[美] Sumit Gupta, Shilpi Saxena 著　张广骏 译

实时大数据分析

基于Storm、Spark技术的实时应用

Real-Time Big Data Analytics

清华大学出版社

实时大数据分析——基于 Storm、Spark 技术的实时应用

[美] Sumit Gupta

Shilpi Saxena 著

张广骏 译

清华大学出版社

北 京

内 容 简 介

本书详细阐述了实时大数据分析的实现过程，主要包括大数据技术前景及分析平台，Storm 的熟悉，用 Storm 处理数据，Trident 概述和 Storm 性能优化，Kinesis 的熟悉，Spark 的熟悉，使用 RDD 编程，Spark 的 SQL 查询引擎，用 Spark Streaming 分析流数据以及 Lambda 架构等内容。此外，本书还提供了相应的示例、代码，以帮助读者进一步理解相关方案的实现过程。

本书适合作为高等院校计算机及相关专业的教材和教学参考书，也可作为相关开发人员的自学教材和参考手册。

北京市版权局著作权合同登记号 图字：01-2017-7944

图书在版编目（CIP）数据

实时大数据分析：基于 Storm、Spark 技术的实时应用/（美）萨米特·古普塔，（美）希尔皮·萨克塞纳著；张广骏译. —北京：清华大学出版社，2018
书名原文：Real-Time Big Data Analytics
ISBN 978-7-302-47728-0

Ⅰ. ①实…　Ⅱ. ①萨…　②希…　③张…　Ⅲ. ①数据处理软件　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2017）第 166833 号

责任编辑：贾小红
封面设计：刘　超
版式设计：李会影
责任校对：赵丽杰
责任印制：杨　艳

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：北京富博印刷有限公司
装 订 者：北京市密云县京文制本装订厂
经　　销：全国新华书店
开　　本：185mm×230mm　　**印　　张**：16.25　　**字　　数**：333 千字
版　　次：2018 年 1 月第 1 版　　**印　　次**：2018 年 1 月第 1 次印刷
印　　数：1～3000
定　　价：79.00 元

产品编号：073176-01

译 者 序

大数据是业内热门的话题，大数据存储后如何做好实时处理是重要的技术焦点。作为当前最受关注的实时大数据开源平台项目，Storm 和 Spark 都能为广大潜在用户提供良好的实时大数据处理功能。除在功能方面的部分交集外，Storm、Spark 还各自拥有独特的特性与市场定位。根据业务应用需求选用恰当的技术平台是大数据应用成功的关键，本书既涵盖了不同实时数据处理框架和技术的基础知识，又论述了大数据批量及实时处理的差异化细节，还深入探讨了使用 Storm、Spark 进行大数据处理的技术和程序设计概念。

本书以丰富的应用场景及范例说明如何利用 Storm 进行实时大数据分析，既涉及了 Storm 的组件及关键概念内部实现的基础，又整合了 Kafka 来处理实时事务性数据，还探讨了 Storm 微小批处理抽象延伸的 Trident 框架和性能优化。此外，包括了使用 Kinesis 服务在亚马逊云上处理流数据的内容。本书后半部分着重介绍了如何利用 Spark 为实时和批量分析开发通用型的企业架构和应用，既可通过 RDD 编程轻松实现数据转换和保存操作，亦介绍了 Spark SQL 访问数据库的实践案例，还扩展了 Spark Streaming 来分析流数据，最后利用 Spark Streaming 和 Spark 批处理等实现了实时批处理兼顾的 Lambda 架构。

本书既包含了易于上手的逐步详细技术指引，也提供了深入浅出的丰富实践范例，学习时要求读者最好拥有 Java 或 Scala 语言的编程经验和 Hadoop 等大数据计算平台的基础知识。

