信息科学与技术丛书

王 飞 编著

数据架构与商业智能

- 数据架构理论与实践
- 大数据架构
- 数据治理
- 商业智能架构理论与实践
- 商业智能——数据仓库架构
- 商业智能——ODS 数据架构
- 商业智能——数据集市架构
- 金融行业数据架构与商业智能案例
- 电力行业数据架构与商业智能案例

机械工业出版社
CHINA MACHINE PRESS

信息科学技术丛书

数据架构与商业智能

王　飞　编著

机 械 工 业 出 版 社

本书是《商业智能深入浅出》一书的姊妹篇，数据架构、商业智能、数据治理和大数据技术是本书的核心。本书共13章，主要内容包括：企业架构总体规划、数据架构现状分析、数据架构目标规划、数据架构案例、大数据架构与实践，数据治理体系、商业智能架构理论、商业智能架构实践、商业智能—数据仓库架构和案例、商业智能—ODS数据架构和案例、商业智能—数据集市架构和案例等。

本书的读者对象包括：公司管理者、IT架构咨询顾问、数据架构师、系统分析师、商业智能架构师以及相关技术爱好者。

图书在版编目（CIP）数据

数据架构与商业智能/王飞编著. —北京：机械工业出版社，2014.10
（信息科学与技术丛书）
ISBN 978-7-111-50289-0

Ⅰ.① 数… Ⅱ.① 王… Ⅲ.① 企业管理-应用软件 Ⅳ.① F270.7

中国版本图书馆CIP数据核字（2015）第104036号

机械工业出版社（北京市百万庄大街22号 邮政编码 100037）
策划编辑：丁 诚 责任校对：张艳霞
责任编辑：丁 诚
责任印制：乔 宇

保定市中画美凯印刷有限公司印刷

2015年6月第1版·第1次印刷
184mm×260mm·22印张·546千字
0001-3000册
标准书号：ISBN 978-7-111-50289-0
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

电话服务
服务咨询热线：(010)88379833
读者购书热线：(010)88379649

网络服务
机 工 官 网：www.cmpbook.com
机 工 官 博：weibo.com/cmp1952
教育服务网：www.cmpedu.com
金 书 网：www.golden-book.com

前　言

本书是《商业智能深入浅出》的姊妹版，数据架构、商业智能、数据治理和大数据技术是本书的核心。

为什么本书将数据架构和商业智能放在一起？本书为什么穿插着大数据和数据治理方面的内容？

传统的商业智能系统是围绕模型设计、数据采集、加工、联机分析和报表生成而设计的，目的是提高企业的运营效率，增强企业的竞争力和领导者的决策能力。而数据架构关注的是数据的分布、流转和数据分类等内容，目的是通过对数据采集、加工、对外服务和数据模型的设计，提高数据处理和加工的效率，提升数据采集的灵活性。

如何建立一个灵活、松耦合、高性能的数据架构规划体系，是很多企业和金融机构必须重视的问题。经过多年的信息化实践，很多企业和金融机构已经逐渐认识到，系统应该具备多渠道数据采集能力、历史与趋势分析能力。数据架构规划在信息化过程中起着非常重要的作用，通过数据架构规划可以推动企业信息化的进程，保证企业通过使用数据，提供更好的产品和服务，降低成本和控制风险，促进企业经营战略的实现，提升企业的核心竞争力。

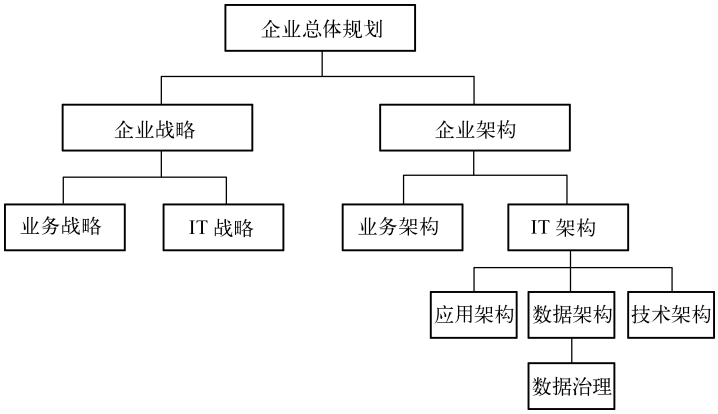
由于激烈的市场竞争和业务的快速发展，很多企业迫切需要改变运营模式，但是由于数据模型的不统一，数据分散，不能共享，严重制约了企业的发展，它们已经充分认识到数据是核心资产，正是这个原因 IT 人员需要了解数据架构方面的知识。数据架构是基础，而商业智能是在数据架构基础之上建立起来的一种解决方案。它们是相辅相成、融汇贯通的，两者之间有相通的地方，又有不同的分析视角和重点。

随着数据采集范围的不断扩大，使得文档、视频等半结构化和非结构化的数据逐渐成为很多企业主要的数据源。我们可以这样说，80% 的数据可能都来自于非结构化数据。包括：图像、音频、微博、网帖、电子邮件等信息。特别是对于商业银行，坐拥大量非结构化数据却未能更好地创造业务价值。对于商业银行来说，大数据更是机遇，客户在不断地与银行交易和交互过程中，会创造出各种类型的数据，这也为商业银行实时或者准实时的数据分析提供了便利，可以对客户进行有针对性的营销，所以，大数据技术也是本书的核心内容之一，穿插在各个章节当中。另外，为了提升数据架构各个层次的管控及其协作能力，也需要相关人员理解数据治理方面的知识，所以本书也穿插着相关内容。

本书的亮点是什么？

本书试图利用公式般的架构推导过程，以企业总体规划为主线，先从企业战略、企业架构出发，逐步细化到业务战略、IT 战略、业务架构和 IT 架构，再细化到应用架构、数据架构、技术架构和数据治理的过程，如下图所示。而商业智能可以看成是帮助用户对自身业务经营做出明智决策的解决方案之一，也可以看作是 IT 战略的一部分。企业 IT 架构的目的是为所有的解决方案提供 IT 支持。最后利用数据架构的方法论讲解关于商业智能的数据模型设计、数据的分布、流转等内容。

这种公式般的推导过程，会让读者真正理解架构的核心思想和方法论，知其然，亦知其所以然，同时可以帮助读者将书中的架构思路和方法应用到具体的项目当中去。



阅读本书应该重点关注哪些内容？

“乱花渐欲迷人眼”，我们不要拘泥于对具体概念的理解，而更应该看重对架构方法和思路的理解，例如，如何对某企业的数据架构现状进行分析，分析的方法和思路是什么；如何对该企业的目标数据架构进行规划，规划的重点和步骤是什么；……。

为什么写这本书？

与本书类似的书籍在国内图书类市场中基本上是一个空白，但是数据架构师的职位在 IT 企业中却越来越受到重视，出现了“喷井”式的局面，数据架构师的理论水平和项目经验也需要达到一定的高度，他们需要掌握数据架构、商业智能、大数据和数据治理方面的知识。

目前现状是商业智能图书不仅小众，在某种程度上甚至可以说是一个珍稀品种。讲解商业智能架构方面知识的书籍更是少之又少，而本书除了讲解企业架构、业务架构、数据架构等方面的知识外，还讲解了商业智能领域的架构知识，更是从企业整体规划的角度去分析商业智能领域的应用，包括围绕商业智能的数据架构等内容。

本书的读者对象有哪些？

本书的读者定位为公司管理者、IT 架构咨询顾问、数据架构师、系统分析师、商业智能架构师以及有志向涉足 IT 架构设计和咨询顾问工作的人们，希望大家都能从本书中获益。

本书编写历时整整一年，其间经历了喜悦、聒噪、痛苦和彷徨，心情是复杂的。如今，伴随着本书最终成稿，复杂的心情烟消云散，自己甚至还有一点成就感。在这里要感谢帮助我完成此书的所有人。

感谢公司的同事，他们以各种方式为本书的编写做出了重要的贡献，感谢他们的技术支持和帮助。

最后，也是最重要的，我要感谢母亲（张丽华）、父亲（王贵林），他们倾注了父母无尽的爱，感谢他们对我的培养和无微不至的照顾，同时对于本书的出版给予了我不懈的支持，还要感谢岳父（丁一贤）、岳母（赵桂荣），书中同样凝聚了他们的心血和付出。感谢二叔（王玉奎），他的鼓励激发了我写作的热情。感谢辛苦的妻子（丁玲玲）和心爱的女儿（王预萱）。他们是我最大的精神支柱，如果没有他们的辛劳和付出，我很难想象能完成这本书的创作。

虽然本人在编著过程中尽了最大努力，但是由于本人的水平和时间有限，本书可能存在不足之处，敬请广大同行和读者批评指正。

作　者

目　　录

第 1 章　企业架构总体规划

本章目标

通过本章的学习，应该理解的内容包括：企业总体架构规划包含哪些内容？关于 IT 战略、业务战略、业务架构、数据架构、应用架构和技术架构的定义是什么？同时我们应该掌握数据架构规划、应用架构规划、技术架构规划的方法论。通过学习，读者应对企业总体规划、企业战略、企业架构、业务架构、应用架构、技术架构和数据架构有一个整体性的认识。

学习本章，读者将掌握：

- 企业架构总体规划的概念
- 企业的总体规划包含哪些内容
- 什么是企业战略
- 什么是企业 IT 战略
- 企业架构和企业战略之间的关系
- 什么是业务架构
- 什么是 IT 架构
- 业务架构和 IT 架构之间的关系
- 数据架构规划的方法论
- 应用架构规划的方法论
- 技术架构规划的方法论

1.1　企业总体架构规划基础

1.1.1　企业总体架构规划概念

一、企业总体架构规划定义？

企业总体架构规划是从全局出发，解决现存问题，同时满足现实需求和适应未来发展的需要，有效地对资源进行管控，加强 IT 技术实力，并且指明企业的经营方向和发展目标，对企业远景发展轨迹进行全面规划。

我们可以建立起对企业总体规划的感性认识，把总体规划看成是对城市的战略规划和具有全局性、长远性的建设规划等内容。同时遵循从实际出发，正确处理好各种关系的原则，使局部建设和整体发展能够稳步前行，近期建设和远期规划可以相互支持。

举例来说，城市的总体规划主要关注一个城市的定位、发展方向、功能区域和基础公共设施等方面。

如图 1–1 所示，假设城市的战略规划是建立一个人文都市，打造区域一体化共赢战略，新型城市化、城乡统筹与美好城乡建设战略和交通引领发展战略，这是对城市的发展方向的

