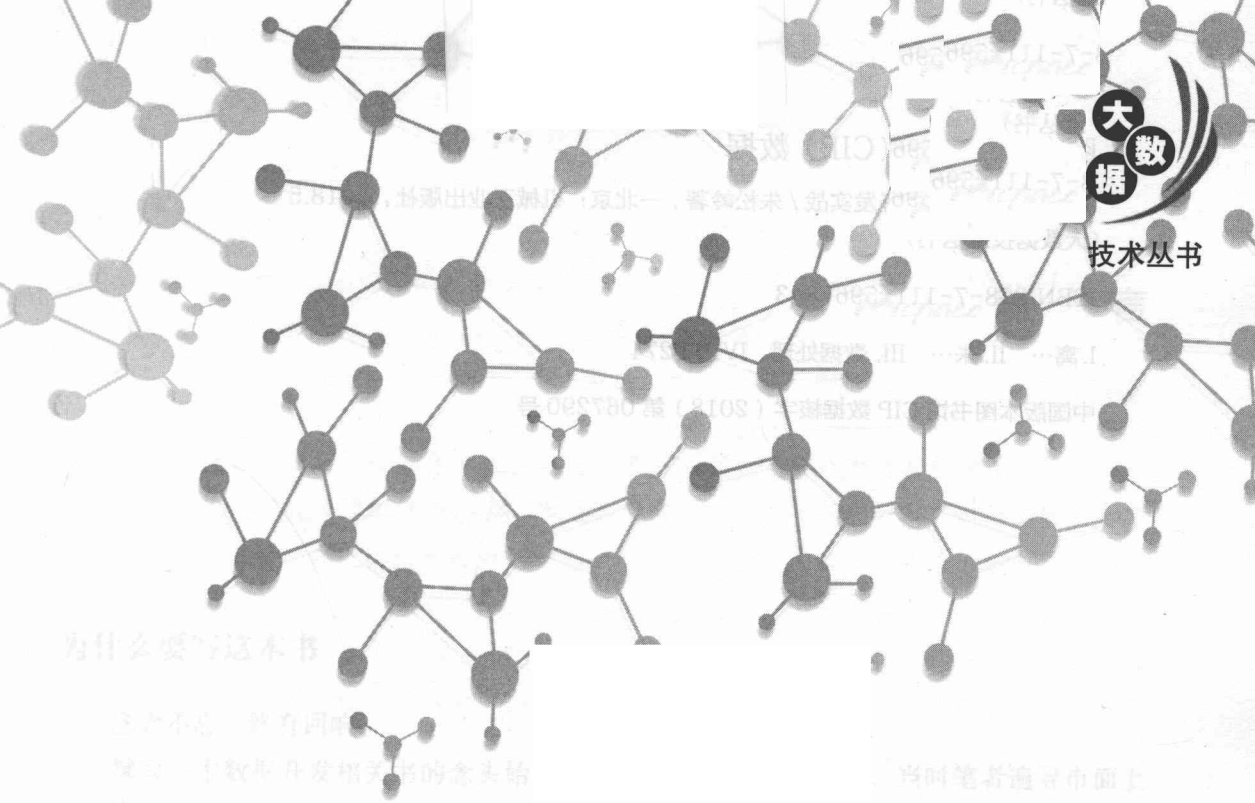
离线和实时大数据开发实战

朱松岭◎著

阿里巴巴大数据开发专家撰写，源于十余年工作实践，只讲实用有效的“招式”

庖丁解牛式讲解离线和实时开发平台架构、原理实现、开发示例，涵盖查询与优化、建模、数仓开发、流计算开发等核心技术

机械工业出版社
China Machine Press



大数据技术丛书

常州大学图书馆
藏书章

离线和实时大数据开发实战

朱松岭◎著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

离线和实时大数据开发实战 / 朱松岭著 . —北京：机械工业出版社，2018.5
（大数据技术丛书）

ISBN 978-7-111-59678-3

I. 离… II. 朱… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字（2018）第 067290 号

离线和实时大数据开发实战

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：高婧雅　　责任校对：殷　虹
印　　刷：北京诚信伟业印刷有限公司　　版　　次：2018 年 5 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：14.75
书　　号：ISBN 978-7-111-59678-3　　定　　价：59.00 元

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）88379426　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzit@hzbook.com

