普通高等教育“十三五”规划教材
国家新闻出版改革发展项目库入库项目
数据科学与大数据技术专业教材丛书

# 流数据分析技术

STREAM DATA ANALYSIS TECHNOLOGY

李静林　袁　泉◎编著

北京邮电大学出版社
www.buptpress.com

DATA SCIENCE

## “数据科学与大数据技术专业教材丛书”

- 《大数据技术基础》
- 《大数据技术基础实验》
- 《网络科学与计算》
- 《数据可视化》
- 《机器学习》
- 《计算机视觉》
- 《R语言编程与数据科学》
- 《Python语言程序设计》
- **《流数据分析技术》**
- 《分布式计算与云计算》
- 《NoSQL数据库技术》
- 《数据仓库与数据挖掘》

“互联网+”创新型教材
北邮智信

扫一扫，下载安装
“北邮智信”APP

北邮智信
1808022165
刮开验证 获取资源

刮开涂层，
在“北邮智信”APP中
验证教材，加载资源

ISBN 978-7-5635-5915-2

定价：38.00元

普通高等教育“十三五”规划教材
国家新闻出版改革发展项目库入库项目
数据科学与大数据技术专业教材丛书

# 流数据分析技术

李静林　袁　泉　编著

北京邮电大学出版社
www.buptpress.com

## 内容简介

流数据分析技术是一种实时或准实时的，对动态数据集合甚至无界时间序列数据进行特征和态势认知的技术，目前已经广泛应用于互联网/移动互联网、物联网、气象、金融等多个领域，支撑运营管理、应用性能管理、监测与测控等多种服务，是大数据的重要研究方向之一。本书以流数据的基本特征为核心内容，突出流数据与传统大数据的联系与区别，介绍流数据的基本处理方法和分析方法。重点内容包括：流数据与流式计算、流数据处理技术、流数据分析技术、流数据处理模型与处理框架等。本书还介绍了流数据分析技术的一些最新进展及流计算框架的最新发展。

本书可作为计算机学科相关专业，特别是数据科学与大数据技术专业的教材。

**图书在版编目(CIP)数据**

流数据分析技术 / 李静林，袁泉编著. -- 北京：北京邮电大学出版社，2020.1

ISBN 978-7-5635-5915-2

Ⅰ. ①流… Ⅱ. ①李… ②袁… Ⅲ. ①数据处理—高等学校—教材 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2019）第 256789 号

**策划编辑**：姚 顺 刘纳新 **责任编辑**：刘春棠 **封面设计**：柏拉图

---

**出版发行**：北京邮电大学出版社

**社 址**：北京市海淀区西土城路 10 号

**邮政编码**：100876

**发 行 部**：电话：010-62282185 传真：010-62283578

**E-mail**：publish@bupt.edu.cn

**经 销**：各地新华书店

**印 刷**：保定市中画美凯印刷有限公司

**开 本**：787 mm×1 092 mm 1/16

**印 张**：12.25

**字 数**：255 千字

**版 次**：2020 年 1 月第 1 版

**印 次**：2020 年 1 月第 1 次印刷

---

ISBN 978-7-5635-5915-2 定价：38.00 元

# 大数据顾问委员会

宋俊德　王国胤　张云勇　郑　宇

段云峰　田世明　娄　瑜　孙少陵

王　柏

# 大数据专业教材编委会

**总主编：** 吴　斌

**编　委：** 宋美娜　欧中洪　鄂海红　双　锴

于艳华　周文安　林荣恒　李静林

袁燕妮　李　劼　皮人杰

**总策划：** 姚　顺

**秘书长：** 刘纳新

# 前　言

2016年2月，教育部公布新增数据科学与大数据技术专业，到2019年3月获批此专业的院校已经达400多所，相应的专业课程建设亦提上日程。在大数据研究领域中，流数据作为一种特殊的动态数据集合甚至无界时间序列数据，是互联网/移动互联网、物联网、气象、金融等多个领域大数据的重要表现形式。对流数据进行实时或准实时的分析，以快速地获取其动态特征并进行态势认知，已经成为支撑运营管理、应用性能管理、监测与测控等多种服务的基础技术。如何考虑流数据时变性的特点对其进行时延敏感的建模分析，一直是大数据技术领域的重点研究方向。因此，在数据科学与大数据技术专业培养方案中，设置了“流数据分析技术”这门重要的专业课程。但是，国内目前缺乏专门的流数据相关书籍或研究报告，《流数据分析技术》这本书就是为这门课程编写的一本参考教材。