在本书的翻译过程中，除张广骏之外，潘玉兰、张华锋、朱仁杰、潘玉芳、张广容、陈长滨、张广芮、白宸蜚等人也参与了翻译工作，在此一并表示感谢。

译 者

前　言

对于现代企业而言，处理过去 10～20 年的历史数据并进行分析以获得提升业务的洞见是当今最为热门的用例。

企业过去曾执迷于数据仓库的开发。通过这些数据仓库，企业努力从每个可能的数据源获取数据并存储下来，再利用各种商业智能工具对数据仓库中存储的数据进行分析。但是开发数据仓库是一个复杂、耗时和大开销的过程，需要相当程度的资金和时间投入。

Hadoop 及其生态系统的涌现无疑为海量大数据问题的处理提供了一种新的方法或架构，通过这种低成本、可伸缩的解决方案，过去需要数天时间处理的成 TB 数据将在几小时内被处理完毕。尽管有着这样的优势，在其他一些需要实时或准实时（如亚秒级服务等级协议 SLA）执行分析及获得业务洞见的应用场景中，Hadoop 还是面临着批处理性能方面的挑战。这类应用需求可称为实时分析（RTA）或准实时分析（NRTA），有时又被称为“快数据”，后者意味着做出准实时决策的能力，即要在有限的商务决策时间内提供卓有成效的数据支持。

为应对这些企业实时数据分析的应用场景，出现了一些高性能、易于使用的开源平台。Apache Storm 和 Apache Spark 是其中最为引人注目的代表性平台，能够为广大相关用户提供实时数据处理和分析功能。这两个项目都归属于 Apache 软件基金会。尽管有部分功能重叠，这两个工具平台仍保持着各自的特色和不同功能。

考虑到以上的大数据技术背景，本书结合实际用例介绍了应用 Apache Storm 和 Apache Spark 进行实时大数据分析的实现过程，为读者提供了快速设计、应用和部署实时分析所需的技术。

本书内容

第 1 章“大数据技术前景及分析平台”奠定了全书的知识背景，主要包括大数据前景的综述、大数据平台上采用的各种数据处理方法、进行数据分析所用的各种平台。本章也介绍了实时或准实时批量分布式处理海量数据的范式。此外，还涉及处理高速/高频数据读写任务的分布式数据库。

第 2 章“熟悉 Storm”介绍了实时/准实时数据处理框架 Apache Storm 的概念、架构及编程方法。这里涉及多种 Storm 的基本概念，诸如数据源（spouts）、数据流处理组件（bolts）、并行度（parallelism）等。本章还以丰富的应用场景及范例说明如何利用 Storm 进行实时大数据分析。

第 3 章“用 Storm 处理数据”着重于介绍 Apache Storm 中用于处理实时或准实时数据流的内部操作，如过滤（filters）、连接（joins）、聚合（aggregators）等。这里展示了 Storm 对 Apache Kafka、网络通信接口、文件系统等多种输入数据源的集成，最后利用 Storm JDBC 框架将处理过的数据保存起来。本章还提到 Storm 中多种企业关注的数据流处理环节，诸如可靠性、消息获取等。

第 4 章“Trident 概述和 Storm 性能优化”验证了实时或准实时事务数据的处理。这里介绍了实时处理框架 Trident，它主要用于处理事务数据。在此提到使用 Trident 处理事务应用场景的几种架构。这一章还提到多种概念和可用参数，进而探讨了它们对 Storm 框架与其任务的监测、优化以及性能调整诸方面的可用性。本章还涉及 LMAX、环形缓冲区、ZeroMQ 等 Storm 内部技术。

第 5 章“熟悉 Kinesis”提到了在云上可用的实时数据处理技术 Kinesis，此技术是亚马逊云计算平台 AWS 中的实时数据处理服务。这里先说明了 Kinesis 的架构和组成部分，接着用一个端到端的实时报警发生范例阐明了 Kinesis 的用法，其中使用到 KCL、KPL 等客户端库。

