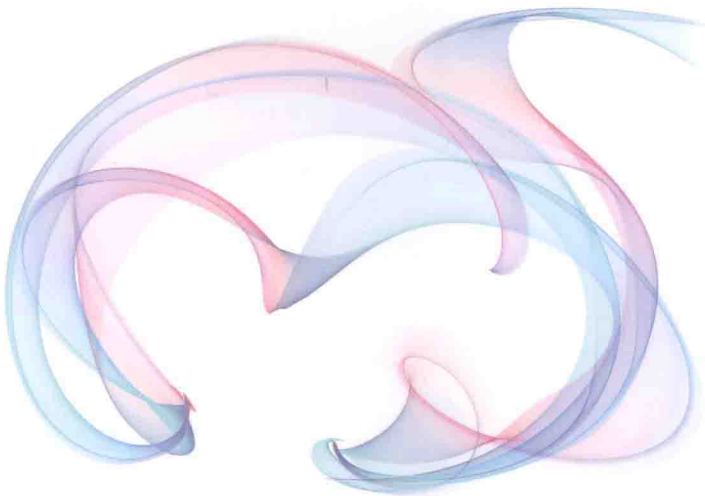
华章科技
Apress

全面展示Splunk的强大功能，系统讲解利用Splunk进行大数据分析的各种技术细节、方法和最佳实践

从数据导入、访问、挖掘和可视化角度，通过大量典型大数据分析案例，全面而系统地介绍Splunk基础知识和使用方法，为快速掌握并灵活运用Splunk从海量数据中挖掘潜在商业价值提供最佳指导

大数据技术丛书



Big Data Analytics Using Splunk

Splunk大数据分析

（美） Peter Zadrozny　Raghu Kodali　著

唐宏　陈健◎译

机械工业出版社
China Machine Press

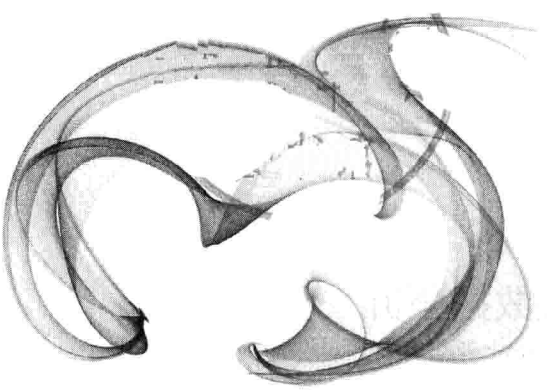
技术丛书

Big Data Analytics Using Splunk

Splunk大数据分析

（美） Peter Zadrozny Raghu Kodali 著

唐宏 陈健◎译



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Splunk 大数据分析 /（美）扎德罗津尼（Zadrozny, P.）等著；唐宏，陈健译 . —北京：机械工业出版社，2014.5
（大数据技术丛书）
书名原文：Big Data Analytics Using Splunk

ISBN 978-7-111-46429-7

I. S… II. ① 扎… ② 唐… ③ 陈… III. 数据处理软件 IV. TP274

中国版本图书馆 CIP 数据核字（2014）第 072758 号

本书版权登记号：图字：01-2013-9164

Splunk 大数据分析

[美] Peter Zadrozny 等著

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：秦 健
印 刷：北京市荣盛彩色印刷有限公司
开 本：186mm × 240mm 1/16
书 号：ISBN 978-7-111-46429-7
版 次：2014 年 5 月第 1 版第 1 次印刷
印 张：19
定 价：69.00

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991 88361066
购书热线：（010）68326294 88379649 68995259
投稿热线：（010）88379604
读者信箱：hzjsj@hzbook.com