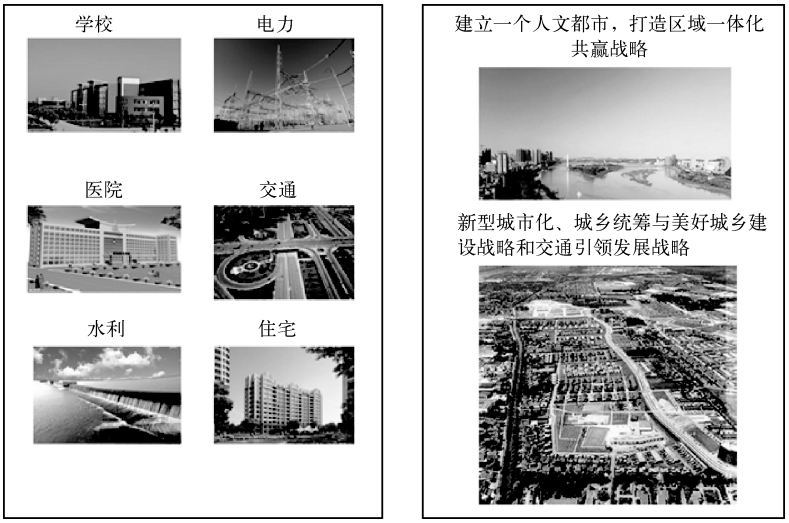


图 1–1　城市总体规划

定位。

城市的建设规划包括水利规划，电网规划，建筑和小区的布局，道路交通，煤气等基础设施的规划。

企业总体规划类似于城市总体规划，包含了企业战略、企业架构和实施解决方案等。

二、企业总体规划包含的内容

企业的总体规划包括企业战略、企业架构和实施解决方案等内容。

如图 1–2 所示，企业战略包含业务战略和 IT 战略，是对企业业务发展方向和 IT 发展方向的描述。它们都属于企业宏观的管理范畴，与城市的战略规划类似。企业架构规划包括业务架构和 IT 架构，是连接企业战略和实施解决方案的核心纽带，类似于对城市的建设规划。

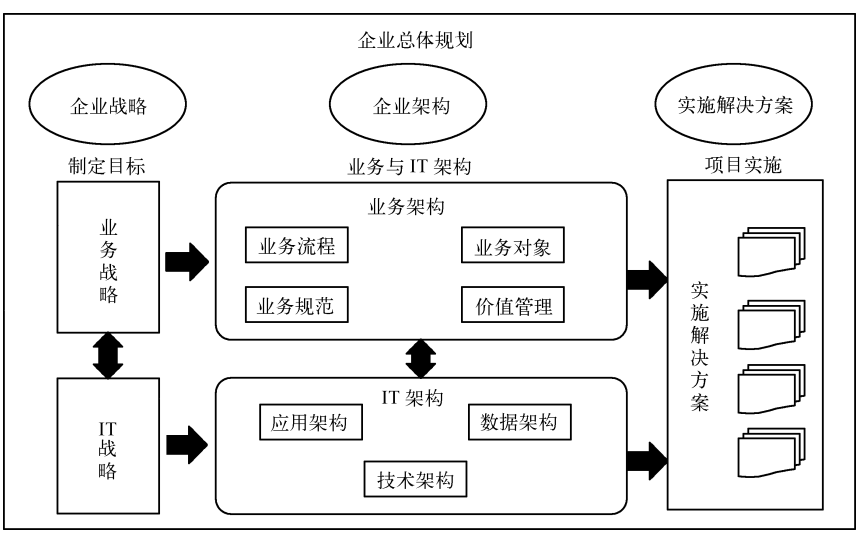


图 1–2　企业的总体规划

其中业务架构包含了对业务流程、业务对象、业务规范和价值管理的描述。IT 架构包含了对应用架构、数据架构和技术架构的描述，而数据架构是本书的核心和重点。

企业具体的实施解决方案是在 IT 架构满足企业战略的基础上，通过数据架构、应用架构和技术架构指导企业具体实施的过程。

1.1.2 企业战略

首先了解一下什么是企业战略。美国 90% 的企业家认为："最占时间、最为重要、最为困难的事就是制定战略规划"，据相关机构统计，发达国家的企业领导一年中要花大约 40% 的时间去研究企业战略。

一、企业战略的定义

企业战略是对企业发展目标，包括达成目标的方法和途径的总体谋划。企业战略的实质就是企业的发展方向和定位，如果企业的战略目标不明确，定位和发展方向不清楚，企业的中层管理人员和普通员工就很难领悟企业高层的战略意图和任务实质。

上面这种状况会导致企业大部分成员丧失方向感，个人的努力和发展方向不明确，就造成了企业无论是技术路线、服务方向，还是组织架构、企业文化等诸多方面，都会产生价值冲突。

企业战略的作用就是企业能够运筹帷幄，根据自身的资源和环境选择合适的经营发展方向，它是一个长远、持续的发展过程，具有一定的稳定性。

例如，企业战略可以包括：企业的信息化战略、竞争战略、营销战略、技术开发战略、人才培养战略等方面，它们都是从不同的维度去描述企业整体性、长期性和基本性的问题，都属于企业战略的范畴。如果企业是一艘船，那么企业的战略就是航海图，引领企业到达目标。

二、企业战略的特征

企业战略属于企业的宏观管理范畴，具有指导性、长远性、系统性、风险性、全局性和竞争性等主要特征。

1）指导性

企业的战略明确了企业的经营方针和远景发展目标，在企业的生产和管理活动中起着指导作用。

2）全局性

企业战略具有全局性，通过对政治、经济、文化以及周边经营环境的深入分析，并且结合自身条件，从系统全局的角度对企业的发展进行全面规划。

3）长远性

企业战略基于企业长期生存和长远发展的需要，确立企业的战略方向和远景目标。企业战略是一个长期、持续的过程，具有一定的稳定性。

4）系统性

企业战略属于决策层的战略，企业的经营方针、投资规模、经营方向和发展目标是企业战略的核心部分。企业战略围绕着发展目标设立各个阶段的经营策略，并且构成一个个环环相扣的企业战略体系。

5）风险性

企业的战略决策具有一定的风险性，如果经过深入的市场研究，客观地设立远景目标，并且资源调配使用得当，制定的企业战略就会起到促进的作用。反之，战略制定出现偏差，就会为企业带来相应的高风险。

6）竞争性

企业战略需要考虑各种的内外环境，明确自身的发展优势，改善相应的经营模式，增强企业的竞争力，只有这样才能在市场竞争中处于领先地位，保证企业长远健康的发展。

一般来说，企业战略包括业务战略和 IT 战略。

1）业务战略

企业的业务战略是指企业拥有的所有资产，通过多种方式进行有效的运营，以实现利润的最大化和资本的增值。它强调了企业在各自生产领域中的发展之道和发展方向，包括如何创造价值，并且以更好的服务去满足客户，这是企业业务战略的核心和重点。

2）IT 战略

企业的 IT 战略是指在充分研究企业发展愿景、业务策略和管理的基础上，形成信息系统的远景、组成架构、逻辑关系等内容，以支撑企业战略目标的实现。从功能划分的角度来看，IT 战略是一类独立的战略，为了明确未来 IT 的发展定位和战略目标，可以从应用系统建设、信息治理、基础设施、IT 管理体系、IT 队伍建设等几个方面进行全面规划。

IT 战略的实质就是关于信息系统功能目标及其实现的总体规划。

IT 战略的目的是指导系统的建设，通过明确相应的优化机制、保障规划和工作计划，并且根据外部环境的变化，不断地修改 IT 战略规划，以适应未来业务发展的需要。IT 战略是保证信息化建设全面性、前瞻性的重要手段之一。

1.1.3 什么是企业架构

一、企业架构的概念

企业总体规划包括企业的战略、企业架构和企业具体的实施解决方案。企业架构又包括业务架构和 IT 架构，本书重点关注的是企业架构中的数据架构部分。我们先了解一下什么是企业架构？企业架构和企业战略的关系是什么？

关于企业架构，不少的学术研究机构、标准组织和大厂商，都给出了各自的定义。

（1）微软公司的定义

企业架构是对一个公司的核心业务流程和 IT 能力的组织逻辑，通过一组原理、政策和技术选择来获得，以实现公司运营模型的业务标准化和集成需求。

（2）IBM 公司的定义

企业架构是记录企业内所有的信息系统，系统之间的相互关系以及系统如何完成企业使命的蓝图。

（3）Zachman 的定义

企业架构是构成组织的所有关键元素和关系的综合描述。企业架构框架（EAF）是个描述企业架构方法的蓝图。

二、企业架构的实质

企业架构实质上就是对企业多角度的一种描述，它反映了企业的业务流程、技术的组织和安排，是对企业关键性业务和技术的整体性描述。

如果我们把企业当做一栋建筑，信息技术就是一些建筑材料，在建造的过程中，应该根据建筑的功能定位并且结合现有的资源进行总体的架构设计，用架构来指导建造的过程。其中对建筑的功能定位类似于企业的战略，对建筑的总体架构设计类似于企业架构。其实“架构”一词最早来源于建筑行业，它描绘了事物的本质结构和内在规律，例如城市需要城市架构。

企业架构先从企业战略出发去梳理业务架构，然后进一步分析和规划 IT 架构，通过对企业架构的分析，将企业的业务战略、业务流程紧密结合起来，为企业描绘一个业务、信息、技术互动的蓝图。企业架构实质上就是企业的全景图，从战略、愿景，到业务、IT 等各个方面展示企业的结构和内部关系，从而指导企业开展信息化建设，最终实现业务和 IT 的融合。

举例来说，修建一栋房子，需要进行很多的架构设计工作，首先要进行外部的效果设计，当客户满意之后，再进行下内部设计，以及配套的线路、上下水管等方面的规划。同样，在进行企业架构设计的时候，也需要像房屋架构设计一样从不同的层次去描述企业的特征，如图 1–3 所示。