第 6 章“熟悉 Spark”介绍了 Apache Spark 的基础知识，其中包括 Spark 程序的高级架构和构建模块。这里先从 Spark 的纵览开始，接着提到了 Spark 在各种批处理和实时用户场景中的应用情况。这一章还深入讲到 Spark 的高级架构和各种组件。在本章的最后部分讨论了 Spark 集群的安装、配置以及第一个 Spark 任务的执行实现。

第 7 章“使用 RDD 编程”对 Spark RDD 进行了代码级的预排。这里说明了 RDD API 提供的各种编程操作支持，以便于使用者轻松实现数据转换和保存操作。在此还阐明了 Spark 对如 Apache Cassandra 这样的 NoSQL 数据库的集成。

第 8 章“Spark 的 SQL 查询引擎——Spark SQL”介绍了 Spark SQL，这是一个和 Spark 协同工作的 SQL 风格的编程接口，可以帮助读者将 Parquet 或 Hive 这样的数据集快速应用到工作中，并支持通过 DataFrame 或原始 SQL 语句构建查询。本章同时推荐了一些 Spark 数据库的最佳实践案例。

第 9 章“用 Spark Streaming 分析流数据”介绍了 Spark 的又一个扩展工具 Spark Streaming，用于抓取和处理实时或准实时的流数据。这里顺承着 Spark 架构简明扼要地描述了 Spark Streaming 中用于数据加载、转换、持久化等操作的各种应用编程接口。为达成

实时查询数据，本章将 Spark SQL 和 Spark Streaming 进行了深入集成。本章最后讨论了 Spark Streaming 任务部署和监测等方面的内容。

第 10 章“介绍 Lambda 架构”引领读者认识了新兴的 Lambda 架构，这个架构可以将实时和预计算的批量数据结合起来组成一个混合型的大数据处理平台，从其中获得对数据的准实时理解。本章采用了 Apache Spark 并讨论了 Lambda 架构在实际应用场景中的实现。

本书阅读基础

本书的读者最好拥有 Java 或 Scala 语言的编程经验，对 Apache Hadoop 等代表性分布式计算平台的基础知识亦有一定了解。

本书适用读者

本书主要面向应用开源技术进行实时分析应用和框架开发的大数据架构师、开发者及程序员群体。这些有实力的开发者阅读本书时可以运用 Java 或 Scala 语言的功底来进行高效的核心要素和应用编程实现。

本书会帮助读者直面不少大数据方面的难点及挑战。书里不但包括应用于实时/准实时流数据及高频采集数据处理分析的大量工具和技术，而且涵盖了 Apache Storm、Apache Spark、Kinesis 等各种工具和技术的内存分布式计算范式。

本书约定

本书应用了一些文本格式以区分不同类型的信息。以下是这些文本格式范例和含义说明。

文中的代码、数据库表名称、文件目录名称、文件名、文件扩展名、路径名、伪 URL、用户输入以及推特用户定位采用如下方式表示：

“The PATH variable should have the path to Python installation on your machine.”

代码块则通过下列方式设置：

```
public class Count implements CombinerAggregator {
    @Override
    public Long init(TridentTuple tuple) {
      return 1L;
    }
```

命令行输入和输出的显示方式如下所示：

```
> bin/kafka-console-producer.sh --broker-list localhost:9092 --topic test
```

 图标表示警告提醒或重要的概念。

 图标表示提示或相关操作技巧。

读者反馈

欢迎读者对本书反馈意见或建议，以便于我们进一步了解读者的阅读喜好。反馈意见对于我们十分重要，便于我方日后工作的改进。

读者可将这些反馈内容发送邮件到 feedback@packtpub.com，建议以书名作为邮件标题。

若读者针对某项技术具有专家级的见解，抑或计划撰写书籍或完善某部著作的出版工作，则可阅读 www.packtpub.com/authors 中的 author guide 一栏。

客户支持

感谢您购买本社出版图书，我们将竭诚对每一名读者提供周到的客户服务支持。

示例源码下载

读者可访问 http://www.packtpub.com 登录您的账户下载本书中的示例代码文件。无论以何种方式购买本书，都可以访问 http://www. packtpub.com/support，注册后相关文件会以电子邮件方式直接发送给您。