版权所有 · 侵权必究
封底无防伪标均为盗版
本书法律顾问：北京大成律师事务所　韩光 / 邹晓东

Preface 前　言

为什么要写这本书

念念不忘，终有回响。

撰写一本数据开发相关书的念头始于笔者学习数据知识的早期，当时笔者遍寻市面上所有的数据书籍，却没有发现一本系统化且从项目实践角度突出重点的数据开发书籍。

笔者非常理解某领域初学者的苦衷，对于他们来说，最重要的不是具体的 API、安装教程等，而是先找到该领域的知识图谱，有了它，就可按图索骥，有针对性地去学。

对于大数据技术来说，上述需求更甚。一方面，由于社区、商业甚至个人原因，大数据的技术可以说是五花八门、琳琅满目，初学者非常容易不知所措，不知从哪里下手。另一方面，从理论上来说，互联网上几乎可以查到所有的大数据技术，比如在百度上搜索、问知乎，但这些都是碎片化的知识，不成体系，初学者需要先建立自己的大数据知识架构，再进一步深入。

本书正是基于这样的初衷撰写的，旨在帮助和加快初学者建立大数据开发领域知识图谱的过程，带领初学者更快地了解这片领域，而无须花更长的时间自己去摸索。

当然，未来是 DT（Data Technology）时代，随着人工智能、大数据、云计算的崛起，未来数据将起到关键的作用，数据将成为如同水、电、煤一样的基础设施。但是，实际上目前数据的价值还远远没有得到充分的挖掘，如医疗数据、生物基因数据、交通物流数据、零售数据等。所以笔者非常希望本书能够对各个业务领域的业务分析人员、分析师、算法工程师等有所帮助，让他们更快地熟悉和掌握数据的加工处理知识与技巧，从而能够更好、更快地分析、挖掘和应用数据，让数据产生更多、更大的价值。

通过阅读本书，读者能建立自己的大数据开发知识体系和图谱，掌握数据开发的各种技术（包括有关概念、原理、架构以及实际的开发和优化技巧等），并能对实际项目中的数

据开发提供指导和参考。

大数据技术日新月异，由于篇幅和时间限制，书中仅讲述了当前主要和主流的数据相关技术，如果读者对大数据开发有兴趣，本书将是首选的入门读物。

本书特色

本书从实际项目实践出发，专注、完整、系统化地讲述数据开发技术，此处的数据开发技术包括离线数据处理技术、实时数据处理技术、数据开发优化、大数据建模、数据分层体系建设等。

我们处于一个信息过度的时代，互联网涵盖了人类有史以来的所有知识，浩如烟海。对大数据开发技术来说，更是如此。那么，大数据相关人员如何吸收、消化、应用和扩展自己的技术知识？如何把握相关的大数据技术深度和广度？深入到何种程度？涉猎到何种范围？

这是很有意思的问题。笔者认为最重要的是找到锚点，而本书的锚点就是数据开发技术。所以本书的另一个特点是以数据开发实战作为锚点，来组织、介绍各种数据开发技术，包括各种数据处理技术的深度和广度把握等。比如在离线数据处理中，目前事实的处理标准是 Hive，实际项目中开发者已经很少自己写 Hadoop MapReduce 程序来进行大数据处理，那是不是说 MapRedue 和 HDFS 就不需要掌握了呢？如果不是，又需要掌握到何种程度呢？笔者的答案是，对于 Hive 要精深掌握，包括其开发技巧和优化技巧等。MapReduce 要掌握执行原理和过程，而 MapRedue 和 HDFS 具体的读数据流程、写数据流程、错误处理、调度处理、I/O 操作、各种 API、管理运维等，站在数据开发的角度，这些都不是必须掌握的。

本书还有一个特点，就是专门讲述了实时数据处理的流计算 SQL。笔者认为，未来的实时处理技术的事实标准将会是 SQL，实际上这也是正在发生的现实。

读者对象

本书主要适合于以下读者，包含：

- ❑ 大数据开发工程师
- ❑ 大数据架构师
- ❑ 数据科学家

此为试读，需要完整PDF请访问：www.ertongbook.com

- ❑ 数据分析师
- ❑ 算法工程师
- ❑ 业务分析师
- ❑ 其他对数据感兴趣的人员

如何阅读本书

本书内容分为三篇，共 12 章。

第一篇为**数据大图和数据平台大图（第 1 章和第 2 章）**，主要站在全局的角度，基于数据、数据技术、数据相关从业者和角色、离线和实时数据平台架构等给出整体和大图形式的介绍。