业界目前尚无准确的流数据的概念定义，本书中将流数据定义为一种实时到达的具有规模大、基数高、统计特征复杂变化特性的数据流。与传统的大数据不同，流数据具有实时性、易失性、突发性、无序性的特点。

实时大规模抵达特性：流数据是持续产生的，持续产生的数据意味着无法预期数据的边界，无法像传统大数据一样采用批处理模式。流数据处理系统为了能赶得上流数据的产生速率，实现实时或准实时的处理，需要采用流式处理或微批处理的模式，使用有限的存储资源进行数据处理。有限的资源会进一步导致超期的数据被丢弃，因此流数据处理仅能对缓存的数据进行有限次数访问，这导致一些需要数据全集或需要多次重复遍历数据的传统大数据挖掘分析算法难以应用在流数据环境中，必须有针对性地优化。同时，流数据处理系统也必须具有容错能力，能够在一次数据处理过程中尽可能全面、准确、有效地获得有价值的信息。而流数据的突发性和无序性意味着无法准确预期抵达数据的量和顺序，这就要求流数据处理系统具备更大的弹性，且在进行流数据分析过程中

不能过多地依赖数据流间的顺序或内在逻辑。

高基数特性：流数据与大数据一样可能存在高基数特征，即数据中的不同类别数量很大。但是与传统大数据不同的是，由于流数据的持续产生，某些数据的类别可能会延迟很久才出现，而另一些早早出现的数据类别却可能是“小众”数据类别。这都导致流数据的处理面临处理性能与资源消耗的权衡。

统计特征复杂变化特性：由于流数据的时间跨度大，从流数据中获得的统计特征可能随时间而变化，而这种特征变化是传统大数据分析过程所不会考虑的。流数据的这种统计特征随时间变化的特性被称为“概念漂移”。概念漂移导致流数据分析方法必须考虑如何随时间进行调整，以适应统计特征的变化。而流数据处理模型和处理方法也需要能够配合流数据分析方法，使之能够获得足够量的数据，能够体现出统计特征变化。

“流数据分析技术”课程旨在培养学生掌握流数据分析与流式数据处理的基本知识和方法，以及运用流计算模式去思考和解决现实问题的能力，提高学生的创新意识，开阔学生研究视野，为学生的进一步深造打下基础。

流数据处理方法和流计算模型在20世纪90年代即已经开始被研究，但在本书之前，国内仅有一些流计算框架的工具书。国内有关流数据的内容都分散在各种与大数据或大数据挖掘工具相关的书籍中，不成体系。国外有一些专门阐述流数据分析与处理的书籍，但外文著作的组织方式和内容并不符合我国教材的需求。同时，随着近年来大数据研究的持续发展，一些新型的流数据分析与处理方法层出不穷，需要一本针对流数据的专门著作，对流数据与大数据的联系与区别、流数据的分析方法与处理方法、流数据处理模型与框架等进行系统化的集中阐述。

本书共分为7章。

第1章为流数据与流计算。这一章首先介绍大数据与流数据的联系与区别，提出了流数据的基本概念和特征，并综述了流数据分析方法和流数据处理方法所需要考虑的问题及相应的方法基础。

第2章为流数据概要结构构建技术。这一章主要介绍流数据处理模型，重点阐述了流数据处理模型中概要数据结构的目的和意义，并对抽样、草图、小波、直方图等多种概要数据结构的构建方法进行了系统化阐述。

