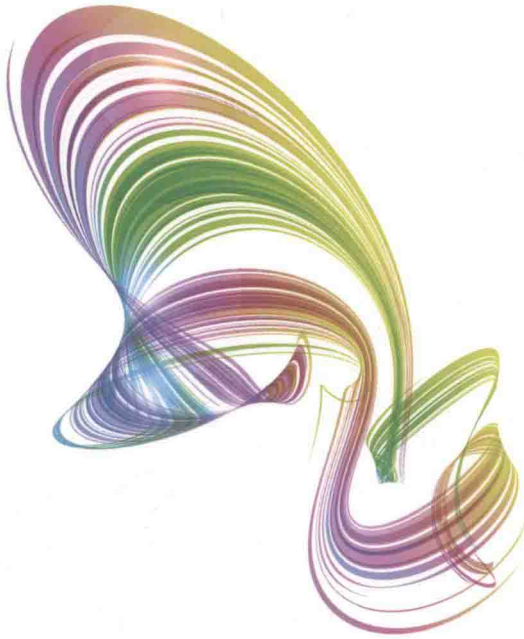
HZ BOOKS
华章IT

从数据、基础平台、分析方法、行业应用4个维度，以场景化方式讲解数据从获取、预处理、挖掘、建模、结论分析与展现到系统应用的流程，以及机器学习的重要技术

三位金融领域的大数据专家近10年行业实战经验总结，包含大量行业解决方案和案例，并公开源代码

大数据技术丛书



Analytics and Applications with Business Cases

Guide to Big Data & Machine Learning

大数据与机器学习

实践方法与行业案例

陈春宝 阙子扬 钟飞◎著

机械工业出版社
China Machine Press

大数据技术丛书

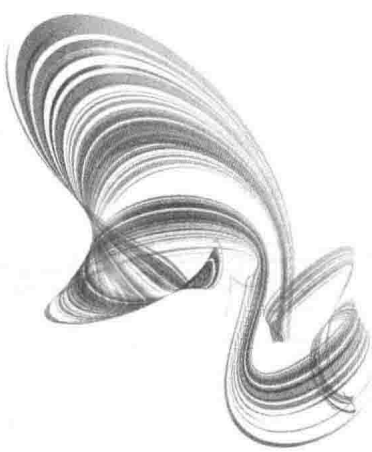
Analytics and Applications with Business Cases

Guide to Big Data & Machine Learning

大数据与机器学习

实践方法与行业案例

陈春宝 阙子扬 钟飞◎著



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

大数据与机器学习：实践方法与行业案例 / 陈春宝，阙子扬，钟飞著 . —北京：机械工业出版社，2017.1
（大数据技术丛书）

ISBN 978-7-111-55680-0

I. 大… II. ① 陈… ② 阙… ③ 钟… III. ① 数据处理 ② 机器学习 IV. ① TP274 ② TP181

中国版本图书馆 CIP 数据核字（2016）第 323978 号

大数据与机器学习：实践方法与行业案例

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：李　艺
印　　刷：北京文昌阁彩色印刷有限责任公司　　版　　次：2017 年 1 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：19.25
书　　号：ISBN 978-7-111-55680-0　　定　　价：69.00 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

Preface 前　言

不畏浮云遮望眼，只缘身在最高层。

——王安石《登飞来峰》

数据科学家＝统计学家＋程序员＋讲故事的人＋艺术家

——Shlomo Aragmon

本书的创作初衷

大数据方面的书籍可谓琳琅满目，有的讲解理论，有的介绍方法，有的传播理念。但是，大数据从业人员（如数据工程师、数据分析师、业务分析师、算法设计师等）应该掌握哪些知识与技能，如何应用数据解决现实的业务问题呢？恐怕最能给出答案的还是实际的数据从业者。为此，三位作者基于近 10 年的数据分析与应用经验，融合各自在商业银行、互联网金融和电商领域的切身体验，寓理论于实战，选取多个详实的案例，站在企业实际应用的角度介绍数据分析应用过程并公布源代码，并最终形成本书。本书对于读者开展数据分析工作能够提供直接帮助，为有志于在大数据领域发展的读者启航。