第 1 章 站在数据的全局角度，对数据流程以及流程中涉及的主要数据技术进行介绍，还介绍了主要的数据从业者角色和他们的日常工作内容，使读者有个感性的认识。

第 2 章 是本书的纲领性章节，站在数据平台的角度，对离线和实时数据平台架构以及相关的各项技术进行介绍。同时给出数据技术的整体骨架，后续的各章将基于此骨架，具体详述各项技术。

第二篇为**离线数据开发：大数据开发的主战场（第 3 ～ 7 章）**，离线数据是目前整个数据开发的根本和基础，也是目前数据开发的主战场。这一部分详细介绍离线数据处理的各种技术。

第 3 章 详细介绍离线数据处理的技术基础 Hadoop MapReduce 和 HDFS。本章主要从执行原理和过程方面介绍此项技术，是第 4 章和第 5 章的基础。

第 4 章 详细介绍 Hive。Hive 是目前离线数据处理的主要工具和技术。本章主要介绍 Hive 的概念、原理、架构，并以执行图解的方式详细介绍其执行过程和机制。

第 5 章 详细介绍 Hive 的优化技术，包括数据倾斜的概念、join 无关的优化技巧、join 相关的优化技巧，尤其是大表及其 join 操作可能的优化方案等。

第 6 章 详细介绍数据的维度建模技术，包括维度建模的各种概念、维度表和事实表的设计以及大数据时代对维度建模的改良和优化等。

第 7 章 主要以虚构的某全国连锁零售超市 FutureRetailer 为例介绍逻辑数据仓库的构建，包括数据仓库的逻辑架构、分层、开发和命名规范等，还介绍了数据湖的新数据架构。

第三篇为**实时数据开发：大数据开发的未来（第 8 ～ 12 章）**，主要介绍实时数据处理的各项技术，包括 Storm、Spark Streaming、Flink、Beam 以及流计算 SQL 等。

第 8 章 详细介绍分布式流计算最早流行的 Storm 技术，包括原生 Storm 以及衍生的 Trident 框架。

第 9 章 主要介绍 Spark 生态的流数据处理解决方案 Spark Streaming，包括其基本原理介绍、基本 API、可靠性、性能调优、数据倾斜和反压机制等。

第 10 章 主要介绍流计算技术新贵 Flink 技术。Flink 兼顾数据处理的延迟与吞吐量，而且具有流计算框架应该具有的诸多数据特性，因此被广泛认可为下一代的流式处理引擎。

第 11 章 主要介绍 Google 力推的 Beam 技术。Beam 的设计目标就是统一离线批处理和实时流处理的编程范式，Beam 抽象出数据处理的通用处理范式 Beam Model，是流计算技术的核心和精华。

第 12 章 主要结合 Flink SQL 和阿里云 Stream SQL 介绍流计算 SQL，并以典型的几种实时开发场景为例进行实时数据开发实战。

勘误和支持

本书是笔者对大数据开发知识的“一孔之见”，囿于个人实践、经验以及时间关系，难免有偏颇和不足，书中也难免出现一些错误、不准确之处和个人的一些主观看法，恳请读者不吝赐教。你可以通过以下方式联系笔者。