译 者 序

绝大多数物理现象、人类活动都会记录在各种媒介中，而随着数字化的普及，这一切又都将转化为数据，人类正在从“卷宗”社会走向“数字”社会。尤其是近年来伴随着智能终端、移动互联网以及物联网等信息技术的发展，数字社会中的数据无论是在类型还是规模方面都在飞速发展，大数据以一种迅疾的速度渗透到我们生活、工作的各个领域。据统计，目前全球被创建和复制的数据总量已超过 2ZB（1021B），远远超过人类有史以来所有印刷材料的数据总量（约 200PB）。想要从庞大的数据库中提取有用的信息或知识，就离不开大数据分析技术和工具。有观点认为，对于已经颠覆或将要颠覆传统行业的应用（如电子商务、互联网金融、物联网），其核心竞争力之一就是大规模的数据分析能力，也就是我们说的大数据能力。

相比传统数据，大数据具有规模大、类型广、时效高等特点，存储和处理这些数据必须引入新的技术和机制。Splunk 是一种典型的大数据处理工具，能够非常高效地按时序对数据进行存储、索引、访问，已广泛应用在多个领域。为此，本书全面系统地介绍了大数据挖掘工具 Splunk，从数据导入、访问、挖掘等角度系统介绍 Splunk 的原理和使用方式，以帮助读者快速掌握 Splunk。

在过去几个月中，黄琰、凡蕙铭、韩超、赖旦冉、何君、蓝贤赟参与了部分翻译，傅桔选、沈书毅、叶玮成担任了部分校审工作，在此感谢他们认真的态度和极大的耐心。当然，本书的翻译工作得以顺利完成，还要感谢机械工业出版社的编辑以及其他所有工作人员在各方面的支持和帮助。最后，对给予我们无私帮助的那些人致以诚挚的谢意。

译者水平有限，书中疏漏在所难免，敬请读者批评指正。

致　谢

首先，我们感谢 Splunk 的工程团队，他们构建了一个一流的产品并且不断地改进它。以下排名不分先后：Stephen Sorkin，感谢他对于全书所有的指导、反馈和意见；Siegfried Puchbauer，感谢他任何关于 DB 连接的宝贵意见；Alice Cotti，感谢她给我们提供了 Foursquare 数据；Sunny Choi，感谢她的支持以及提供的所有性能信息；Dhivya Srinivasan，感谢他让各个应用程序都正确工作；David Carasso，感谢他对于情感分析应用程序的帮助；David Foster，感谢他对于 Twitter 应用程序的帮助。我们还感谢 Rob Das，他编写了第 7 章的草稿。Omcar Paradkar，感谢他作为我们的顾问并否定了我们一些马虎和不好的想法。最后，我们感谢 Rob Reed，他给了我们鼓励并提供了编辑的技巧。

特别感谢 GoGrid（www.gogrid.com）为 Peter 在圣荷西州立大学的课程“Introduction to Big Data Analytics”提供了免费的云服务，本书中所有项目的开发也用到了这个服务。我们也感谢在 Peter 的 2012 年秋季和 2013 年春季学期中参与到确定本书内容的课程实验中的学生们。

Peter Zadrozny 的致谢

在这里不提及我的家人是非常不公平的，他们默默地承受着每个事件中我的缺席。虽然我实际上还在他们身边，但他们也只能看到我的背影。我感谢他们的支持和理解。

Raghu Kodali 的致谢

我不得不说，我很幸运地得到了几位不可思议的人的大力支持，这些人使编写这本书成为可能。感谢 Peter 关于章节不断提供的指导，帮助我把重点放在正确的事情上。

感谢 Paul Stout 在技术审校和确定尽可能准确描述 Splunk 功能的章节上的意见。

感谢 Tom Welsh 在编辑过程中真诚的反馈。

感谢我的父母 Laila 和 Chandra Sekhara Rao Kodali 给予了我无限的耐心，他们是我最

忠实的支持者。他们一直相信我，并且他们从印度到加州来给我鼓励。在我潜心写这本书的时候，他们照顾我的孩子。

感谢我的妻子 Lakshmi，她鼓励我进入这个项目并且支持我。对于我 8 岁的儿子 Yash，他对于我如何开展这个项目有着浓厚的兴趣，不停地查看他是否能提供一定的帮助。对于我 5 岁的儿子 Nikhil，他询问我大数据是什么并且不断地提醒我，要我快点完成写作，这样他就能在书中看到他的名字了。