本书特点

本书有三大特点。其一，内容全面，覆盖大数据生态中的数据、基础平台、分析方法和应用四个领域，对数据应用从业务需求、数据准备、数据分析、挖掘建模、演示报告、成果应用等全流程进行了详细阐述；其二，以业务场景为主线，精选银行和互联网方面最具代表性的案例，站在数据消费者和分析师的角度，身临其境地介绍了数据如何产生价值，

寓理论于实战，让读者能知其所以然；其三，写作手法上遵循大道至简原则，用浅显的语言介绍复杂的数据分析应用过程，归纳数据分析师乃至数据科学家应该修炼的要点，既关注技术细节，又不拖泥带水，能为读者提供直接帮助。

本书定位

本书既可作为数据分析与商业分析人员的入门指引和案头工具，亦可为统计学、计算机科学、市场营销等专业研究生拓宽视野。

源代码下载

对书中源代码感兴趣的读者，可与作者联系（邮箱：64346837@qq.com）。

Contents 目　录

前言

第一部分　数据与平台篇

第 1 章　数据与数据平台 ……3

1.1　数据的基本形态 ……4
1.1.1　数据环境与数据形态 ……4
1.1.2　生产数据 ……5
1.1.3　原始数据 ……5
1.1.4　分析数据 ……6
1.2　数据平台 ……7
1.2.1　数据仓库平台 ……9
1.2.2　大数据平台 ……13
1.2.3　MPP 数据库 ……22
1.2.4　NoSQL 数据库 ……23
1.3　应用系统 ……24
1.4　本章小结 ……25

第 2 章　数据体系 ……26

2.1　数据闭环 ……27
2.2　数据缓冲区 ……28
2.2.1　系统解耦 ……29
2.2.2　批量导出 ……31
2.2.3　FTP 传输 ……40
2.2.4　批量导入 ……42
2.3　ETL ……49
2.3.1　ETL 工具 ……50
2.3.2　ETL 作业 ……52
2.4　作业调度 ……56
2.5　监控和预警 ……56
2.5.1　使用监控工具进行监控 ……57
2.5.2　使用 BI 工具进行监控 ……57
2.6　本章小结 ……57

第 3 章　实战：打造数据闭环 ……59

3.1　数据缓冲区的基本规则 ……60
3.1.1　文件存储规则 ……61
3.1.2　文件命名规则 ……61
3.1.3　文件清理规则 ……62
3.2　自动加载的流程 ……62
3.2.1　扫描文件 ……63
3.2.2　下载文件 ……64

3.2.3 解压文件……65
3.2.4 加载文件……65
3.3 自动加载程序的数据库设计……66
3.3.1 数据文件信息表……67
3.3.2 数据文件状态表……68
3.3.3 加载配置信息表……69
3.3.4 数据缓冲区信息表……70
3.3.5 目标服务器表……70
3.4 自动加载程序的多线程实现……71
3.4.1 ScanFiles……72
3.4.2 DownLoadAndUnZip……75
3.4.3 LoadToHive……77
3.4.4 LoadToOracle……78
3.4.5 自动加载程序的部署架构…79
3.4.6 程序的维护和优化……80
3.5 本章小结……80

第二部分 分 析 篇

第 4 章 数据预处理……83
4.1 数据表的预处理……84
4.2 变量的预处理……85
4.2.1 缺失值的处理……85
4.2.2 极值的处理……90
4.3 变量的设计……91
4.3.1 暴力衍生……91
4.3.2 交叉升维……92
4.4 变量筛选……95
4.4.1 筛选显著变量……95
4.4.2 剔除共线性……96
4.5 本章小结……100