- 微信号：yeshubert
- 微博：hubert_zhu
- 邮箱：493736841@qq.com

希望与大家共同交流、学习，共同促进数据技术和数据行业的发展，让数据发挥更大的价值。

致谢

首先非常感谢 Apache 基金会，在笔者撰写各个开源技术框架相关内容的过程中，Apache 官方文档提供了最全面、最深入、最准确的参考材料。

感谢互联网上无名的众多技术博客、文章撰写者，对于数据技术和生态的繁荣，我们所有人都是不可或缺的一分子。

感谢阿里巴巴公司智能服务事业部数据技术团队的全部同事，尤其薛奎、默岭、萧克、延春、钟雷、建帧、思民、宇轩、赛侠、紫豪、松坡、贾栩、丘少、茅客等，与他们的日

常交流让我受益颇多。

感谢机械工业出版社华章策划编辑高婧雅，从选题到定稿再到本书的出版，她提供了非常专业的指导和帮助。

特别致谢

特别感谢我的妻子李灿萍和我们的女儿六一，你们永远是我的力量源泉。

同时感谢我的父母和岳父岳母，有了你们的诸多照顾和支持，我才有时间和精力去完成额外的写作。

谨以此书献给我的家人，以及直接或间接让数据发挥价值的所有朋友们！

朱松岭（邦中）

目　录 *Contents*

第二篇 离线数据开发：大数据开发的主战场

此为试读，需要完整PDF请访问：www.ertongbook.com

第一篇 Part 1

数据大图和数据平台大图

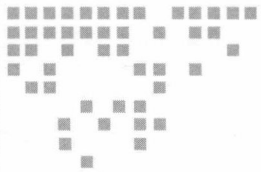
“不谋万世者，不足谋一时；不谋全局者，不足谋一域。”作为本书的开篇，本篇正是基于此考虑撰写的。本篇分为两章，主要站在全局的角度对数据、数据技术、数据相关从业者和角色、离线和实时数据平台架构等给出整体和大图形式的介绍。

本篇不会详细深入具体的各项数据技术内部，这些内容将交由第二篇和第三篇的各个具体章节来完成。

第 1 章为数据大图，主要从数据整体角度，结合数据从采集到消费的四大流程，对相关的数据技术和人进行介绍和刻画。

第 2 章为数据平台大图，主要从数据平台的角度对离线和实时数据平台架构以及相关的各项技术进行介绍。就业界数据平台的现状来讲，离线数据平台仍然是众多公司和组织的数据主战场，但是实时数据越来越重要，也越来越得到重视并被放在战略地位，可以说实时数据平台是数据平台的未来，未来也许将会颠覆离线数据平台。

第 2 章是本书的纲领，同时给出了数据技术的整体骨架，后续的第二篇和第三篇将基于此骨架，具体介绍各个数据开发技术和框架。



Chapter 1 第 1 章

数据大图

数据是原油，数据是生产资料，数据和技术驱动，人类正从 IT 时代走向 DT 时代，随着数据的战略性日渐得到认可，越来越多的公司、机构和组织，尤其是互联网公司，纷纷搭建了自己的数据平台。不管是基于开源技术自研、自建还是购买成熟的商业解决方案，不管是在私有的数据中心还是在公有云端，不管是自建团队还是服务外包，一个个数据平台纷纷被搭建，这些数据平台不但物理上承载了所有的数据资产，也成为数据开发工程师、数据分析师、算法工程师、业务分析人员和其他相关数据人员日常的工作平台和环境，可以说数据平台是一个公司、机构或组织内“看”数据和“用数据”的关键基础设施，已经像水电煤一样不可或缺，正是它们的存在才使得数据变现成为可能。

数据从产生到进入数据平台中被消费和使用，包含四大主要过程：数据产生、数据采集和传输、数据存储和管理以及数据应用，每个过程都需要很多相关数据技术支撑。了解这些关键环节和过程以及支撑它们的关键技术，对一个数据从业者来说，是基本的素养要求。因此本章首先对数据流程以及相应的主要数据技术进行介绍。

同时，本章也将介绍数据的主要从业者，包括平台开发运维工程师、数据开发工程师、数据分析师、算法工程师等，并对他们的基本工作职责和日常工作内容等进行介绍，使读者对数据相关的职位有基本的认识和了解。