目　录

第 1 章　大数据和 Splunk

本章将讨论什么是大数据以及处理大数据的几种方法（包括 Splunk 在内）。

1.1　什么是大数据

无可否认，“大数据”已成为软硬件公司为促销产品而过度使用的术语。然而，在炒作的背后，确实蕴含着极其重要的技术趋势和极大的潜在商机。尽管人们经常把大数据与社会化媒体联系在一起，但我们将展开说明大数据远不止于此。在引入定义之前，让我们先来看一些关于大数据的事实。

回到 2001 年，来自麦塔集团（Meta Group，2005 年被高德纳公司收购的一个 IT 研究公司）的 Doug Laney 在一篇研究论文中写道，电子商务大大加速了数据管理朝 3 个方面的发展：数据量（volume）、速率（velocity）和多样化（variety），它们被称为大数据的 3 个 V。正如期望的那样，不少公司在其对大数据的定义中加入了更多的 V。

当提到大数据时，数据量大是第一感觉，也就是大数据的“大”。一些专家认为大数据的起点为 10 亿兆字节（Petabyte，PB）。随着我们产生的数据越来越多，我们相信这个起点肯定会继续增加。然而，数据量本身并不是判断大数据的完美指标，而另外两个指标 V 对大数据的定义有更直接的影响。

速率指的是数据产生的速度或者数据传输的频率。想象一下从洛杉矶高速公路上的传感器传来的数据流，或者从某些机场用于扫描和处理人脸数据的摄像机传来的数据流。此外，还可想象一下流行的电子商务网站用户单击行为所触发的数据流。

多样化则是指信息系统产生的不同数据和文件类型。想象一下 iTunes 商店的音乐文件（大约有 2800 万首歌曲和超过 300 亿次下载），或者 Netflix 服务存储的电影（超过 75 000 部），或者纽约时报网站的文章（从 1851 年至今超过 1300 万篇），推文（每天超过 5 亿条推文），Foursquare 用户的签到地理位置信息（每天超过 500 万条），以及所有包含内置计算机的系统产生的不同系统日志文件。当把这 3 个指标的 V 组合在一起时，你将开始对大数据有更完整的认识。

人们经常和大数据联系在一起的另一个特征是：数据是非结构化的。我们认为，不存在所谓的非结构化数据。我们的观点是：这个困惑来源于一个常见的认知，如果某种数据不符合预定义格式、模型或者结构，那么这种数据被视为非结构化数据。

电子邮件消息通常被当做非结构化数据的典型例子，而邮件的正文可被视为非结构化的，它部分遵循了一个良好定义的数据结构，RFC-2822 规范，同时包含一组字段：From、To、Subject 和 Date。Twitter 消息的结构也一样：消息主体或者叫推文，可被视为非结构化

数据，也可被视为部分结构化的数据。

一般来说，自由文本可被视为非结构化的，因为正如之前所提到的，它不必遵循某个预定义模型。要对文本执行不同的操作，有很多种处理方法，且大多数方法都不需要预定格式。

关系数据库强制要求预定义的数据模型和模型的表中清晰定义的字段，目的是表达它们之间的关系。我们把它称为早期结构绑定（Early Structure Binding），而且在这种设计中，我们必须预先知道这些数据要回答哪些问题，这样才能设计相应的数据模式或结构来回答这些问题。

因为人们常常把大数据与富文本信息的社会化媒体流关联在一起，所以很容易理解为什么人们喜欢将“非结构化”与大数据联系在一起。在我们看来，“多结构化”也许是更准确的描述，因为大数据可以包含多种格式（3 个 V 中的第三个 V）。