第3章为流数据频繁模式挖掘技术。这一章介绍了流数据分析中最基础的频繁项和频繁模式挖掘算法。频繁模式挖掘目的是找出数据流中出现频率大于一定阈值的数据或模式。本章对经典的黏性抽样、有损计数等方法进行了详细的阐述。

第4章为流数据聚类分析技术。这一章介绍了流数据分析过程中的聚类算法。聚

类是将数据对象集合中相似的对象元素划分为同一簇的过程，也是研究得最深入、最广泛的一类问题。本章对基于划分的、基于层次的、基于密度的、基于网格的等多种不同的流数据聚类算法进行了阐述。

第5章为流数据分类分析技术。这一章介绍了流数据分析过程中的分类算法。分类是应用最广泛的数据挖掘技术之一，流数据的概念漂移特点也对流数据分类技术的实现影响很大。本章重点阐述了两种分类技术，包括基于贝叶斯的分类和基于决策树的分类，并阐述了这些方法的进一步扩展。

第6章为流数据学习与时间序列分析技术。这一章介绍了流数据分析过程中的回归方法。回归是对多变量关系的统计学习，可以用来进行预测。本章重点阐述了增量式参数回归所需的最优化方法，并对最新的基于非参数回归的模型树学习、规则学习方法进行了阐述。

第7章为流数据处理模型与框架。这一章介绍了近年来工业界较为流行的Storm、Spark、Flink等多种流处理框架。本章的侧重点并不在工具的使用，而是聚焦在不同流计算框架的设计思路、计算模型、容错模型、实现机制上，通过对比流计算引擎的设计实现思路，加深对流数据处理模型和方式的理解。

本书可以作为数据科学与大数据技术专业的本科高年级专业课教材，也可以作为研究生相关课程的参考材料。

本书的编写得到了北京邮电大学网络与交换技术国家重点实验室交换与智能控制研究中心教师与研究生的支持，他们是：魏晓娟、李梓延、莫浩杰、薛亚青、李冠略、冯亦瑄，在此一并表示感谢。另外还要感谢共同完成多项相关科研项目，进行流数据相关分析算法和处理引擎研发的已经毕业的博士和硕士研究生们，他们曾经的努力是本书能够推出的前提。

作为在计算机领域从事科研和教学的教师，专业知识的深度和广度的局限性使得本书仍存在不足之处，欢迎广大读者反馈对本书的意见和建议，我们将随着“流数据分析技术”专业课程的建设，不断改进本书的质量。

**李静林**
**北京邮电大学**

# 目　录

# 第1章 流数据与流计算

## 1.1 大 数 据

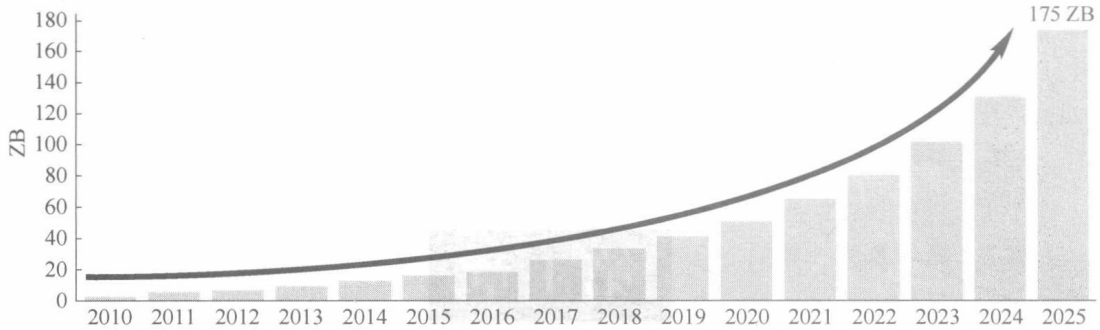
### 1.1.1 大数据的发展

2012年5月，联合国发表名为《大数据促发展：挑战与机遇》的政务白皮书，其中引用国际数据公司(IDC)的数据预测，指出“全球可用数据量从2005年的150 EB(Exabyte)增长到2010年的1 200 EB，预计未来几年每年将增加40%。基于这一增长率，到2020年，全球数据量将增长到2007年的44倍，平均每20个月增长一倍”。在白皮书中，联合国首次提出“大数据(Big Data)”的概念，并指出“当前世界正在经历数据革命，或称‘数据洪流’”。