1.1 数据流程

不管是时髦的大数据还是之前传统的数据仓库，不管是目前应用最为广泛的离线数据还是越来越得到重视的实时数据，其端到端流程都包含：数据产生、数据采集和传

输、数据存储处理、数据应用四大过程，具体的数据流程图及其包含的关键环节如图 1-1 所示。

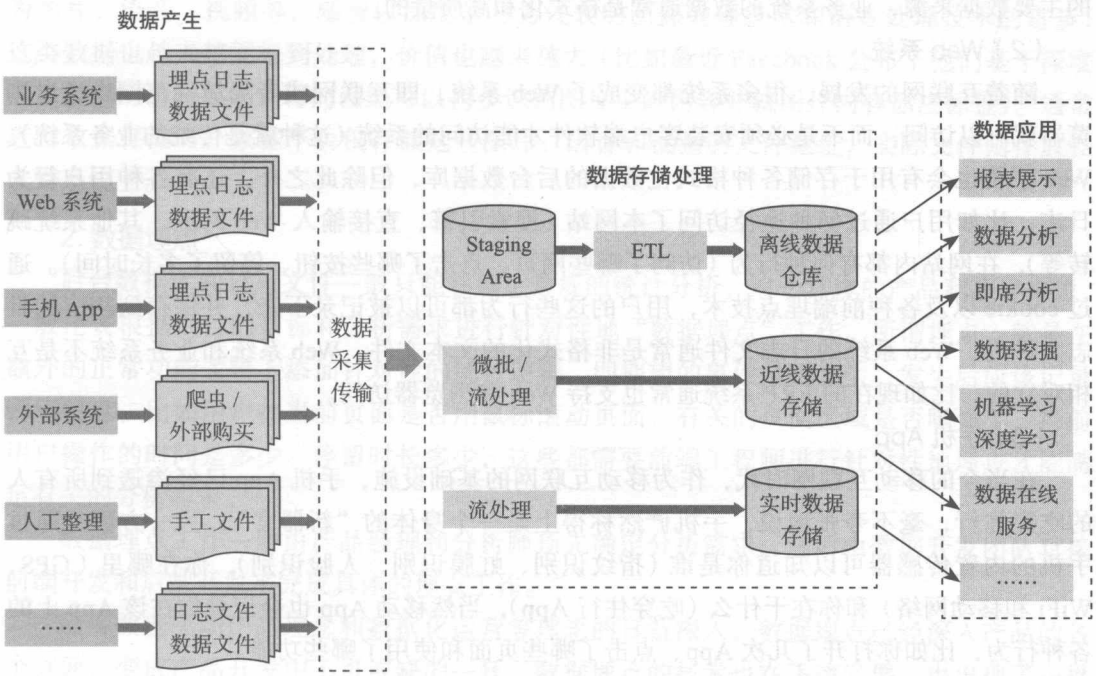


图 1-1 数据流程大图

下面详述图 1-1 所示的各个关键关节。

1.1.1 数据产生

数据产生是数据平台的源头，没有数据，所谓的大数据也无从谈起。所以首先要保证有数据。

随着近年来互联网和移动互联网的蓬勃发展，数据已经无处不在，毫无疑问，这是一个数据和信息爆炸的时代。所以，即使一个企业和个人没有数据，通过爬虫工具和系统的帮助，也可以从互联网上爬取到各种各样的公开数据。但是更多的、高质量的数据是爬取不到的，这些数据存在于各个公司、企业、政府机关和机构的系统内部。

1. 数据分类

根据源头系统的类型不同，我们可以把数据产生的来源分为以下几种。

（1）业务系统

业务系统指的是企业核心业务的或者企业内部人员使用的、保证企业正常运转的 IT 系统，比如超市的 POS 销售系统、订单 / 库存 / 供应链管理的 ERP 系统、客户关系管理的 CRM 系统、财务系统、各种行政系统等。不管何种系统，后台的数据一般都存在后台数据

库内。早期的大部分数据主要来源于这些业务系统的数据库，管理人员和业务运营人员查看的数据报表等基本来源于此。即便是目前，企业的业务系统依然是大部分公司数据平台的主要数据来源，业务系统的数据通常是格式化和高质量的。

（2）Web 系统

随着互联网的发展，很多系统都变成了 Web 系统，即互联网或者局域网范围内通过浏览器就可以访问，而不是必须安装客户端软件才能访问的系统（这种就是传统的业务系统）。Web 系统也会有用于存储各种格式化数据的后台数据库，但除此之外，还有各种用户行为日志，比如用户通过何种途径访问了本网站（搜索引擎、直接输入 Web 网址、其他系统跳转等），在网站内都有何种行为（访问了哪些网页、点击了哪些按钮、停留了多长时间）。通过 cookie 以及各种前端埋点技术，用户的这些行为都可以被记录下来，并保存到相应的日志文件内。Web 系统的日志文件通常是非格式化的文本文件。Web 系统和业务系统不是互相对立的，比如现在的 ERP 系统通常也支持 Web 和浏览器访问。

（3）手机 App

在当今的移动互联网时代，作为移动互联网的基础设施，手机 App 已经渗透到所有人的吃穿住行，毫不夸张地说，手机俨然称得上是一个身体的“新器官”；另一方面，通过手机的内置传感器可以知道你是谁（指纹识别、虹膜识别、人脸识别），你在哪里（GPS、WiFi 和移动网络）和你在干什么（吃穿住行 App），当然移动 App 也会记录你在该 App 上的各种行为，比如你打开了几次 App、点击了哪些页面和使用了哪些功能。