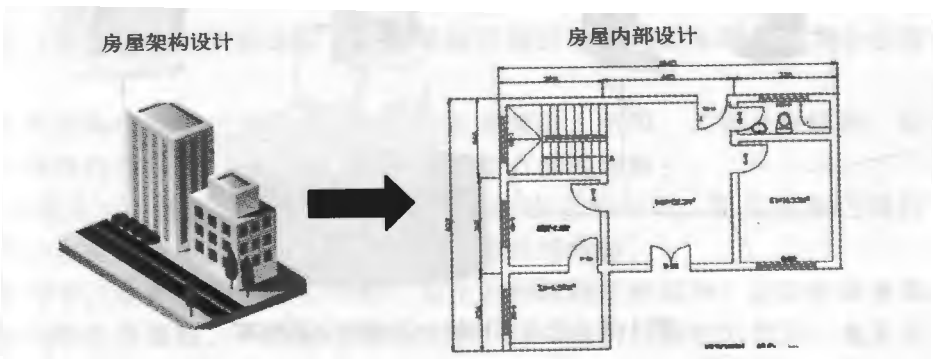


图 1–3　企业架构的形象比喻

企业架构是对真实世界中的企业的业务流程和 IT 设施的抽象，主要包括企业组织、职能、业务流程、IT 系统、数据、网络部署等的完整的、一体化的描述。企业架构反映了企业业务的状况，并体现了业务与 IT 的映射关系，明确各类 IT 基础设施对业务的支撑关系。企业架构就像城市的“总体规划蓝图”，在它的指导下，各个 IT 系统的建设得以有序进行。归根结底，企业架构的目的是将跨企业的、零散的业务流程优化成一个集成的环境，同时帮助企业执行业务战略及 IT 战略规划。

如图 1–4 所示，缺乏企业架构的 IT 系统犹如一个个的“竖井”结构，各个部门难以保持信息的一致性。

企业架构统一关键的企业数据，确保跨部门之间信息的一致性，保证了数据的完整性和准确性，如图 1–5 所示。

为了满足中国人民银行或者中国银监会的监管要求，增强核心竞争力并满足现实需求，很多金融机构也在进行企业架构的建设。

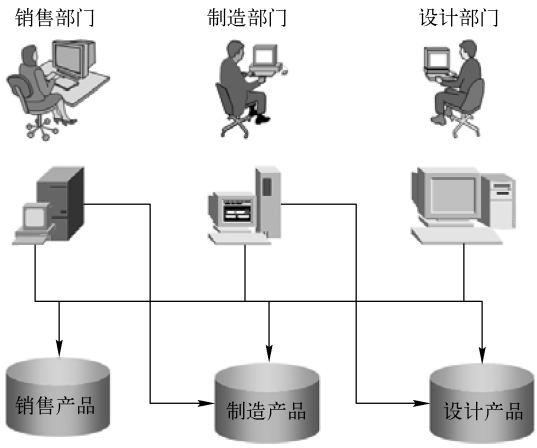


图 1–4 “竖井”架构

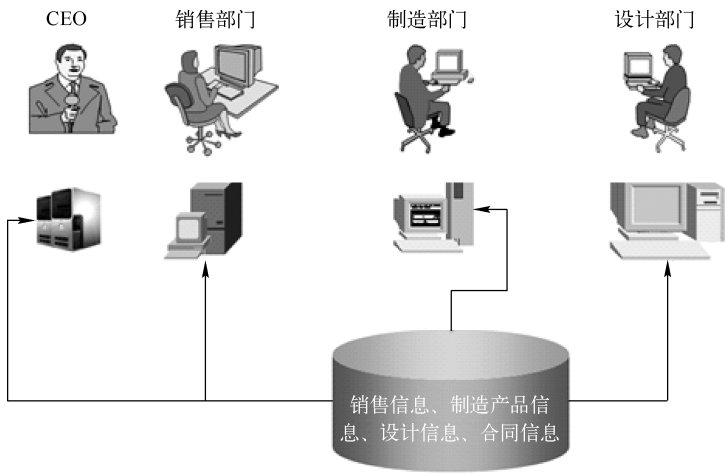


图 1–5 企业架构统一关键的企业数据

● 监管要求

由中国银监会颁布的《中国银行业信息科技十二五发展规划监管指导意见》中已经说明了“信息科技规划要与业务战略保持一致，以业务架构为基础，科学设计应用架构、数据架构和基础架构”，这说明政府监管部门已经越来越重视对规划的要求，同时也提高了监管力度。

● 增强核心竞争力

企业架构可以保证 IT 系统能够快速响应市场需求，使系统设计灵活、先进，具备良好的扩展性。同时 IT 与业务的有效融合，有力地提升了企业核心竞争力，并且支持未来业务和规模的扩张。

● 满足现实需求

企业架构可以帮助企业统一各类概念与术语，梳理现有系统，提取可重用的 IT 资产，加快积累，有效降低应用的开发成本，提高设计、开发效率和质量。

三、企业架构的价值

企业架构的价值可以分为有形价值和无形价值。

企业架构的有形价值体现在以下几个方面：

1）有效利用现有的架构，缩短系统开发和部署的时间，构建灵活的系统环境。

2）减少系统的重复建设，节约并且降低系统设计和开发的成本。

3）有效利用现有资源，减少设计和开发人员的学习周期。

企业架构的无形价值体现在以下几个方面：

1）有效达成业务人员和 IT 技术人员之间的共识。

2）加强业务人员和技术人员的沟通。

3）保证信息的集中，增加知识的积累。

四、企业架构的组成

企业架构的过程实质上就是对现实世界中企业的业务流程和 IT 设施抽象的过程。它反映了企业的业务流程和 IT 架构之间的关系。

一般来说，企业架构包括业务架构和 IT 架构。我们先了解一下什么是业务架构？

1. 业务架构

一个优秀的架构师和咨询顾问，不在于他有多厉害的技术手段，重要的是他对业务的理解有多深。通常来说，业务架构可以作为 IT 架构的输入部分。广义的业务架构包括产品、销售、财务、人力资源、客户服务等企业核心的业务功能和职责。并且将企业战略转化成企业运营的目标和形式，同时明确相关人员角色、企业资源、IT 资源和服务是如何协调和部署的。我们可以认为由企业战略决定了业务架构的模式，同时业务架构又是企业战略实现的手段之一。

狭义的业务架构包含了企业运营活动中的业务策略、组织、关键业务流程、组织架构以及人员组织结构等内容。我们对业务架构有以下两方面的理解：

① 业务架构是对业务规划的一种描述，主要解决业务布局，以及业务之间的关系，包括制定什么样的业务策略、建立什么样的机制和流程等内容。

在企业架构中，业务架构是核心内容，是企业相对稳定的部分，企业在业务架构的基础上可以建立相应的业务流程，不断满足市场需求，可以做到差异化的竞争。业务架构决定了 IT 架构的内容，同时 IT 架构又推动了业务架构的规划，它们是相互支持和促进的关系。

② 业务架构定义了企业如何创造价值以及企业内部的协作关系。它描述了企业如何满足客户需求，如何进行市场竞争，如何达成与其他企业之间的合作关系，如何建立相应的业务运营体系和绩效考核等内容。

业务架构是基于企业战略的，它决定了企业各组成部分是如何运转的。同时业务架构建立了企业战略和日常运营活动中的关联关系，它是连接企业战略和具体项目实施的一座桥梁，通过业务架构的支持，达到企业战略中预先设定的战略目标。

举例来说，假设企业的战略目标是将成本降低 10%，要实现该目标，就需要对现有的运营机制进行改进，可以通过在线自助服务减少人力成本，或者是优化现有的业务流程，提升运营效率。一般来说，日常运作的组织、业务流程和 IT 运营系统都应该在业务架构的框架下运转，如果没有业务架构，就会导致运营与企业战略方向的脱节，使每个业务环节存在缺乏统一调度等问题。

2. IT 架构

IT 架构是对企业系统的 IT 规划，是建立企业信息化系统的综合性的蓝图，IT 架构可以

帮助企业获得最优的投资回报，同时实现业务和技术接口之间的标准化，保证企业运营和企业战略之间的一致性。

IT 架构又承担了 IT 战略与 IT 项目实施、执行的桥梁作用，它主要包含应用架构、数据架构和技术架构。

IT 架构主要解决以下问题：

- 提供明确的技术解决方案，和企业的战略目标保持一致。
- 保证业务需求和技术支持之间转换的高效性，实现企业资源的最优配置。

IT 架构的原则：

（1）法律法规遵循原则

系统的建设应该符合相关法律法规的要求，如一些行业法规要求、信息安全要求等方面。

（2）架构及标准遵循原则

对于未来系统的建设，应该遵循架构及标准的原则。例如，技术解决方案、功能范围等方面需要和企业业务战略保持一致。

（3）数据整合原则

如果存在多个数据源，特别是在数据处理、存储和数据服务过程中有相同的部分，抽象出来形成统一的数据管理模块。

（4）资产重用原则

在系统建设过程中涉及的所有设备、软件或者组件，都需要进行管理，特别需要考虑在未来系统架构过程中这些资产的重用性，从而降低系统建设的成本。

（5）灵活高效原则

系统的架构需要满足一定的灵活性，以适应外部环境和业务需求的变化。同时，要能够保证系统处理数据的高效性，以满足客户的各种需求。

IT 架构的作用：

- 理解 IT 的价值。帮助企业高层理解 IT 的价值，为企业未来的发展提供信息化支持。
- 构建灵活的环境。有效利用现有的资源和已有的架构，缩短部署和开发的时间，构建灵活的环境。
- 降低成本。减少系统重复建设，降低系统建设成本。
- 规避各种风险。
- 有效地促进业务和 IT 之间的融合。
- 加强沟通。加强业务人员和 IT 人员的沟通，建立共同交流的平台。

IT 架构包含应用架构、数据架构和技术架构。下面分别进行描述。

五、应用架构

1. 什么是应用架构？

应用架构是对实现业务能力、支撑业务发展的应用功能结构化的描述方法。

系统的应用架构可以从功能和应用两个不同的视觉角度描述系统各个组件的构成以及组件之间的关系。功能组件模型侧重于业务功能，而应用组件模型则侧重于应用系统设计。

应用架构是业务架构和技术架构之间的“桥梁”，如图 1–6 所示。

图 1-6　应用架构是“桥梁”

2. 应用架构的目标

- 为业务发展和业务战略的实现提供有力的架构支撑和保障。
- 提供对业务架构的应用支撑。
- 描述应用系统的实现方式。
- 描述应用系统间的交互关系。
- 描述应用与核心业务的对应关系。

