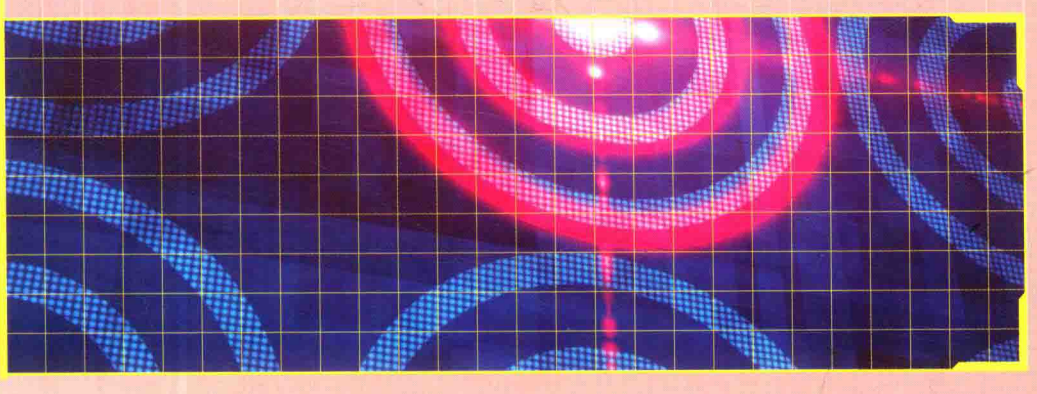
先·进·信·号·处·理·系·列

大数据技术体系与开源生态

刘 驰 罗 童 林秋霞 王新科 樊路遥 | 编著



中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

先·进·信·号·处·理·系·列

大数据技术体系与开源生态

刘 驰 罗 童 林秋霞 王新科 樊路遥 | 编著

人民邮电出版社
北京

图书在版编目（CIP）数据

大数据技术体系与开源生态 / 刘驰等编著. -- 北京：人民邮电出版社，2018.8
（先进信号处理系列）
ISBN 978-7-115-49223-4

Ⅰ. ①大… Ⅱ. ①刘… Ⅲ. ①数据处理－研究 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第197403号

内 容 提 要

本书从大数据生命周期的角度阐述大数据技术体系与开源生态的发展。全书分为七篇，包括大数据技术体系与开源生态概述、大数据获取技术、大数据管理技术、大数据处理技术、大数据分析与挖掘技术、大数据可视化与交互技术、大数据安全与治理技术。又分为15章，详细介绍大数据的技术概况、发展近况和技术优势、软件架构和应用场景等内容。

本书适合大数据和人工智能业内人员、各大高校相关专业的高年级本科生和研究生，以及对大数据应用中各类框架组件的爱好者阅读。

◆ 编　　著　刘　驰　罗　童　林秋霞　王新科　樊路遥
责任编辑　代晓丽　马　涵
责任印制　彭志环

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路11号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京鑫正大印刷有限公司印刷

◆ 开本：700×1000　1/16
印张：23.25　　2018年8月第1版
字数：456千字　　2018年8月北京第1次印刷

定价：159.00元

读者服务热线：(010)81055488　印装质量热线：(010)81055316
反盗版热线：(010)81055315

前　言

大数据是以容量大、类型多、存取速度快、应用价值高为主要特征的数据集合。随着信息技术和人类生产生活的交汇融合，互联网快速普及，全球数据呈现爆发增长、海量集聚的特点，对经济发展、社会治理、国家管理、人民生活都产生了重大影响。大数据技术的应用和发展，为人类提供了认识复杂系统的新思维和新手段，成为推动经济转型升级的新动力，也是提升社会治理能力的新途径，以及提升国家综合能力和保障国家安全的新利器。

美国政府最先对大数据技术革命做出战略反应，从2012年到2016年共实施了4轮政策行动，美国白宫成立了“美国大数据研发高级指导小组”，实施了《大数据研究和开发计划》，加强了在大数据研发和应用方面的布局。欧盟发布了《数据驱动经济战略》，大力推动“数据价值链战略计划”，倡导欧洲各国抢抓大数据发展机遇。英国政府发布了《把握数据带来的机遇：英国数据能力战略》，从提升数据分析技术、加强国家基础设施建设、确保数据安全和共享等方面作出部署。此外，日本、韩国、澳大利亚等国均出台了相关政策，大力推动大数据应用及产业发展。

我国在党的十八届五中全会上将大数据上升为国家战略。2015年8月，国务院印发了《促进大数据发展行动纲要》；2017年1月，工信部印发了《大数据产业发展规划（2016—2020年）》等政策指导文件，为大数据战略的实施指明了具体方向。在党的十九大报告中，习近平总书记明确指出：“推动互联网、大数据、人工智能和实体经济深度融合。”2017年12月8日，在中共中央政治局第二次集体学习时，习近平总书记发表了“大数据发展日新月异，我们应该审时度势、精心谋划、超前布局、力争主动，深入了解大数据发展现状和趋势及其对经济社会发展的影响，分析我国大数据发展取得的成绩和存在的问题，推动实施国家大数据战略，加快完善数字基础设施，推进数据资源整合和开放共享，保障数据安全，加快建设数字中国，更好服务我国经济社会发展和人民生活改善”的讲话。自此，我国大数据发展开启了新的篇章。