把大数据局限在所谓的非结构化数据的范畴是不公平的。结构化数据同样也可以是大数据，特别是暂时保存在辅助存储设备中的数据，一旦装载到数据仓库进行分析即可发现它们所蕴含的巨大价值。这种数据经常被忽略的主要原因是它们的数据量，这类数据的量级通常超过了一般关系数据仓库的容量。

这里介绍一个叫高德纳的 IT 咨询公司在 2012 年提出的定义：“大数据就是高容量、高速度、和 / 或高多样化的信息资产，需要新的处理技术来增强决策能力、原理分析和流程优化。”我们喜欢这个定义，因为它不仅关注实际数据，而且注重大数据的处理方法。本书后面的章节将对这个定义进行更详细的介绍。

为了提高我们对事物的理解能力，我们还喜欢将大数据分类。在我们看来，大数据可分成两个大类：人类产生的数据足迹和机器自动生成的数据。随着互联网活动的持续增加，我们的数字足迹也持续增长。尽管我们每天与数字系统互动，但是大部分人没有意识到哪怕是琐碎的单击或交互都会留下很多信息。在了解互联网相关的统计数据前，我们必须承认，我们唯一熟悉的大数字是麦当劳的口号“亿万招待”以及偶尔曝光的美国政客谈论的万亿级的预算和赤字。为了给读者一个认识，下面展示一小部分互联网统计数据，用以说明网络活动所产生的数据量。我们很清楚，当我们写下这些数据的那一刻它们就已成为旧数据，但是它们的确存在：

- 到 2013 年 2 月为止，Facebook 已经有超过 10 亿用户，其中每天活跃的用户有 6.18 亿。他们每天分享 25 亿条消息并且“喜欢”其他 27 亿条消息，每天产生超过 500TB 的新数据。
- 在 2013 年 3 月，面向商务的社交网站 LinkedIn 拥有超过 2 亿成员，并以每秒 2 个新成员的速度增长，在 2012 年其用户群共进行了 57 亿次职业相关的搜索。
- 照片是个很热门的主题，这是因为大部分人的手机都带有相机功能。这些照片的数量是惊人的。Instagram 用户每天上传 4000 万张照片，每秒“喜欢”其中 8500 张照片，并每秒创建大约 1000 条评论。在 Facebook 上，上传照片的速率保持在每天 3 亿张，一个月将产生 70PB 数据。到 2013 年 1 月为止，Facebook 已经存储了 2400

亿张照片。

- Twitter 有 5 亿用户，并以每天 15 万人的速度增长，其中有 2 亿活跃用户。在 2012 年 10 月，Twitter 每天会产生 5 亿条推文。
- Foursquare 在 2013 年 1 月庆祝其签到数量达到 30 亿，每天大约有来自 2500 万名用户产生的 500 万次签到，这些用户创建了 3000 万条消息。
- 在博客方面，一个叫做 WordPress 的热门博客平台在 2013 年 3 月的报道称，该平台每个月产生将近 4000 万篇新博文和 4200 万条评论，并且每个月超过 3.88 亿用户查看超过 36 亿个页面。Tumblr，另外一个热门的博客平台，同样在 2013 年 3 月的报道称，其用户共创建近 1 亿个博客和超过 440 亿篇文章。在当时，Tumblr 上通常一天内用户共发布 7400 万篇文章。
- 个性化网络电台 Pandora 报道称，在 2012 年他们的用户共收听了 130 亿小时的音乐，也就是大约总时长为 13 700 年时间的音乐内容。
- 与此类似，Netflix 也宣称在 2012 年 7 月间他们的用户观看了超过 10 亿小时的视频，这相当于美国 30% 的网络流量。不仅如此，在 2013 年 3 月，YouTube 也宣称说他们的视频每个月有超过 40 亿小时的观看量，并且每分钟上传 72 小时的视频。
- 在 2013 年 3 月，互联网上共有差不多 1.45 亿个互联网域名，其中大约 1.08 亿使用流行的顶级域名“.com”。互联网是一个非常活跃的领域，在 3 月 21 日，有 167 698 个域名被创建，有 128 866 个域名被删除，净增 38 832 个新域名。
- 在更平常的电子邮件世界，来自 Mashable 的 Bob Al-Greene 称，在 2012 年 11 月，每天有超过 1440 亿封电子邮件被发送，其中大约 61% 来自企业。领先的电子邮件服务是 Gmail，它拥有 4.25 亿活跃用户。