第 5 章 聚类，简单易用的客户细分方法……101
5.1 从客户细分说起……102
5.1.1 为什么要做客户细分……102
5.1.2 怎么做客户细分……103
5.1.3 聚类分析，无监督的客户细分方法……107
5.2 谱系聚类……107
5.2.1 基本步骤……107
5.2.2 案例：公司客户差异化服务……110
5.2.3 谱系聚类方法的题外话…115
5.3 K-means 算法……116
5.3.1 基本步骤……116
5.3.2 案例：电商卖家细分……117
5.3.3 K-means 算法的题外话…121
5.4 本章小结……121

第 6 章 关联规则挖掘，发现产品加载和交叉销售机会……122
6.1 销售的真谛：让客户买得更多…123
6.1.1 案例：电商的生意经……123
6.1.2 案例：富国银行的“商店”经营模式……124
6.1.3 案例总结……125
6.2 交叉销售……126
6.2.1 为什么要做交叉销售……126
6.2.2 怎么做交叉销售……126
6.3 关联规则挖掘，发现交叉销售机会……128
6.3.1 Apriori 算法……129

6.3.2 Apriori 算法的主要指标…129
6.3.3 Apriori 算法的基本步骤…131
6.4 案例：信用卡产品交叉销售…131
6.4.1 准备数据…132
6.4.2 SAS 实现…132
6.4.3 结果分析…133
6.4.4 序列关联分析…136
6.4.5 结果应用…137
6.5 本章小结…138

第 7 章 社交网络分析，从“关系”的角度分析问题…139

7.1 先看几张美轮美奂的图片…140
7.2 社交网络分析方法…142
7.2.1 定义…142
7.2.2 应用场景…142
7.2.3 网络识别算法…143
7.3 案例：电商通过订单数据识别供应链…144
7.3.1 供应链及供应链金融…144
7.3.2 识别核心企业及其上下游关系…144
7.3.3 分析结果的业务应用…149
7.4 案例：P2P 投资风险防范…151
7.4.1 案例背景…151
7.4.2 防范方法…152
7.5 本章小结…153

第 8 章 线性回归，预测客户价值…155

8.1 数值预测…156
8.2 回归与拟合…157
8.2.1 回归就是拟合…157
8.2.2 在 Excel 中添加趋势线预测…158
8.3 案例：信用卡客户价值预测…159
8.3.1 确定预测目标…159
8.3.2 准备建模数据…161
8.3.3 模型拟合…163
8.3.4 模型评估…165
8.4 基于客户价值分层的业务策略…167
8.5 本章小结…167

第 9 章 Logistic 回归，精准营销的主要支撑算法…169

9.1 大数据时代的精准营销…170
9.1.1 精准营销…170
9.1.2 基于大数据的精准营销模式…171
9.1.3 如何做到精准…172
9.2 Logistic 回归算法介绍…173
9.2.1 算法原理…173
9.2.2 关键步骤…174
9.3 案例：信用卡消费信贷产品的精准营销…176
9.3.1 案例背景…176
9.3.2 数据准备…176
9.3.3 数据预处理…180
9.3.4 建模…182
9.3.5 模型评估…185
9.4 预测模型的应用与评估…189
9.5 本章小结…189

第 10 章 决策树类算法，反欺诈模型“专家”…191

10.1 决策树，重要的分类器…191

10.2 决策树的关键思想 ……192
10.2.1 理财客户画像案例背景…192
10.2.2 关键思想一：递归划分…194
10.2.3 关键思想二：剪枝……197
10.3 案例：电商盗卡交易风险识别…198
10.3.1 案例背景 ……198
10.3.2 以 SAS 实现 ……199
10.3.3 以 Clementine 实现 ……201
10.3.4 以 R 实现……204
10.4 随机森林……208
10.5 本章小结……209