读者还可经由以下步骤下载源码文件：

（1）通过电子邮件加密码方式注册登录我们的网站。

（2）用鼠标切换上方的 Support（支持）标签页面。

（3）单击 Code Downloads & Errata（源码下载和勘误表）。

（4）在搜索框输入书名。

（5）在搜索结果列表中选择希望下载源码的图书项。

（6）在所购图书的下拉菜单中进行选择。

（7）单击 Code Download（源码下载）菜单。

文件下载到本地计算机之后，请使用下列软件的最新版本将文件内容解压到文件夹：

- Windows 操作系统下的 WinRAR 或 7-Zip 软件
- Mac 操作系统下的 Zipge 或 iZip 或 UnRarX 软件
- Linux 操作系统下的 7-Zip 或 Peazip 软件

勘误表

尽管我们努力争取做到尽善尽美，书中错误依然在所难免。如果读者发现谬误之处，无论是文字错误抑或是代码错误，都欢迎您不吝赐教。对于其他读者以及本书的再版工作，这将具有十分重要的意义。对此，读者可访问 http://www.packtpub.com/submit-errata，选取对应书籍，单击 Errata Submission Form 链接，并输入相关问题的详细内容。经确认后，输入内容将被提交至网站，或添加至现有勘误表中（位于该书籍的 Errata 部分）。

另外，读者还可访问 http://www.packtpub.com/books/content/support 查看之前的勘误表。在搜索框中输入书名后，所需信息将显示于 Errata 项中。

版权须知

一直以来，互联网上所有媒体的版权问题从未间断，Packt 出版社对此类问题异常重视。若读者在互联网上发现本书任何形式的非法副本，请及时告知网络地址或网站名称，我们将对此予以处理。

对于可疑的盗版资料链接，读者可将其通过邮件发送至 copyright@packtpub.com。

衷心感谢读者们对作者的爱护，这也有利于我们日后提供更为精彩的作品。

问题解答

若读者对本书有任何疑问，欢迎发送邮件至 questions@packtpub.com，我们将竭诚为您提供优质服务。

目　录

第 1 章　大数据技术前景及分析平台

大数据在突飞猛进的发展中已成为下一代数据存储、管理和分析方面最强大的计算范式。IT 巨头们实际上都已接纳了这种变化，而且格外重视大数据技术在业务中的应用。

代表性的数据存储和分布式处理平台 Hadoop 已然成熟并在应用中继续改进提升。目前不仅可以从大数据全局视野了解各种工具，还可以从大数据空间各个具体角度来审视特定技术的应用效果。

读者可以通过本章来熟悉大数据技术前景及分析平台。首先介绍了大数据的基础框架、处理模块组件及未来的发展。接着讨论了大数据近实时分析的需求及应用场景。

如下内容有助于理解大数据技术前景：

- 大数据基础框架
- 大数据生态系统的组件
- 分析框架
- 分布式批处理
- 分布式数据库（NoSQL）
- 实时数据处理及流数据处理

1.1　大数据的概念

大数据并不是一个突如其来的时兴科技词语，而是在厚积薄发中不断演变，时机到来时一下变得广为人知。传统数据库和数据仓库的统治地位本来看上去牢不可破，随着 Hadoop 等大数据技术的日趋成熟，这种情况到了终结的时候。

如今，在社会和经济的各个方面都会接触到海量的数据信息。举凡像生产制造、汽车、金融、能源、日用品、交通运输、安保、信息技术及网络等工业中都已积累大量行业数据信息，并且新数据还在源源不断地产生。大数据作为一种学科（抑或领域、概念、理论、思想）出现，通过对海量数据的存储、处理、分析获取智能见解，使得信息化及自动计算决策成为可能。这些决策全方位促进了经济领域的推广、增长、规划和预测等应用，这也是大数据像风暴一样席卷全世界的原因。