回顾这些统计信息，人类网络行为产生的数据足迹毫无疑问是巨大的。我们能快速从中看出 3 个 V，为了让读者了解大数据如何影响经济，我们来分享一个来自基于用户评论的网站 Yelp 在 2013 年 1 月（当时他们有 1 亿个独立访客和超过 100 万条评论）发布的公告：“Yelp 的企业主的调查报告称，平均而言，所有受调查类别中的顾客第一次访问 Yelp 时消费 101.59 美元。这些花费的来源可以从雇佣一个盖房顶的人到买一个新床垫，甚至可以在早晨买一杯咖啡。如果这 1 亿个独立访问者 1 月份在本地商家每人消费 100 美元，那么 Yelp 将会给本地商业带来超过 100 亿美元的影响。”

我们不会拿互联网环境下每天生活中的每一分钟或每一秒的统计数据来烦你。然而，举几个相关的大数据的例子有助于巩固这个概念。当我们访问 Amazon 网站或者在 Netflix 选择电影时，所得到的推荐基于大数据分析，沃尔玛也采用同样的方法了解一个区域的消费者偏好并根据此分析来安排库存。现在，你一定对人类数据足迹的大数据量有了很好的认识，并清楚认识到这些数据对经济和社会产生的影响。社会化媒体只是大数据的一部分。

大数据的第二个分类是机器数据。人类的数字化生活需要大量的防火墙、负载平衡器、路由器、转换器和电脑来支持。所有这些系统都会产生日志文件，从安全和审计的日志文件到描述网站访问者行为的网站日志文件，包括臭名昭著并被抛弃的购物车功能。

要计算出需要多少服务器才能支持人类产生的数字足迹几乎不可能，因为所有的公司都对其服务器数量保密。许多专家试图根据用电量（根据某些公司愿意公布的电源使用效率指标）对最有名的公司，例如 Google、Facebook 和 Amazon，来计算它们拥有的服务器数量。利用这个方法，James Hamilton 在 2012 年 8 月发表的一篇博文中推测，Facebook 大约有 180 900 台服务器，Google 有超过 100 万台服务器。其他专家表示，在 2012 年 3 月 Amazon 大约有 5 亿台服务器。2012 年 9 月，《纽约时报》刊登了一篇煽动性质的文章，声称在美国有成千上万的数据中心，它们的耗电量大约占美国总耗电量的 2%，但是这 2% 的耗电量中有 90% 或更多被浪费，因为很多服务器实际上没有使用。

我们只能猜测，全世界处于工作状态的服务器数量在百万台左右。当再加上所有其他的典型数据中心基础设施组件，如防火墙、负载平衡器、路由器、交换机和其他也产生日志文件的组件，我们可以看到支撑我们数字足迹的基础设施所产生的日志格式的机器数据量是巨大的。

有趣的是，在不久前，大多数包含了机器数据的日志文件都大半被人们忽略。这些日志文件都是包含有用数据的金矿，因为它们包含对 IT 和商业活动的重要情报，因为这些日志文件忠实地记录了客户活动和习惯以及产品和服务的使用情况。上述日志文件使得公司的端到端业务变得透明化，可以用来改善客户服务，确保系统安全，也可以帮助系统设计满足法律法规的要求。更重要的是，日志文件可以帮助我们找到已经发生的问题，并且可以协助预测相同的问题未来将发生的时间。