第 11 章 数据可视化，是分析更是设计……210

11.1 数据演示之道 ……210
11.1.1 好“色”之图……211
11.1.2 版式有形 ……212
11.1.3 数据发声 ……214
11.2 个性化地图 ……215
11.2.1 案例背景：存款增长率指标展示 ……215
11.2.2 获取地理位置的经纬度数据 ……216
11.2.3 定制地图背景和图标……217
11.2.4 生成地图 ……220
11.3 文本分析……222
11.3.1 案例：电商的客户评价分析 ……222
11.3.2 分词 ……223
11.3.3 词云制作 ……224
11.3.4 情感分析 ……225
11.4 本章小结……227

第三部分 应 用 篇

第 12 章 标签系统……231

12.1 认识标签系统 ……231
12.2 标签系统的设计 ……233
12.2.1 标签系统的层次结构……233
12.2.2 标签系统的更新规则……233
12.2.3 机器学习模型转化为标签 ……235
12.3 标签系统的实现 ……236
12.3.1 标签映射表 ……237
12.3.2 标签系统的前端实现……238
12.3.3 标签系统的数据后端实现 ……238
12.3.4 标签系统的在线接口实现 ……242
12.4 本章小结……242

第 13 章 数据自助营销平台……244

13.1 数据自助营销平台的价值所在…245
13.1.1 自动化营销，提升工作效率 ……245
13.1.2 降低营销成本，提升用户体验 ……247
13.1.3 个性化营销，提升响应率 ……248
13.1.4 统一管理，便于效果追踪 ……249
13.2 数据自助营销平台的实现原则…249
13.2.1 数据营销活动的节点……249
13.2.2 数据自助营销平台的基础：标签系统……251

13.2.3 数据自助营销平台的批量任务……252
13.2.4 实时数据营销……254
13.3 数据自助营销平台的场景实例…254
13.3.1 客户生命周期管理……254
13.3.2 用卡激励计划……257
13.4 本章小结……260

第 14 章 基于 Mahout 的个性化推荐系统……261

14.1 Mahout 的推荐引擎……262
14.1.1 Mahout 的安装配置……262
14.1.2 Mahout 的使用方式……263
14.1.3 协同过滤算法……264
14.1.4 Mahout 的推荐引擎……265
14.2 规模与效率……268
14.2.1 Mahout 推荐算法的适用范围……268
14.2.2 通过分布式解决规模和效率的问题……270
14.3 实现一个推荐系统……275
14.3.1 系统框架……275
14.3.2 推荐系统的刷新……276
14.3.3 部署一个可用的推荐系统……276
14.4 本章小结……280

第 15 章 图计算与社会网络……281

15.1 社会网络和属性图……282
15.2 Spark GraphX 与 Neo4j……283
15.2.1 Scala 编程语言……284
15.2.2 Cypher 查询语言……285
15.3 使用 Spark GraphX 和 Neo4j 处理社会网络……286
15.3.1 背景说明……286
15.3.2 数据准备……286
15.3.3 Spark GraphX 处理原始网络……287
15.3.4 Neo4j 交互式查询分析…291
15.3.5 更多的应用场景……295
15.4 本章小结……296

第一部分 Part 1

数据与平台篇
(Data & Infrastructures)

迟序之数，非出神怪，有形可检，有数可推。

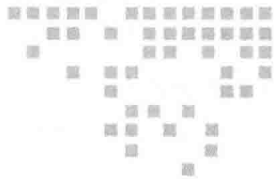
——祖冲之

数学是知识的工具，亦是其他知识工具的泉源。所有研究顺序和度量的科学均和数学有关。

——笛卡儿

对于大部分非计算机专业出身的分析人员和业务人员来说，数据库领域的专业术语简直让人抓狂，非要搞得那么高深吗？大可不必。

数据科学家是数据的应用者，以最大限度来提炼数据价值为目的，不必像数据仓库开发者那样对数据的存储、结构以及数据仓库的内生技术一清二楚，但应该站在找到数据、拼接数据、使用数据的角度，大体了解数据的分布、处理逻辑，以便为分析快速地准备素材。