（4）外部系统

很多企业除了自己内部的数据，还会通过爬虫工具与系统来爬取竞争对手的数据和其他各种公开的数据，此外企业有时也会购买外部的数据，以作为内部数据的有益补充。

（5）人工整理

尽管通过各种内部和外部系统可以获取到各种数据，但有些数据是无法直接由系统处理的，必须通过人工输入和整理才能得到，典型的例子是年代较久的各类凭证、记录等，通过 OCR 识别软件可以自动识别从而简化和加快这类工作。

以上从 IT 系统的类型介绍了数据产生的源头，从数据的本身特征，数据还可以分为以下几类。

（1）结构化数据

结构化数据的格式非常规范，典型的例子是数据库中的数据，这些数据有几个字段，每个字段的类型（数字、日期还是文本）和长度等都是非常明确的，这类数据通常比较容易管理维护，查询、分析和展示也最为容易和方便。

（2）半结构化数据

半结构化数据的格式较为规范，比如网站的日志文件。这类数据如果是固定格式的或者固定分隔符分隔的，解析会比较容易；如果字段数不固定、包含嵌套的数据、数据以 XML/JSON 等格式存储等，解析处理可能会相对比较麻烦和繁琐。

（3）非结构化数据

非结构化数据主要指非文本型的、没有标准格式的、无法直接处理的数据，典型代表为图片、语音、视频等，随着以深度学习为代表的图像处理技术和语音处理技术的进步，这类数据也越来越能得到处理，价值也越来越大（比如最近 Facebook 公布了他们基于深度学习技术的研究成果，目前已经可以初步识别图片中的内容并就图片内容标注标签）。通常进行数据存储时，数据平台仅存储这些图片、语音和视频的文件地址，实际文件则存放在文件系统上。

2. 数据埋点

后台数据库和日志文件一般只能够满足常规的统计分析，对于具体的产品和项目来说，一般还要根据项目的目标和分析需求进行针对性地“数据埋点”工作。所谓埋点，就是在额外的正常功能逻辑上添加针对性的统计逻辑，即期望的事件是否发生，发生后应该记录哪些信息，比如用户在当前页面是否用鼠标滚动页面、有关的页面区域是否曝光了、当前用户操作的时间是多少、停留时长多少，这些都需要前端工程师进行针对性地埋点才能满足有关的分析需求。

数据埋点工作一般由产品经理和分析师预先确定分析需求，然后由数据开发团队对接前端开发和后端开发来完成具体的埋点工作。

随着数据驱动产品理念和数据化运营等理念的日益深入，数据埋点已经深入产品的各个方面，变成产品开发中不可或缺的一环。数据埋点的技术也在飞速发展，也出现了一批专业的数据分析服务提供商（如国外的 Mixpanle 和国内的神策分析等），尽管这些公司提供专门的 SDK 可以通用化和简化数据埋点工作，但是很多时候在具体的产品和项目实践中，还是必须进行专门的前端埋点和后端埋点才可以满足数据分析与使用需求。

1.1.2 数据采集和传输

业务系统、Web 系统、手机 App 等产生的数据文件、日志文件和埋点日志等分散于各个系统与服务器上，必须通过数据采集传输工具和系统的帮助，才能汇总到一个集中的区域进行关联和分析。

在大数据时代，数据量的大当然重要，但更重要的是数据的关联，不同来源数据的结合往往能产生 1+1 > 2 甚至 1+1 > 10 的效果。另一个非常关键的点是数据的时效性，随着时间的流失，数据的价值会大打折扣，只有在最恰当的时间捕捉到用户的需求，才会是商机，否则只能是错失时机。试想一下，如果中午用户在搜索“西餐厅”，只有在这一刻对用户推送西餐厅推荐才是有效的，数分钟、数十分钟、数小时后的推送效果和价值将逐步递减。而实现这些必须借助于专业的数据采集传输工具和系统的帮助。

过去传统的数据采集和传输工具一般都是离线的，随着移动互联网的发展以及各种个性化推荐系统的兴起，实时的数据采集和传输工具越来越得到重视，可以毫不夸张地说，

此为试读，需要完整PDF请访问：www.ertongbook.com