3. 应用架构的原则

应用架构的原则主要包括业务前瞻性、应用企业化、系统平台化、系统整合化和适度松耦合。

- 业务前瞻性

能够适应未来业务发展的要求，保证应用架构对于企业战略和业务架构的支持能力，应用架构应该具备一定的前瞻性，同时保证架构的灵活性和可扩展性。应用架构在覆盖现有业务的基础上，能够满足未来业务发展的可扩展性，并且考虑现有的资源配置，保证架构的可落地性。

- 应用企业化

通过应用架构的设计，解决系统多、功能分散或者界限不清晰的问题，推动企业进行集中的应用建设。并且全面考虑到业务的需求，增强对外服务相关的组件设计，提升系统对外服务的能力。

- 系统平台化

将相同的业务逻辑抽象出来，形成公共的服务组件，采用平台化的策略，形成基础平台，并且针对业务功能的差异，进行个性化的配置和开发，从而实现系统的灵活性和扩展性，支持快速产品的研发。

- 系统整合化

将相同的业务组件抽象出来，统一建设，在此基础上，考虑系统差异化的需求。例如数据报送规则的差异、产品加工逻辑的差异和服务对象的差异等。实现机构、用户、权限等公共组件和技术组件的整合。

- 适度松耦合

减少组件间的相互依赖，提高系统的故障防范和隔离的能力，同时参考最佳实践，结合业务的特点，合理划分应用架构的各个层次，提高组件的内聚性。

4. 创建应用架构的整体步骤

如图 1-7 所示。以业务战略为出发点，形成企业的业务能力和组件化业务模型，参考业务需求，梳理未来应用功能模型，在应用架构设计原则的指导下，形成未来的应用架构，最后进行未来应用场景的验证。

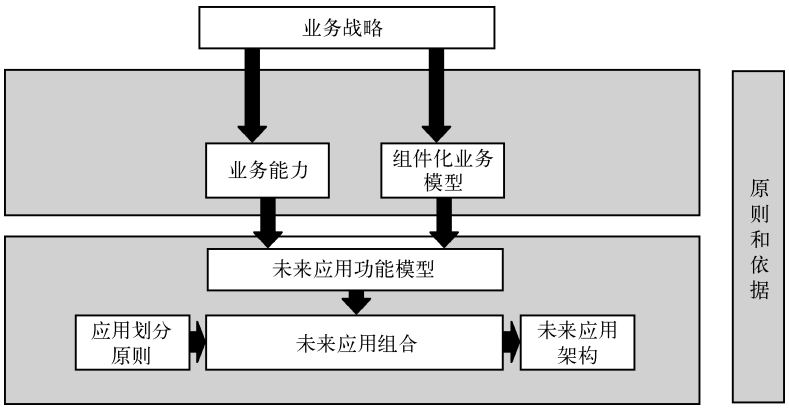


图 1-7　创建应用架构的整体步骤

5. 应用架构相关案例

某金融机构应用架构示例，如下图所示：

（1）客户服务层

客户可以包括对公客户和个人客户，包括综合前端服务平台、网上银行系统、电话银行系统、自助服务、手机银行系统等内容。

（2）渠道整合层

渠道整合层主要包括综合前置平台，ECIF 等。

（3）业务处理层

业务处理层主要包括：总账系统、核心业务系统、信贷管理等。其中总账系统主要是对整个银行财务状况的管理。核心业务系统包括总账接口、瘦核心和应用接口。总账系统通过总账接口与核心业务系统相连。瘦核心主要是银行的会计核算功能，账户管理和客户信息管理等。

应用架构不是本书重点，所以不做赘述。

六、数据架构

1. 什么是数据架构

从概念上来说，数据架构是指与数据相关的各种架构组件的排列顺序，其中架构组件主要实现数据的存储、交互、分布、流转和应用等功能。

数据架构的核心主要包括数据层次的划分、数据的分布、各层次的数据模型和数据的转换等。数据架构是企业架构中最重要的组成部分之一，也是本书的重点内容之一。

数据架构主要研究和解决如何管理和使用数据。主要内容包括数据从源系统经过各种处理、加工而达到目标系统的布局与流向的框架结构。

数据架构的目标是为了实现企业数据的标准化、一致性和准确性，在此基础上，充分挖掘数据的价值，有效支持企业的数据管理和经营决策分析，实现企业数据的统一规划体系。

数据架构可以帮助企业消除信息孤岛，建立一个共享、通用的企业级基础数据平台。

2. 数据架构包含的内容

数据架构主要包含数据定义、数据分类、数据分布、数据 CRUD 等内容。

• 数据定义

所谓数据定义就是数据模型。数据模型是指用实体、属性及其关系对企业生产运行过程中涉及的所有业务概念和逻辑规则进行统一的定义、命名，包括数据概念模型、数据逻辑模型、数据物理模型。

数据模型是数据架构规划中最重要的内容之一，良好的数据模型可以反映业务模式的本质，确保数据架构为业务需求提供全面、一致、完整的高质量数据，从架构规划以及设计层面，明确数据概念模型，制定数据逻辑模型和物理模型。数据模型是业务人员、IT 人员进行沟通的一套语言。

• 数据分类

数据分类是根据业务特征对数据进行归类和划分，用层级列表的方式展现业务的规则，数据分类的规范需要满足各种数据需求对数据组织的要求，它独立于具体的数据模型和数据分布。

• 数据分布

数据分布包括数据的业务分布与数据的系统分布。数据分布主要分析数据在业务各个环节中的创建、引用、更新和删除，并且根据业务对数据的处理特点，规划合理的数据分布，考虑相关的数据流向，以满足相关的业务需求。

在规划设计数据分布的时候，我们需要考虑以下几个方面。

① 明确系统不同位置之间的数据定位，以及数据的内容和数据流向。

② 考虑海量数据在不同数据库之间的快速增量迁移。

③ 考虑数据的快速加工。

④ 能够适应数据采集的多元化。

⑤ 需要考虑特殊情况下的数据纠错更

• 数据 CRUD

CRUD 是建立（Create）、读取（Read）、更新（Update）及删除（Delete）这 4 项操作的英文缩写。数据的 CRUD 可以明确系统核心的数据由哪些系统产生，哪些系统有权限读取这些核心数据，而这些数据的更新和删除的权限属于哪些系统，数据 CRUD 是为了确保数据的安全性和一致性。

• 数据管控

数据管控包含数据质量管理、数据生命周期管理、数据标准管理、元数据管理等多个管控专项，如图 1-8 所示。

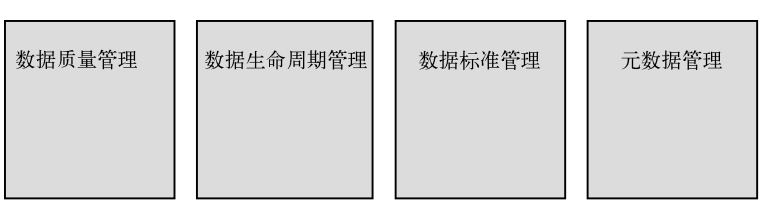


图 1-8　数据管控包含的内容

• 数据质量管理

数据质量管理是指通过一系列技术手段或者管理手段提高数据质量的过程。数据质量管

理是循环管理的过程，目的是通过提升数据的使用价值，为系统赢得经济效益。

● 数据生命周期管理

数据生命周期管理是按照数据的业务属性划分数据的几个阶段：数据的创建、数据的使用、数据的归档和数据的销毁。

数据生命周期管理的目的是为了满足对历史数据查询的要求，减少数据冗余，提高数据的一致性，并且提升系统的性能和响应速度。减少数据存储、运维等方面的基础设施投入。

● 数据标准管理

数据标准是统一对数据的理解和使用，为数据的业务属性、业务规则、管理属性和技术属性制定统一的规范。

通过数据标准管理，可以加强对业务的标准化工作，强化对业务的管理，完成对重点数据的统一管理。数据标准管理的原则：保证数据标准命名、编码的唯一性，维护数据标准的权威性和稳定性，保证数据标准的准确性和可执行性。

● 元数据管理

元数据管理是指管理数据的数据，负责记录和管理系统中所有数据的定义、规则、规范和流程。元数据管理可以清晰、直观地了解数据的来源、变化过程等信息。当数据发生变化时，用户可以借助元数据管理工具分析出这些数据变化带来的影响。

3. 数据架构的目标

实现企业数据的标准化、一致性、准确性和可靠性。制定实现企业数据统一管理的规划体系。有效支撑企业信息数据管理和经营决策分析。

4. 创建数据架构的整体步骤

数据架构对于企业有效地分配、部署和使用数据，实现数据的组织、共享，从而保证数据在各系统之间的一致性、有效性和完整性都有重要的指导意义。

创建数据架构的整体步骤（见图1-9）包括在了解数据架构现状的基础上，参考系统需求，借鉴行业先进的数据分类方法和参考架构，分别从三个体系进行数据架构的规划。

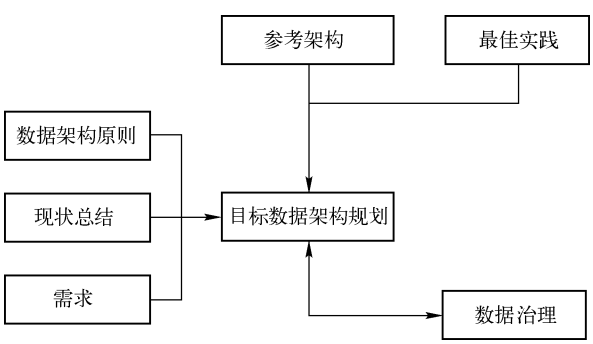


图 1-9　创建数据架构的整体步骤

5. 数据架构规划工作思路及方法

首先，数据架构从业务特征和业务需求出发，明确数据主题域的划分和数据的分类，主题域是从较高层级对业务进行抽象和归纳，是从概念层面上对系统的全面描述，需要考虑业务的扩展性。主题域划定后，一般较少变更。