第 1 章 Chapter 1
数据与数据平台

合抱之木，生于毫末；九层之台，起于垒土；千里之行，始于足下。

——《老子》

世界的本质是数。

——毕达哥拉斯

数据时时刻刻在伴随着我们的工作和生活，就像空气围绕着我们一样，以致于我们常常忽略了它的存在。但如果你立志做一个崇尚数据的人，静下心来像科学家研究空气一样研究数据，就会发现数据为我们认知事物打开了一条全新的途径。

可以通过数据认识自身：人类全身的肌肉大约有 639 块，由 60 亿条肌纤维构成，而起着重要作用的大脑则由 140 亿个细胞构成。

可以通过数据描述我们的工作：周一上午 10:00~11:30 召开会议，讨论公司第三季度的销售目标。

可以通过数据描述我们的行为：花费 6088 元购买一台 iPhone 6 手机，中速游泳 60 分钟消耗约 1000 千卡的热量。

还可以通过数据认识我们所处的环境：现在时间是 14:00，当前温度为 28℃，上个月的 CPI 同比上涨 1.4%，蔬菜和水果价格上涨了 6.7%。

甚至可以通过数据认识遥不可及的物体：太阳直径为 1 392 000 公里，表面温度达 57 809 开尔文……数据让我们认识了世间万物，那么我们该如何认识数据本身呢？

数据的本质是一个十分深奥且宽泛的话题，甚至带有哲学的意味。作为技术类书籍，

本书不尝试从哲学的角度研究数据，而是基于实践，从思维和技术手段出发来认识、理解、处理并分析周围的数据。为了更加具体，本书研究的数据定位于企业经营数据。

本章首先将从数据的基本形态入手，介绍企业中数据的来源和表现形态；然后介绍与之相关的数据平台，并简单介绍两类应用系统。在着手处理数据之前，让我们先对数据有一个清晰的认识。

1.1 数据的基本形态

我们不是自然科学家，但是可以借鉴自然科学的思路来看待数据问题。问题是数据具有形态吗？虽然数据并不具有固态、液态或气态等形态，但是可以根据需要为数据定义属于自己的专属形态。

一旦为数据赋予了恰当的形态，并在一定范围内（比如在一个公司内部）达成共识，形成对数据的系统化认识，就可以基于这些数据形态提出相应的管理和使用方案，提升数据的效率和价值。

一般情况下，对于企业经营中产生的数据，可以定义为三种形态：生产数据、原始数据和分析数据。这些数据形态的产生，是基于企业应用系统所在的生产环境和分析环境而存在的，在深入讨论数据形态之前，我们先来熟悉一下数据所在的环境。

1.1.1 数据环境与数据形态

数据环境是指数据存储、处理、转换所处的物理环境，常见的数据环境有生产环境、分析环境和测试环境。

生产环境是生产应用系统实时运行所在的环境，而生产应用系统则是一系列业务逻辑的组合。我们可以把生产环境想象成人的身体，生产应用系统就是人体中的各个系统（消化系统、呼吸系统等），业务逻辑则是这些系统中的“经络”，而数据便是运行于经络之中的“气血”。数据从“经络”中的一个“穴位”流转到另一个“穴位”，并在“流淌”中发生变化，所以，生产环境中的数据是“动态变换”的数据，我们称为生产数据。

分析环境是与生产环境物理解耦的一个数据环境。在生产环境中，由于数据总是处于不停变化中，这些数据的变化将直接反映为业务逻辑结果的变化，因此不应该尝试在生产环境中对数据进行分析处理。为了不影响生产环境的正常运行，需要将生产环境中的“动态”数据的快照保存下来（例如每日凌晨将时间戳为昨日的数据导出），这些数据快照是“静态”的，我们称为分析数据，保存分析数据的物理环境即我们所说的分析环境。

在实际中，还有另外一个环境，即测试环境。测试环境中的数据也是独立于生产环境和分析环境的，由于测试环境的数据通常不是有效的数据，因此本书不关注测试环境的数据。

至此，根据数据所处的环境，我们将数据定义为三种基本形态：生产数据、原始数据和分析数据。

生产数据存在于生产环境之中，分析数据存在于分析环境之中。此外，在生产数据和分析数据之间，还存在一种过渡形态的数据，即原始数据。图 1-1 展示了数据环境及其对应的数据形态。