在这样的大背景下，全球大数据产业日趋活跃，技术演进和应用创新加速发展，全世界各地的创新人才不断地涌入到大数据产业中来，进一步反哺大数据技术的发展。再加上开源生态的不断壮大，进一步促进了大数据技术的分享和发展。

大数据技术已成体系，从大数据生命周期的角度，其可以划分为大数据获取、大数据管理、大数据处理、大数据分析与挖掘、大数据可视化、大数据安全与治理 6 个方面，本书也是从这 6 个方面来对大数据技术体系和开源生态的建设进行相应的介绍。

本书共 7 篇 15 章，其中：

第一篇：大数据技术体系与开源生态概述，包括第 1~4 章。第 1 章为大数据技术体系概述。第 2 章描述了开源生态与代码托管平台简介。第 3~4 章着重介绍了大数据技术和云计算技术这两类技术的开源生态建设及发展情况，包括 Apache 软件基金会、Linux 基金会、OpenStack 基金会、Cloud Native Computing Foundation 和开源中国。其中，第 1 章由段雄编写，第 2~4 章由罗童编写。

第二篇：大数据获取技术，包括第 5 章，从当前主流的消息队列相关技术的角度阐述数据如何从异构系统采集并融合在一起，包括 ZeroMQ、RabbitMQ、ActiveMQ、Apache Kafka4 个开源项目。第 5 章由樊路遥编写。

第三篇：大数据管理技术，包括第 6~7 章。第 6 章描述了当前主流数据库技术，包括 MySQL、PostgreSQL、MongoDB、Apache CouchDB、Vertica、Apache HBase、Neo4j、OrientDB、InfiniteGraph、Alluxio、Apache Tajo 11 个开源项目。第 7 章介绍了大数据平台资源管理技术，包括 Apache Zookeeper、Apache Hadoop YARN、Apache Mesos、Apache Mnemonic 4 个开源项目。其中，第 6 章由温琦和胡柏青编写，第 7 章由方久鑫编写。

第四篇：大数据处理技术，包括第 8~9 章。第 8 章介绍了当前主流的大数据批处理平台，包括 Apache Hadoop、Apache Spark、Apache Kylin 这 3 个开源项目。第 9 章介绍了当前主流的实时流处理平台，包括 Apache Storm、Apache Spark Streaming、Apache Flink、Apache Beam 和 Apache Apex 4 个开源项目。其中，第 8 章由王新科编写，第 9 章由林秋霞编写。

第五篇：大数据分析与挖掘技术，包括第 10~11 章。第 10 章介绍了主流大数据分析工具，包括 Apache Mahout、Apache Lens、Apche Spark MLlib、Scikit-Learn4 个大数据开源项目。第 11 章介绍了主流人工智能开源平台，包括 TensorFlow、Caffe、PyTorch、TensorFlow Lite4 个开源项目。其中，第 10 章由陈喆和王新科编写，第 11 章由樊路遥、朴成哲和温琦编写。

第六篇：大数据可视化与交互技术，包括第 12~13 章。第 12 章介绍了两个主流大数据：可视化与交互技术，Tableau 和 Apache Zeppelin。第 13 章介绍了其他大数据可视化与交互技术，包括 BIRT、KNIME、Jaspersoft Community 3 个开源

项目。其中，第 12 章由张晶和罗童编写，第 13 章由陈喆、胡柏青和林秋霞编写。

第七篇：大数据安全与治理技术，包括第 14~15 章。在第 14 章，通过介绍 Apache Falcon 和 Apache Atlas 两个开源项目阐述了大数据治理技术。第 14 章介绍了 Apache Kerberos、Apache Ranger、Apache Sentry 和 Apache Metron 4 个大数据安全开源项目。其中，第 14 章由陈喆编写，第 15 章由陈喆和张晶编写。

由于编者及撰写者的认识和水平所限，本书内容仅从大数据技术体系及其开源生态发展的一个视角写作，既不全面，也难免偏颇。但能够为读者提供一定的参考，则本书目的已达。欢迎广大读者对本书进行批评指正。

作者

2018 年 4 月

目 录

第一篇 大数据技术体系与开源生态概述

第三篇 大数据管理技术

第五篇 大数据分析与挖掘技术

第一篇

大数据技术体系与开源生态概述

随着互联网的飞速发展，我们的生活中到处都充斥着数据，并且数据的规模正在呈指数级增长，显然大数据时代已经到来。对于如此庞大的数据量，我们需要对其进行相应地获取、存储、管理、处理、分析和可视化等操作。大数据技术也就成了解决这些问题的核心。同时随着开源软件的不断发展，大数据技术也得到不断的发展。现如今，大数据技术的发展已经相对成熟，开源生态系统也不断地壮大起来。第一篇，主要对大数据技术体系进行相应介绍，介绍了现在比较流行的一些大数据技术，然后对开源生态中的大数据开源组织和云计算开源组织进行相应介绍。本篇的目的是使读者对大数据技术体系和开源生态有一个理论上的认识，为后面学习具体的大数据技术打下基础。