其次，进行数据模型的设计。对于目标数据架构来说，一般流程是参考行业内先进的架构经验进行目标架构的设计，包括对数据存储、分布和流转的设计。

最后，对数据分布和流转进行场景验证，同时需要考虑各个阶段的数据管控的要求。

具体内容如图 1–10 所示。

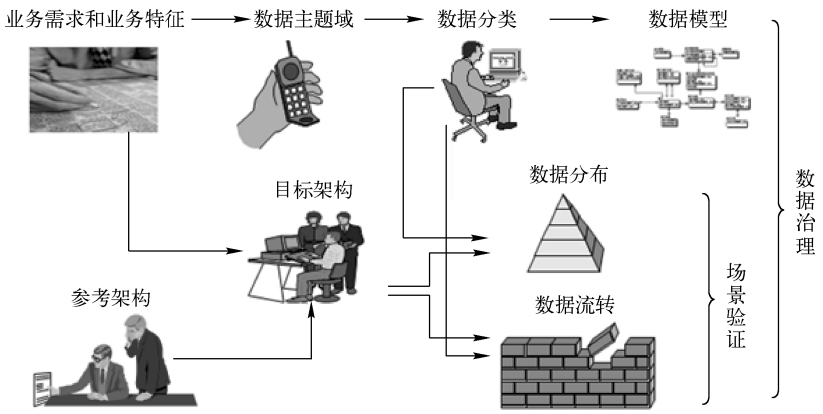


图 1–10　数据架构规划工作思路及方法

6. 数据架构原则

数据架构的原则主要包括灵活性原则、高效性原则、可扩展性原则、数据共享原则、数据可用性原则、数据定义标准原则、数据安全性原则。

● 灵活性原则

对于数据的组织及其架构的划分要充分考虑灵活性。例如，源数据采集格式需要考虑不同业务的需求，能够灵活地适应业务的变更。

● 高效性原则

需要考虑数据校验和数据加载的高效性。例如，各个数据库之间的数据迁移、产品加工和产品的快速生成都需要考虑高效性。

● 可扩展性原则

数据架构整体规划要充分考虑系统未来的可扩展性，在新技术或者新需求、新业务出现时，能够尽量减少数据架构的变更。

● 数据共享原则

数据在系统内可以共享，相同的数据指标需要遵循唯一性，强化对公共需求的加工。

● 数据可用性原则

对数据的采集以能够支撑业务需求为基础。

● 数据定义标准原则

数据项必须有易理解的业务定义，使用户理解数据的意义，同时确保数据的定义遵循统一标准，而且数据标准需要满足完整性、正确性、一致性等要求。

● 数据安全性原则

数据按照非功能性要求，定义数据的安全级别、安全管理等级。并且区分敏感数据和非敏感数据。

7. 数据架构相关案例

数据架构从数据的产生、加工、使用和管理的视角来描述业务系统。数据架构的规划主要包括以下几个方面：

1）数据分类和数据模型化，从数据业务特性出发，规划数据主题域，并且在数据主题域的基础上对数据进一步分类。然后根据数据分类，对关键属性和核心数据关系模型化，形成高阶的数据模型。

2）根据行业先进的数据架构，结合业务数据的加工特征，重点考虑数据架构的灵活性、可扩展性和高效性等几个方面，规划目标数据架构。

3）根据数据分类，规划数据分类在目标架构逻辑数据库存储上的分布与流转，从而对目标数据架构进行验证。

4）结合业务管理要求，规划系统的数据治理架构。

七、技术架构

1. 技术架构概念

技术架构是 IT 架构中比较底层的架构，它定义了如何建立一个 IT 运行环境来支持数据架构和应用架构。

技术架构主要描述业务、数据、应用服务部署的基础设施能力，通过技术架构可以建立一个 IT 平台，涉及对技术的采用、基础设施的建立、产品的选择、系统的管理等方面。

技术架构需要考虑业务架构、数据架构和应用架构，包括一些软硬件、网络技术等方面。技术架构的设计目标就是参考成熟的技术规范，打造一个安全、可靠、灵活、易维护，并且支持业务连续性的 IT 技术架构。

2. 技术架构的目标

1）针对未来系统的技术架构，制定技术架构设计规范、实施规划、决策支持等内容。

2）通过技术架构，提高系统的灵活性、扩展性。

3）通过标准化、组件化和平台化技术打造灵活、可扩展的平台，这样可以快速地满足业务的变化。

3. 技术组件的识别

可以根据技术架构的相关案例和业务组件需求，分层次去识别系统的技术组件。技术组件的描述见表 1–1。

表 1–1　技术组件的描述

技术组件名称	技术组件功能描述
调度服务	提供统一的任务调度服务接口，实现基于平台的作业调度管理功能
元数据管理	元数据是用于描述数据及其环境的数据。一般来说，它有两方面的用途，即业务元数据和技术元数据
加解密	提供标准的加解密技术及接口，能够满足数据安全传输、存储的要求
缓存管理	基于成熟的缓存框架，同时提供数据缓存管理，提高数据的使用和存储效率

4. 技术架构原则

技术架构的原则如图 1–11 所示，主要包括以下几个方面的内容：

（1）安全、可靠性原则

从应用组件到物理基础架构，需要充分考虑系统的可用性，以保证系统运行的连续性和

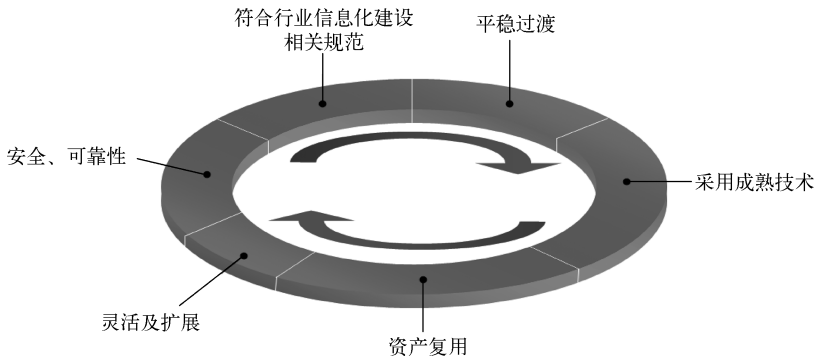


图 1–11　技术架构原则

完整性。安全性应该遵循相关安全政策、标准和法规。

（2）灵活性及扩展性原则

满足灵活加工产品的要求，业务变更或新功能开发能够在短时间内完成，能够适应业务量的变化。

（3）资产复用原则

对已有的成熟技术、规划经验等相关资产进行提炼和重用，降低开发与维护的成本。抽取公共技术组件，使架构能够满足不同业务之间差异化的需求，支持业务的可持续发展。

（4）采用成熟技术原则

选用主流技术，采用成熟的技术平台和开发工具，引入已验证过的开发框架，提升开发效率，平衡成熟产品技术和自主开发能力。基于成熟产品及实施案例，选择合适的技术路线。

（5）平稳过渡原则

能够支撑业务的连续性，保证未来系统的过渡和切换必须是阶段化可控的和低风险的。

（6）符合行业信息化建设相关规范

遵循统一认证规范、容灾规范、安全规范、广域网安全规范等，加强系统设计、开发等规范管理，在已有规范的基础上形成并完善整体架构方案。

5. 技术架构规划工作思路及方法

技术架构规划的工作思路和方法如图 1–12 所示。

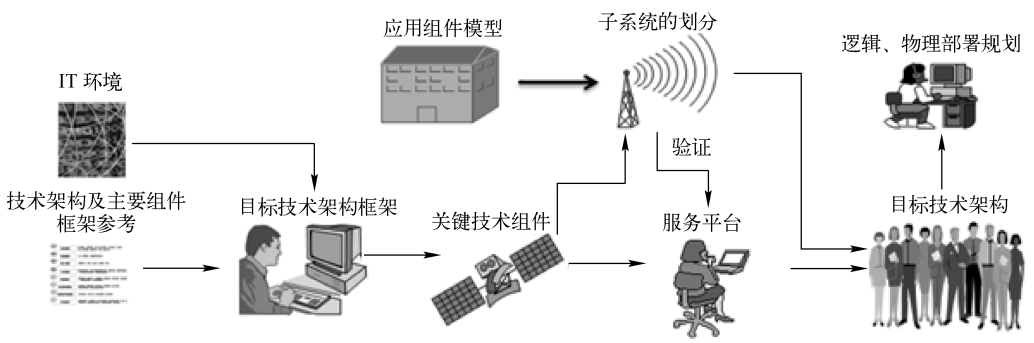


图 1–12　技术架构规划工作思路及方法

1）参考先进的技术架构，结合现有的 IT 环境，采用分层的方式设计目标技术架构。目标技术架构提供高度的灵活性和可扩展性。

2）参考技术架构和已定义的业务组件需求，分层次识别未来系统关键技术组件。

3）根据应用架构组件分组，按照业务特点和技术实现考量，划分子系统。

4）根据技术组件的服务能力，按照 SOA 的思路划分为几个服务平台，为规划子系统提供基础的公共服务。

5）子系统的划分也为了验证服务平台中的服务能力是否有缺失。间接验证技术组件是否有缺失。

6）提供标准化服务的技术组件与子系统的结合，形成完整的目标技术架构。

7）最后参考最佳实践，对逻辑部署和物理部署进行规划。

技术架构主要包括安全管理、集成服务、接入渠道、公共服务等方面的内容，如图 1–13所示。

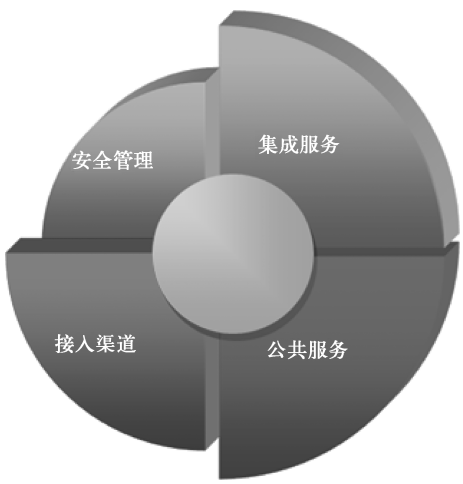


图 1–13　技术架构

（1）安全管理

安全管理主要包括目录服务、身份管理、用户认证 & 授权、单点登录、访问控制、审计服务、数据安全、PKI、操作安全。

（2）集成服务

集成服务主要包括内、外部接口，数据整合和拆分。

（3）接入渠道

接入渠道主要包括客服中心、信件、E – mail、互联网、手机。

（4）公共服务

公共服务主要包括信息服务总线、文件交换服务、流程引擎、规则引擎、批量作业服务、审计服务负载均衡、存储管理及恢复。

其中主要的技术组件包括网络服务、系统管理服务、测试和开发服务、平台服务等内容。

（1）网络服务

网络服务主要包括网络管理、网络安全、传输服务、网络协议、网关服务、路由服务、

网络加速服务、内容网络服务。

（2）系统管理服务

系统管理服务主要包括配置管理、网络管理、软件分发、问题管理、账户管理、高可用性管理、监控及其优化管理。

（3）测试和开发服务

测试和开发服务主要包括开发环境、开发工具、优先级管理、测试环境。

（4）平台服务

平台服务主要包括数据库服务、打印服务、其他设备服务、机房基本设备、UPS、布线等技术、服务器平台架构、高可用性架构、灾难备份机制。

技术架构从应用架构和数据架构实现的角度进行规划。技术架构规划过程主要包括以下几个步骤。

1）参考技术架构，结合现状分析和技术架构原则，识别各种技术组件。

这些组件可能是应用架构或数据架构中某些组件的功能实现，也可能是作为一个系统必须具备的技术组件。针对这些技术组件，考虑各种成熟软件实现技术。

2）根据子系统划分原则，将系统划分为多个子系统和技术平台。

这些技术平台由技术组件构成。通过技术平台构建多个子系统。针对每个子系统，定义包含的应用组件和逻辑数据存储，并描述每个子系统与技术组件之间的关系，保证技术平台所包含的技术组件能够很好地支持所有应用组件的技术实现。