2018年，IDC在《数据时代2025——数字化的世界(从边缘到核心)》白皮书中指出，全世界数据将从2018年的33 ZB(Zettabyte)进一步增长到2025年的175 ZB。这一数据进一步印证了联合国政务白皮书的预测(如图1-1所示)。

基于这一数据增长的预测，IDC白皮书指出“数据驱动的世界将持续在线，持续跟踪，持续监视，持续地听和看，因为它将持续学习”。

“The data-driven world will be always on, always tracking, always monitoring, always listening and always watching—because it will be always learning.”




数据来源：Data Age 2025 IDC, 2018年11月。

图 1-1　全球数据量预测

在数据驱动的世界中，终端、边缘、云端都发挥着关键作用，云端提供集中式的数据存储、归档、服务交付、更高深层次的数据挖掘分析、指挥和控制，以及法规遵从性。边缘则提供了更多的智能和交互性，对数据进行预处理，并将结果发回给云端进行更深入的分析。因此，数据从终端流到网络边缘和云端，进而再从网络云端流回网络边缘和终端。数据的这种传播驱动了数据的进一步增长，并对数据的挖掘分析和利用产生影响，实现了整个网络的智能，如图 1-2 所示。

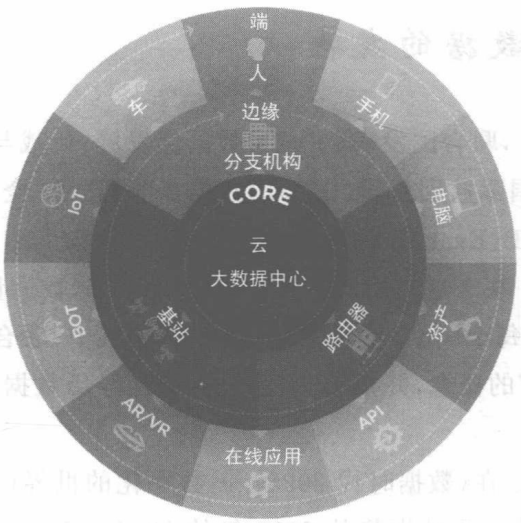
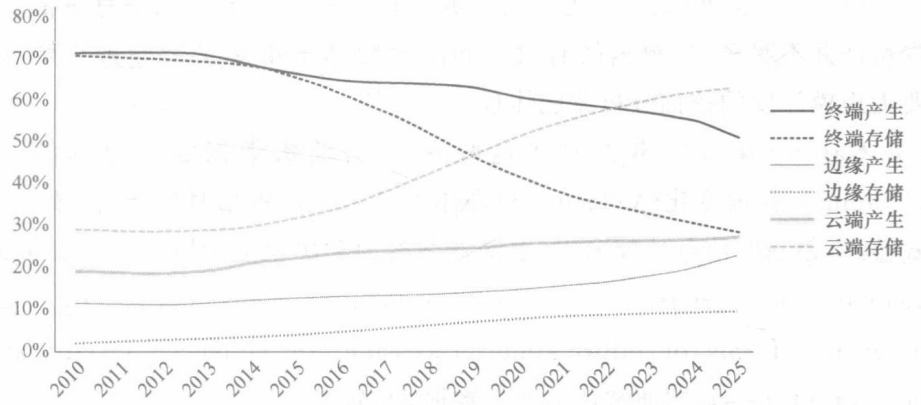



图 1-2　终端-边缘-云端的数据应用模型

随着 5G 技术的普及和边缘计算技术的发展，未来世界中数据的产生与数据的存储将产生显著的变化。终端侧产生和存储的数据将显著降低，网络边缘侧和云端存储和产生的数据则将显著升高，如图 1-3 所示。IDC 预计到 2024 年，云端存储的数据将是终端的 2 倍。