到目前为止除了我们介绍的机器数据外，还有用于实时采集数据的传感器。大多数工业设备都包含能产生大量数据的内置传感器。例如，在燃气机中用于发电的一个叶片每天将产生 520GB 的数据，而在一个这样的燃气机中有 20 个叶片。一个跨大西洋的航班飞机能产生几个 TB 的数据，这些数据可以用来简化维护操作，提高安全性能，以及（对航空公司最重要的是）减少燃料消耗。

另一个有趣的例子是日产汽车的品牌 Leaf，这是一种纯电动汽车。它有一个称为 CARWINGS 的系统，这个系统不仅提供传统的远程信息处理服务和一个智能手机的应用程序来控制汽车的所有功能，而且能无线传输车辆统计数据到中心服务器。每个 Leaf 所有者可以跟踪他们的驾驶效率，并且能和其他 Leaf 车主比较他们的汽车能耗数据。我们不知道日产公司从 Leaf 型号汽车上收集的详细信息是什么，也不清楚他们收集这些信息的用途，但是我们可以清楚地看到大数据的 3 个 V 在例子中的体现。

一般来说，传感器数据属于工业大数据类别，尽管最近“物联网”已经成为一个热门词汇，用来描述带传感器的设备构建的联网世界，世界上从电表到自动售货机有 3 亿多个联网设备。本书中将不会涉及这一类的大数据，但书中描述的方法和技术可以很容易地用于工业大数据分析中。

1.2 非传统的数据处理技术

大数据不仅仅是关于数据的，它还涉及能更好地处理大数据的3个指标（3个V）的非传统数据处理技术，因为这些技术可以增加数据的价值。传统的关系数据库有以下众所周知的特点：

- 事务处理是支持ACID属性的：
  - 原子性（Atomicity）：整个事务中的所有操作，要么全部完成，要么全部不完成，不可能停滞在中间某个环节。
  - 一致性（Consistency）：在任意事务操作结束后，系统将处于有效状态。
  - 独立性（Isolation）：创建结果的操作看起来像顺序执行的，每次一个。
  - 持久性（Durability）：所有对系统的更改都是持久的。
- 在同时处理成千上万用户请求的同时，响应时间通常在毫秒级范围。
- 数据量是TB级的。
- 通常使用SQL-92标准作为主要的编程语言。

一般来说，关系数据库不能很好地处理这3个V。正因为如此，人们发明了许多方法去解决这3个V固有的问题。这些方法将牺牲一个或多个ACID属性，有时候甚至全部都要放弃，以换取处理大容量、速率或多样化的可扩展性。这些非传统的方法也将放弃快速响应时间或处理大量并发用户的能力来支持解决3个V中的一个或多个。

一些人将这些非传统的数据处理方法称为NoSQL，并根据它们存储数据的方式来分类，如键－值对（key-value）存储和文档存储，而文档的定义随产品的不同而不同。根据不同的讨论对象，可能会有更多的分类。

开源的Hadoop可能是大数据世界里最知名的软件框架，但它绝不是唯一的。作为一个框架，它包含了许多旨在解决分布式数据存储、大数据检索和分析的相关问题的组件。它提供了两个基本功能来实现上述目的，这两个基本功能可运行在消费级服务器集群上，它们是：

- 一个称为HDFS的分布式文件系统，它不仅存储数据，而且复制数据，使得数据一直可用。
- 一个叫MapReduce的针对并行化问题的分布式处理系统，它采用两步走的方法。第一步（或称为Map），将一个问题细分成许多小的问题，然后发送到服务器群进行处理。第二步（或称为Reduce），将第一步产生的结果结合起来，得出原问题的最终结果。

Hadoop的一些其他的组件，一般称为Hadoop生态系统，包括Hive，一个基于Hadoop基本功能进行更高层次抽象而衍生的系统。Hive是一个数据仓库系统，在该系统中用户可以编写SQL-92标准的指令并自动转换为MapReduce任务。Pig是另一个与Hive有相似功能且基于Hadoop的高层级抽象系统，只不过它使用一种称为Pig Latin的更加面向数据流的编程语言。