3）系统的实现一般分为展现交互层、逻辑执行层和数据存储层。

在系统部署上由不同的软件技术支持。按照方法论，通常将部署单元分为三大类：展现部署单元、执行部署单元和数据部署单元，根据多个子系统和技术平台所包含的组件，识别对应的部署单元。并根据组件之间的关系定义部署单元之间的关系。

4）根据位置、用户分布、网络连接及接入点等情况，结合参考架构和用户的 IT 环境，规划系统的逻辑架构。

5）参考逻辑架构，结合真实的 IT 环境，包括开发、测试、生产环境，可以采用诸如虚拟化或者资源池技术，规划物理架构和基础架构。

6. 技术架构相关案例

某金融机构技术架构相关案例如图 1–14 所示。

其中技术架构包括：渠道层、应用服务层、公共技术服务层、集成服务层、软件服务层和基础设施层。例如，软件服务层中的内容管理是为未来系统提供更广泛的非结构化内容进行存储、访问和管理，包括业务中涉及的影像，各种格式的办公文档，XML、HTML 文件，各类报表、图像和音频/视频信息等。

八、企业总体规划总结

企业总体规划包括企业战略和企业架构两个部分。企业战略描述的是企业的目标。企业架构描述的是业务流程、运营模式、关键业务指标和企业 IT 系统需要完成哪些工作等内容。

企业战略决定企业架构的模型，同时企业架构又支持企业战略的实现。如果我们把企业战略看成是一个城市的发展方向和战略目标，包括对城市的定位等方面。那么企业架构就是对城市的设计规划，包括城市的组成，每部分是如何构建的，以及它们之间的关系是什么。

从本质上来讲，企业架构是连接企业战略和 IT 项目实施的桥梁。通过企业架构的规划，

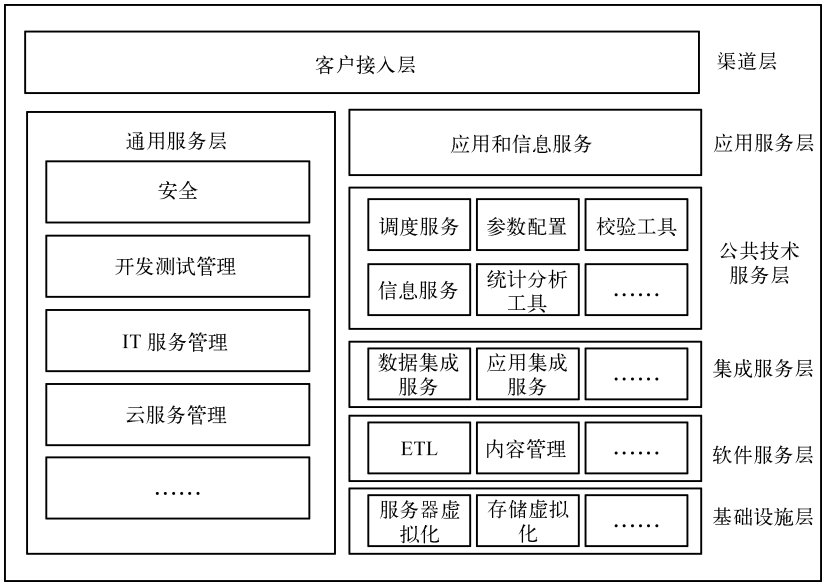


图 1-14　某金融机构技术架构案例

包括在业务战略和 IT 战略理解的基础上，进行自顶向下的设计，形成稳定的 IT 环境，从而将战略、业务流程和具体的 IT 规划三者之间紧密的联系起来。

我们参考企业架构的方法论，IT 架构是业务发展和业务战略的实现提供架构支撑和保障。业务架构可以作为整个 IT 架构的核心输入。业务架构和 IT 架构是相互依赖，相互促进的关系。在 IT 系统的设计和开发过程中，业务架构可以提供完整的业务视图和业务要求，指导 IT 架构的实现，同时 IT 架构保障业务架构的实现。业务架构主要包含了业务流程、业务对象、地域和价值管理的描述，而 IT 架构主要包含了应用架构、数据架构和技术架构等内容。

业务架构对应用架构和数据架构提出业务需求。而应用架构为业务架构提供应用支持，数据架构为业务架构提供数据支持。同时技术架构是数据架构、应用架构到 IT 系统的落地和实现。应用架构和数据架构是业务架构落地到系统架构的一个重要阶段。在企业架构中，数据架构是核心，也是本书的重点内容之一，因为数据是信息系统的重要资源，在构建 IT 架构的时候，首先考虑数据架构对业务的支持，理想的 IT 架构是数据驱动的。数据架构帮助企业消除信息孤岛，建立一个共享、一致的企业数据基础平台。

应用架构是为业务提供哪些应用和功能，它主要连接业务架构中的流程、业务组件、功能和人员等，同时也能连接数据架构中的数据管理部分，还能够提出对技术架构和基础设施的要求。应用架构有着承上启下的作用，可以避免各个部门从自己的角度出发，建立很多重复的，难以共享的应用系统，应用架构在 IT 架构中也发挥着重要的作用。

技术架构是 IT 架构中比较底层的架构，它用来支持数据和应用，以保证业务的正常运转。技术架构需要考虑技术的采用，未来技术的发展等因素。

1.2　国内商业银行战略规划和架构状况剖析

在过去 10 年间，我国银行业在信息化建设过程中，已经基本形成了完整的框架体系，

建成了面向业务的新一代综合业务系统。特别是“十一五”期间，以四大行（工、农、建、中）为代表的国有银行，相继提出了建设国际一流银行的战略目标，其他股份制银行也提出了相应的发展战略，逐步完善了符合本行特色的战略和架构。

总体来说，无论是国有银行还是股份制商业银行，都已经认识到了战略规划和架构的重要性，同时整个银行业基本上实现了核心业务的“数据大集中”，提升了银行的抗灾难能力，并且随着银行业务的增加，对产品的创新和业务流程的改进提出了更高的要求，通过一系列的建设和升级工作，商业银行的业务功能逐步完善，效益不断增加。

但是随着银行信息化建设的深入发展，面临的问题也逐渐暴露出来，在应用系统建设方面，各个银行都追求业务系统的快速开发和产品的快速上市，业务部门和技术部门之间存在着孤立的现象，造成了部分银行的系统数目繁多，系统之间缺少企业级整体架构的思想，如图 1–15 所示。

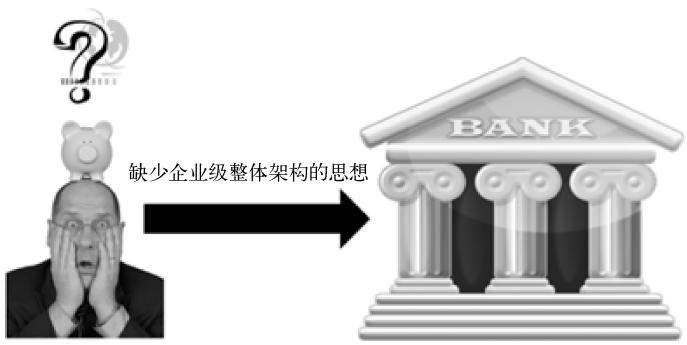


图 1–15　银行信息化面临的问题

在数据架构方面，数据被分散到各个应用系统之间，数据质量较差，数据的使用缺少规范。同时部分银行对战略规划的概念和认识还缺少统一性，很多银行都处于比较片面的阶段，从而影响战略规划对日常系统建设的指导作用，因为战略规划和业务规划的契合度不高，缺少 IT 战略对于业务战略的支持和业务战略对 IT 战略的指导。

因此，在战略规划的过程中，需要业务部门和 IT 部门共同参与，相互合作，达成业务和 IT 技术部门的共识。业务部门和技术部门总是存在着看不见的鸿沟，业务部门经常抱怨技术无法适应市场的需求，而技术部门则经常抱怨业务需求的不确定，需求变更过于频繁。作为技术部门常常被动地接收业务需求，疲于应付，更谈不上技术的创新和引领业务的发展。

以上最主要的原因之一就是缺乏从战略角度出发的总体架构规划，当业务部门提出不同的业务需求，IT 技术部门则以不同的技术框架和软硬件去满足业务，各个系统相互分散，在银行内部形成了一个个的“信息孤岛”，使银行的维护成本大幅提高，不能有效地利用数据资源，从而无法利用这些宝贵的资源去推动业务向前发展。从业务上来说，由于缺乏对全局的把握，无法形成统一的业务视图，降低了业务的灵活性，也就无法支撑日益复杂的业务。

基于以上的现状分析，我们从管理的角度来说，应该从制度上消除技术部门和业务部门之间的“隔膜”，从管理机制上把 IT 技术部门和业务部门的目标统一起来，使业务部门除

了关注业务和经营指标外，也关心具体的操作流程、应用架构、数据架构和技术风险等内容。技术部门除了考虑技术实现外，也应该考虑项目的效益，使技术融入业务，建立相应的考核机制和激励措施，如图 1–16 所示。