图 1-1　数据环境及其对应的数据形态

注意，图 1-1 中所示的原始数据，既不属于生产环境也不属于分析环境，这意味着它不直接用于生产，也不直接用于分析。原始数据作为生产数据到分析数据的中间形态存在，本书随后的章节将进一步讨论原始数据的相关问题。

1.1.2　生产数据

生产数据是应用系统中在线使用的数据，它可能是一个生产系统的生产环境数据库中的数据，比如在一个 P2P 借贷平台的系统中，用户进行注册、充值、投资等行为产生的数据将被记录到生产环境数据库中，这些数据即为生产数据。

生产数据是动态的，会随着业务应用的变化而变化，比如用户账户余额数据，会随着用户投资的变化而变化。任何存在于生产环境中的数据，都在时刻准备发生改变，只不过有些生产数据的变化频率特别低而已，比如用户的年龄信息。

正常情况下，数据分析师并不直接接触生产数据，但需要注意的是，有些生产数据是从分析数据而来的。比如用户标签数据，它本身是从分析数据构建的，属于分析数据。但这些标签数据一旦用于应用系统，例如作为推荐系统的底层数据，即转化为生产数据，这种情况下，应用系统输出结果的质量将受到分析数据的直接影响。

1.1.3　原始数据

由于生产数据是动态的数据，而过去大量的分析工具和分析方法很难处理动态改变的数据（流处理已经改变了这种情况）。为了在不影响生产应用系统的情况下分析和处理这些数据，我们需要将这些数据从生产系统解耦。

从生产系统解耦的数据即是原始数据。数据解耦的过程一般包括数据脱敏（如屏蔽电话

号码、去除住宅详细信息等）、信息筛选（抛弃不需要的字段）、批量导出（如在 T 日凌晨批量导出 T-1 日的交易明细数据）等。

原始数据可以以多种形式存在，例如存储在生产数据库备库中，或者以文本文件的格式存放在文件服务器中。无论以何种形式存在，原始数据都应该独立于生产环境和分析环境，这可以避免分析环境对生产环境的干扰。

存放这些原始数据的地方，我们称为数据缓冲区。在很多企业中，数据缓冲区和原始数据并未得到足够重视，它们大多为了前期的方便，省略了数据缓冲区和原始数据形态，就像图 1-2 所示的那样。

显然，数据直连的方式让生产环境直接暴露在分析环境之上，两者之间的 ETL（Extract-Transform-Load）过程将对双方的性能造成影响。随着数据量的增加，这可能会带来数据管理和应用上的灾难。