数据来源：Data Age 2025, Seagate & IDC, 2018年11月。

图 1-3 数据产生与存储位置

目前，不同的系统还是由自身维护、管理、分析数据，而在未来的数据驱动的世界中，大数据的挖掘将会把产生异构数据的系统整合起来，形成一个整体。进而，真正地改变人们的生活以及理解世界的方式。

## 1.1.2 大数据的概念

大数据的概念

目前大数据还没有统一的定义，2011 年 IDC 的报告[1]中对大数据给出了一个轮廓的描绘："大数据技术描述了一个技术和体系的新时代，被设计于从大规模多样化的数据中通过高速捕获、发现和分析技术提取数据的价值"。这个描绘刻画了大数据的 4V 特性：规模大(Volume)、变化快(Velocity)、多样性(Variety)和价值密度低(Value)。

一般认为大数据的这几个特征的解释如下。

(1) 规模大：由于数据产生的用户多、数量大、位置分布广，终端、边缘、云端都会持续产生数据，因此数据的规模远超以往，这对数据的存储和处理都提出挑战。

(2) 变化快：由于产生和使用数据的用户庞大，数据会在流中持续不断地到达，这对数据的实时处理产生极大的挑战。同时由于用户之间的复杂交互，数据在用户之间快速传播，且传播行为复杂，这就造成数据的易变性(Variability)，进一步加剧了实时提取有价值信息的难度。

(3) 多样性：由于数据的来源多样，数据的类型和数据的结构也多种多样，有结构化的数据，如日志、传感器数据等，也有非结构化的数据，如文本、语音、视频、图形等，还有介于结构化和非结构化之间的半结构化数据。不同数据的处理方法不同，体现出的潜在特征、规律等也不同，这极大地提升了多源数据异构融合处理的门槛。

(4) 价值密度低:未经处理的数据具有高度的冗余,数据信息量低,数据特征并不明显,且数据的有效性和准确性(Validity,Veracity)存疑,需要大规模深度的挖掘分析才能体现出其价值。

Gartner 在 2012 年总结了这些观点,并将大数据定义为"高容量(Volume)、高度变化(Velocity)和多样化(Variety)的信息资产,需要成本效益高、创新的信息处理形式,以增强洞察力和决策能力"(High volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making),即通常认为的大数据 3V 定义。

2015 年李学龙等在论文[2]中对大数据的定义进行了总结,指出"大数据的定义呈现多样化的趋势,达成共识非常困难",并总结了当前较为流行的三种定义方法:

- 属性定义(Attributive Definition),包括前文所述的 4V、3V 定义方法;
- 比较定义(Comparative Definition),即从演化的观点探讨了具有时间和跨领域变化的数据集才能被认为是大数据;
- 体系定义(Architectural Definition),即大数据包含大数据科学(Big Data Science)和大数据框架(Big Data Frameworks)。大数据科学围绕大数据的获取、处理、评估等技术展开研究;大数据框架则侧重于大数据问题的分布式处理和分析方法。

### 1.1.3 大数据思维

由于大数据的规模大、变化快、种类杂、价值密度低这些特性,大数据的挖掘处理方式与传统数据挖掘有较大的区别。

大数据思维

传统的数据处理主要针对抽样数据,通过准确的数据建模,以获得精确的数据处理结果。而大数据的数据价值密度低,导致大数据的数据量大,且数据统计特征分布不均匀,使用传统采样分析的方法难以获取数据的准确特征。同时,由于大数据的变化快,长时间之前的数据特征可能已经无法指导当前的应用行为,因此大数据条件下,精确性已不再是追求的最终目标,更多的时候是挖掘大数据蕴含的变化规律,并对宏观趋势给出快速预测。同样由于大数据的数据种类杂,多源数据间并不一定存在必然的因果关系,而是需要发现数据间存在的关联关系,以挖掘多源数据之间存在的事实规律,从而实现对未来的准确预测。

因此,大数据思维与传统数据处理思维的主要区别体现在以下三个