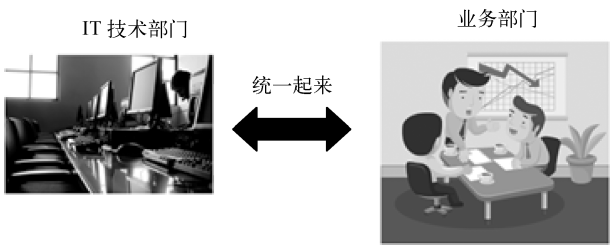


图 1–16　消除技术部门和业务部门之间的“隔膜”

在 IT 技术部门中，首先在需求分析阶段，统筹和优化整体的业务需求。然后根据业务需求，规划项目的设计、开发和运维等活动。技术部门应该主动了解需求，不仅要承担技术的角色，也要考虑业务解决方案和对业务流程的整合。技术部门的真正价值就是利用已有的 IT 技术提供整体的业务解决方案，帮助银行进行业务流程优化和改造。

在业务部门中，不仅需要考虑业务流程的优化、业务的集中处理，更应该将 IT 战略和银行的业务战略融合到一起，从战略、管理变革的角度降低 IT 的风险。IT 与业务的融合，可以促使商业银行适应市场环境的变化，同时也相应地促进了业务的发展，提高了商业银行的竞争力。

因此，金融机构迫切需要企业架构的方法论来解决由于信息化建设带来的各种问题。在银行的信息化建设过程中，企业架构越来越受到大中小银行的重视，它们已经开始从整体架构上规划 IT 系统。

从战略规划的角度来讲，需要遵循以下几个原则：

（1）业务和 IT 的高度融合

对于各个银行来说，IT 战略规划要坚持从自身的业务战略出发，结合行业的发展趋势，全面考虑信息化建设的各项 IT 和业务工作，实现业务和 IT 的高度融合。

（2）借鉴先进经验

根据银行的信息化战略，积极吸取国内外先进的理念、整体框架和先进技术，充分利用已有的资源，提升银行的创新能力，从而推动业务的发展。

（3）分阶段重点实施

信息化的建设不是一蹴而就的，而是逐步完善的过程，根据业务发展的重点方向，利用现有的资源，分阶段重点实施系统规划。

从整体架构的角度来说，企业架构是桥梁，在对业务战略和流程理解的基础上，进行规划，形成灵活的、可扩展的架构。

对于银行来说，其架构设计是否灵活、先进，已经关系到银行核心业务和未来业务的发展，包括是否能够适应市场竞争带来的压力。

明确战略规划，保证战略规划的前瞻性、全面性和统一性，识别未来发展的定位和战略目标，结合银行整体的业务架构，设计应用架构、数据架构和技术架构，并且建立相应的业务流程和决策机制，更好地推动银行战略目标的实现，这个过程已经成为国内外银行当前的

重要任务之一，这也是银行通过信息化转变成“智慧银行”的主要过程。

1.3 数据架构在银行信息化建设中的重要性

数据架构在企业信息化建设中占有非常重要的地位。目前来说，资金、人才和数据是公认的企业的资产。企业可以通过使用数据，提供更好的产品和服务，降低成本和控制风险。

如何建立一个灵活、松耦合、高性能的数据架构规划体系，是很多企业必须重视的问题，经过多年的信息化实践，很多企业已经逐渐认识到，系统应该具备多渠道数据采集能力、历史与趋势分析能力。数据架构规划在信息化过程中起着非常重要的作用，通过数据架构规划可以推动企业信息化的进程，使企业充分利用数据，提供更好的产品和服务，降低成本和控制风险，促进企业经营战略的实现，提升企业的核心竞争力。

一、数据架构在企业总体规划中占有非常重要的地位

（1）数据是信息系统中最重要的资源之一

信息系统就像是数据工厂的流水线，而核心是数据的加工和流转。比较有价值的其实就是数据。

（2）数据是业务和技术沟通的桥梁

当业务需求和技术实现出现脱节的时候，信息系统往往就会出现问题。为了加强彼此的沟通，就会找到一门共同的语言，这种语言就是数据，它承担着业务和技术沟通的桥梁作用。

数据是业务系统真实的记录，可以通过分析数据的过程完成对业务需求的技术性分解，同时数据又是系统功能设计的依据。

（3）数据是企业价值提升的“推进器”

通过对数据的全面分析，可以促进企业的业务发展。

二、良好的数据架构对银行信息化建设的重要性

（1）数据是银行的核心资产

在信息化建设过程中，数据又是信息系统的重要资源，如何提高数据的利用率是数据架构关注的重点之一。另外，在数据架构过程中应该有大局观念和全局意识。优秀的数据架构可以提高银行的服务能力和满足银行多样化的需求。

（2）支持产品的多样化

目前很多银行系统都是从数据源的采集环节到数据终点的发布，整体呈现出了一种紧耦合的关系，经常出现对系统某一功能点的调整修改，都需要对整个系统的多个处理环节进行改造的情况。这种落后的数据架构，已经严重制约了硬件性能的发挥，最后只能靠打补丁的方式对现有系统进行改造。也就是在现有系统上增加新的功能点，或者开发新的产品，采集新的数据源，每个系统都自成体系，这种方式会造成严重的重复建设的问题，资源也会严重浪费，同时也无法支持产品的多样化。

（3）消除信息孤岛

数据架构可以帮助银行消除信息孤岛，建立共享、通用的企业数据基础平台。没有好的数据架构，同样也不会有好的数据质量，这样会降低银行的社会公信力和权威性，也就降低了社会的认同感。

小结

- 数据和空气有着类似的功能，不同的企业和个人需要不同类型的数据，数据就是价值。大数据即将开启一个新的时代，无论知识普及、技术共享，还是人才培养，都需要国家从战略层面上去支持。
- 很多企业已经充分认识到数据是核心资产和竞争力，正是这个原因，IT 人员才需要了解数据架构方面的知识，并且能够利用数据架构提升数据分布的合理性，降低数据存储的成本。
- 从概念上来说，数据架构是指与数据相关的架构组件的排放顺序，架构组件负责数据的存储、交互、应用等功能。同时数据架构是企业架构的重要组成部分，对于企业有效地分配、部署和使用数据，实现数据的合理组织、有效共享，具有重要的指导意义。
- 对于企业架构来说，它可以从全局出发，统一各类概念和术语，梳理现有的系统，提取可重用的 IT 资产，从而降低开发的成本，提高数据质量。
- 企业架构包含业务架构和 IT 架构，我们可以参考先进的架构实践，对 IT 架构进行优化，确保 IT 架构能够很好地支持未来业务的发展。而 IT 架构又包含了应用架构、数据架构和技术架构。
- 企业的总体架构就是从全局出发，解决现存问题，同时满足现实需求和适应未来发展的需要，有效地对资源进行管控，加强 IT 技术实力，并且指明了企业的经营方向和发展目标，对企业远景发展轨迹进行全面的规划。
- 企业总体规划包括企业的战略、企业架构和企业具体的实施解决方案。
- 企业战略是对企业发展目标，包括达成目标的方法和途径的总体谋划。企业战略的实质就是企业的发展方向和定位。
- 企业战略的作用和目标就是企业能够运筹帷幄，根据自身的资源和环境选择合适的经营发展方向，它是一个长远、持续的发展过程，具有一定的稳定性。企业战略属于企业的宏观管理范畴，具有指导性、长远性、系统性、风险性、全局性和竞争性等主要特征。
- 企业的业务战略是指企业拥有的所有资产，通过多种方式进行有效的运营，以实现利润的最大化和资本的增值。
- 企业的 IT 战略是指在充分研究企业的发展愿景、业务策略和管理的基础上，形成信息系统的远景、组成架构、逻辑关系等，以支撑企业战略目标的实现。
- 企业架构实质上就是对企业多角度的一种描述，它反映了企业的业务流程、技术的组织和安排，是对企业关键性业务和技术的整体性描述。企业架构的目的是将跨企业的、零散的业务流程优化成一个集成的环境，同时帮助企业执行业务战略及 IT 战略规划。企业架构的过程实质上就是对现实世界中企业的业务流程和 IT 设施抽象的过程。它反映了企业的业务流程和 IT 架构之间的关系。一般来说，企业架构包括业务架构和 IT 架构。
- 广义的业务架构包括产品、销售、财务、人力资源、客户服务等企业核心的业务功能

和职责。并且将企业战略转化成企业运营的目标和形式，同时明确相关人员、企业资源、IT 资源和服务是如何协调和部署的。我们可以说由企业战略决定了业务架构的模式，同时业务架构又是企业战略实现的手段。而狭义的业务架构包含了企业运营活动中的业务策略、组织、关键业务流程、组织架构以及人员结构等内容。

- IT 架构是对企业系统的 IT 规划，是建立企业信息化系统的综合性的蓝图，IT 架构可以帮助企业获得最好的投资回报，同时实现业务和技术接口之间的标准化，保证企业运营和企业战略之间的一致性。
- 应用架构是对实现业务能力、支撑业务发展的应用功能结构化的描述方法。系统的应用架构可以从功能和应用两个不同的视角描述系统各组件构成以及组件之间的关系。功能组件模型侧重于业务功能，而应用组件模型侧重于应用系统设计。
- 数据架构是数据在信息系统中的布局与流向的框架和与数据相关的架构组件的摆放。数据是指系统所处理的所有信息和数据。而架构组件负责数据的存储、交互和应用等功能。
- 技术架构是 IT 架构中比较底层的架构，它定义了如何建立一个 IT 运行环境来支持数据架构和应用架构，技术架构主要描述业务、数据、应用服务部署的基础设施能力，通过技术架构可以建立一个 IT 平台，涉及对技术的采用、基础设施的建立、产品的选择、系统的管理等方面。
- 我们从管理的角度来说，应该从制度上消除技术部门和业务部门之间的“隔阂”，从管理机制上应该把 IT 技术部门和业务部门的目标统一起来，使业务部门除了关注业务和经营指标外，还要关心具体的操作流程、应用架构和技术风险等内容。技术部门除了考虑技术实现外，还要考虑项目的效益，使技术融入业务，建立相应的考核机制和激励措施。
- 明确战略规划，保证战略规划的前瞻性、全面性和统一性，识别未来发展的定位和战略目标，结合银行整体的业务架构，设计应用架构、数据架构和技术架构，并且建立相应的业务流程和决策机制，更好地推动银行战略目标的实现，这个过程已经成为国内外银行当前的重要任务之一，这也是银行通过信息化转变成“智慧银行”的主要过程。
- 国内商业银行 IT 架构的变革主要表现在以下几个方面：商业银行的 IT 架构必须建立“以客户为中心”的原则，以市场为导向的业务流程。基于“以客户为中心”的思想，建立一系列产品创新的快速响应机制。商业银行的 IT 架构应该满足低成本、灵活性和抗风险性等三个基本要求。
- 数据架构在商业银行的信息化建设中占有非常重要的地位。目前来说，资金、人才和数据是公认的企业的资产。企业可以通过使用数据，提供更好的产品和服务，降低成本和控制风险。