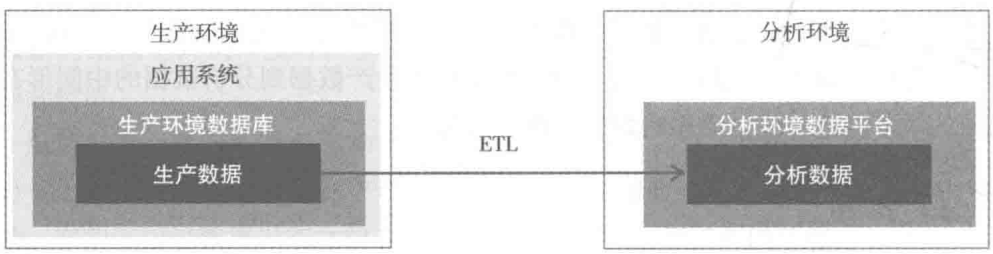


图 1-2 省略数据缓冲区的数据直连

本书极力推荐图 1-1 所示的方式，虽然它比图 1-2 要复杂，但在数据管理和可扩展性方面有非常大的优势。后面的第 2.1 节中会深入讨论该问题。

1.1.4 分析数据

分析数据是对原始数据进行属性筛选、标准统一之后，使用优化存储的方式存放于分析环境中的数据。从原始数据到分析数据的关键步骤在于 ETL 过程。

比如，原始数据中的一张表 A 可能包含 100 个字段，经过 ETL 之后，得到了一个包含 45 个字段的表 B，其中的日期格式进行了统一，且滤除了一些特殊字符，并将表 B 存放于分析环境数据平台的关系数据库 Oracle 中。这样，原始数据中的表 A 完成了属性筛选和标准统一（日期格式），转换成了分析数据表 B。

另一种需要标准统一的情景根源于原始数据本身的多样性。由于原始数据来源于不同的生产应用系统，其数据格式及字段含义均存在差异。例如，原始数据存放的格式可能有 Windows 文本、Linux 文本、主机格式文本、数据库文件等多种形式；字段含义上的差异则更加多样，比如，由原始数据文件 A 中性别字段使用 1 表示男性、2 表示女性，而原始数

据文件 B 中性别字段使用 M 表示男性、F 表示女性。通过标准统一，可以约定所有的分析数据统一使用 1 表示男性、2 表示女性。数据统一可为数据分析和数据应用铺平道路。

经过 ETL 之后的分析数据，为了进一步提高存储效率和读取效率，需要使用技术手段进行存储优化，比如创建索引、进行分区、分表存储、使用大数据平台等。

通过对原始数据的提炼和优化，分析数据具有了信息集中、标准统一、分析效率高等特点，便于数据进一步的分析和应用。

分析数据需要依托数据平台而存在，数据平台的性能对其上的数据分析和应用有决定性影响。数据平台是分析环境的基础，在随后的“数据平台”章节中，我们将详细介绍。

1.2　数据平台

数据平台是存放分析数据的平台，也是支持大多数数据分析和数据挖掘应用的底层平台，它使用了统一的数据清洗与处理规则，因而可以保证从基础平台上输出的数据内容是一致的。

传统的数据平台基本等同于大家熟悉的“数据仓库”，但互联网浪潮让人们对数据采集、存储和应用提出了越来越高的要求，传统数据仓库平台独力难支，因此“现代化”的数据平台是多种数据库产品的融合。图 1-3 是一个精简化的现代数据平台架构图。

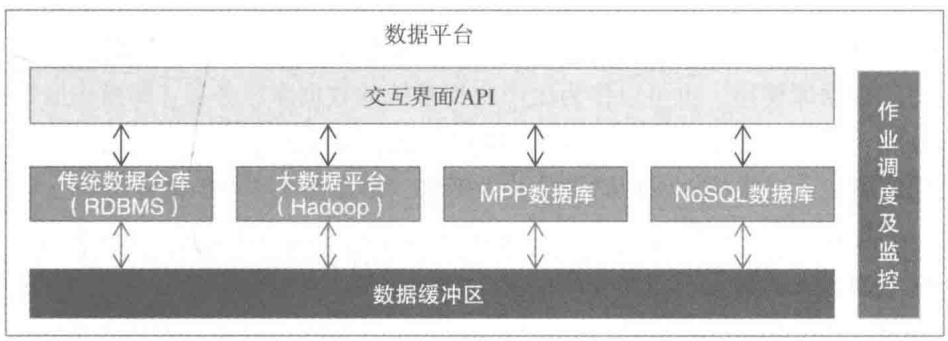


图 1-3　数据平台架构示意图

现代的数据平台融合了传统数据仓库、大数据平台、MPP 数据库、NoSQL 数据库等多种数据产品，这些数据库产品之间互为补充，组成统一的数据平台。

从传统的关系型数据库开始，数据库产品逐渐细分，这些细分产品在特定场景中比传统的关系型数据库表现出了更好的性能。图 1-4 展示了一些主流的数据库产品，注意到有很多数据库产品是“跨界”产品，例如，Oracle 同时属于关系型、分析型、操作型三类数据库。