第 2 章　数据架构现状分析

本章目标

通过第 1 章的学习，我们已经了解了什么是数据架构、企业总体架构规划包含哪些内容、什么是企业战略和业务架构，以及应用架构、数据架构和技术架构的定义是什么。还了解了数据架构规划、应用架构规划和技术架构规划的方法论、企业总体架构和数据架构之间的关系等重要内容。

从本章开始将正式学习数据架构方面的知识。

本章重点介绍对数据架构现状分析的工作方法。现状分析主要发生在项目的初始阶段，主要分析现状数据架构存在哪些问题，如何对现状数据进行分类，结合对战略的理解，明确下一阶段的工作重点。掌握数据架构现状分析的相关案例，如何进行数据分布、流转的现状分析，关于数据治理现状分析的工作方法，数据质量管理的现状分析方法，数据生命周期管理的现状分析方法，数据标准管理的现状分析方法，元数据管理的现状分析方法等内容。它是项目成功的关键环节之一。

学习本章后，读者将掌握：

- 对数据架构现状分析的工作方法
- 对于现状调研和高层访谈来说，我们可以集中于哪些问题
- 对现状的数据分类的原则和方法
- 如何对现状数据进行分类
- 如何基于数据分类进行现状分析
- 如何对现有系统进行梳理
- 掌握数据架构现状分析的相关案例
- 学习数据处理架构的先进经验
- 如何进行数据分布的现状分析
- 如何进行数据流转的现状分析
- 关于数据治理现状分析的工作方法
- 关于数据质量管理的现状分析方法
- 关于数据生命周期管理的现状分析方法
- 关于数据标准管理的现状分析方法
- 关于元数据管理的现状分析方法

2.1　对数据架构现状分析的工作方法

数据架构现状的分析主要通过现状调研、资料的分析、高层领导访谈，了解数据架构的现状。

现状分析，主要以发现问题、分析问题为主，在理解现状的基础上，借鉴行业内先

进的经验，从数据分类、数据分布及其存储、数据处理架构和数据管控等几个方面对数据架构现状进行描述，从而发现数据架构存在哪些问题，同时提出改进的方向，如图 2-1 所示。

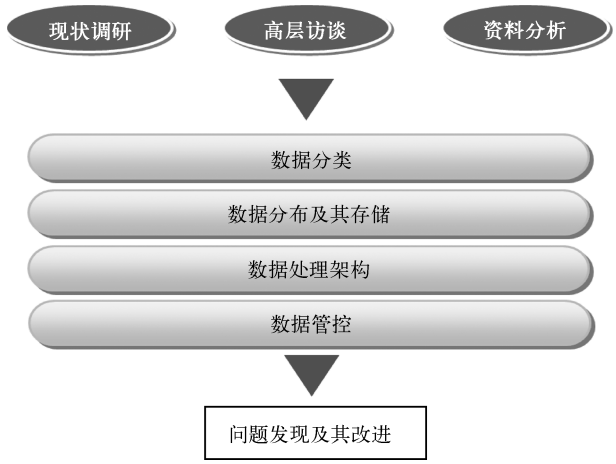


图 2-1　数据架构现状分析的工作方法

最后把发现的问题和数据架构改进的方向作为未来数据架构规划的依据和重要输入部分，如图 2-2 所示。

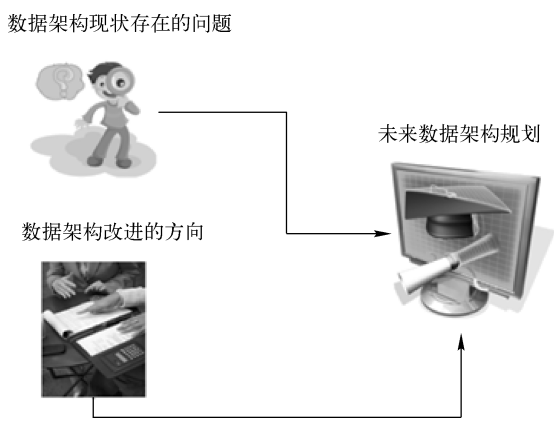


图 2-2　未来数据架构规划的依据和重要输入

对于现状调研和高层访谈来说，可以集中于以下几个方面的问题：

1）高层领导认为现有的核心业务系统有哪些？

2）现有系统能否支撑现有及未来业务发展？是否存在需要改进的地方？

3）在产品和服务方面有哪些思路？对 IT 方面有什么新的期望？

4）未来 IT 建设应达到什么样的水平？未来几年业务发展的目标是什么？

5）在新产品方面，未来的产品有哪些特点？

6）未来 3 ~ 5 年会拓展哪些数据？

7）在数据治理方面，哪些工作需要提高？是否能够提高数据质量？质量保证手段有

哪些？

8）技术发展很快，如大数据处理方式。对于新技术，领导层有什么看法？

9）目前系统和业务发展的优势和劣势是什么？

10）目前 IT 系统存在哪些问题？对 IT 架构的期望是什么？目前 IT 规划的目标有哪些？

11）数据采集、加工、对外服务上有哪些问题？

12）系统运维上存在哪些问题？

对数据架构现状分析的工作方法可以总结如下：

首先，可以先从数据分布、存储和流转等几个方面对系统现状进行描述，其中数据分布的现状分析是对现有系统的梳理，描述数据分类在各个数据库中的分布。

其次，对于数据架构的现状分析，可以参考行业内先进的实践经验，分别从数据的采集、加载、数据加工等几个方面对数据处理架构进行抽象和归纳。分析它存在哪些不足。

最后，从数据治理和管控的角度，对现状数据的数据质量、数据标准、元数据管理、数据的生命周期管理等几个方面存在哪些问题进行分析，发现现状存在哪些问题。

2.2 对现状的数据分类的原则和方法

2.2.1 对数据分类的说明

首先了解一下什么是数据分类。

数据分类是按照选定的属性（或特征）区分分类对象，将具有某种共同属性（或特征）的分类对象集合在一起的过程。

数据分类是在业务层面上将数据按照某种属性进行归类和划分，它是按照业务特征进行分类的，数据分类促进业务沟通，现状的分类有利于分析，规划的数据分类有利于设计。

数据分类最终可以形成数据大类和数据小类，数据大类是从全局角度理解业务，数据小类是从微观角度对同一大类的进一步细分。

数据分类的原则和方法主要包括以下几个部分：

（1）分类应该按照业务特征对数据进行划分。

（2）企业数据执行同一个分类标准。

（3）分类应该满足可维护性和可扩充性。

（4）分类没有二义性。

（5）分类应该满足业务需求对于数据组织的要求。

（6）分类是业务和技术沟通的桥梁。

2.2.2 现状数据的分类

一、数据分类——大类

数据大类是从宏观的角度理解企业全局的业务情况，我们可以在现状分析的基础上，对数据大类进行主题域的划分。主题域是从较高层次上对业务的一种抽象和归纳。在主题域的划分过程中，需要全面考虑业务的扩展性，当确定后，主题域很少发生变更。

通过对系统现状分析，并结合现有的业务，将数据分为几个较大的主题域。我们结合金融行业的业务活动特点，参考最佳行业实践和 Teradata 金融业逻辑数据模型，可以将数据大类分成 8 个部分：当事人、产品、渠道、合约、财务、机构、事件、活动。

（1）当事人

银行所服务的任意对象，如个人、客户和员工等。

（2）产品

银行提供给客户的产品和服务信息。

（3）渠道

渠道是客户和银行之间进行交互的方法和手段。通过渠道，客户与银行进行接触，购买相关产品和服务。

（4）合约

银行与客户之间、银行内部员工之间签订的协议信息。例如，银行和个人签订的贷款合同。

（5）财务

主要包括银行的总账科目余额、财务预算等信息。

（6）机构

是指银行内部的机构，如银行所属的分行机构、支行等。

（7）事件

基于合约的协议信息，有主体触发事件类信息，如存取款、收费、投诉等内容。

（8）活动

主要是银行对客户所做的各种宣传和促销活动，目的是将产品推销给客户，加强银行与客户之间的关系。

数据大类之间的关系如图 2-3 所示：当事人签订合约的信息，同时主动触发事件，事件的发生基于合约的内容，事件信息、机构的信息和合约的信息可以加工成产品等内容。

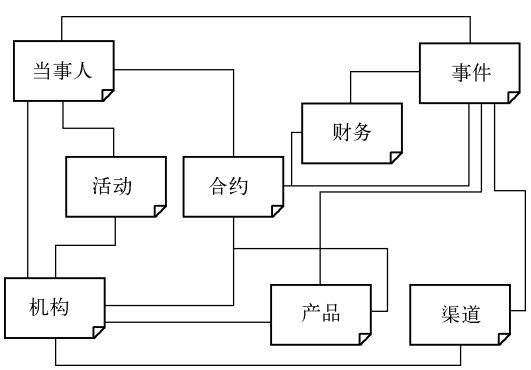


图 2-3　数据大类之间的关系

二、数据分类——小类

数据小类是在同一大类内，按照业务的特性进行进一步的细分。例如，我们按照数据 8 大类继续细分，举例见表 2-1。

表 2-1　数据分类

大　类	小　类	描　述
当事人	个人客户	包括个人客户的身份信息、职业信息、联络信息等内容
	企业客户	包括企业概况信息、身份信息等内容
	雇员	包括雇员身份信息、联系方式等内容
产品	服务类产品	包括查询报告、个人信息查询等内容
	统计类产品	包括管理统计报表等内容
	分析类产品	包括风险分析报告、一些分析挖掘类产品等内容
渠道	ATM	略
	柜面	略
	POS 终端	略
财务	银行的总账科目余额	略
	银行财务预算	略
合约	贷款合同信息	包括贷款合同的合同编号、合同授信额度、金额、币种、合同生效日期等信息
	担保合同信息	分为保证合同、抵押合同和质押合同等信息
机构	分行	描述分行的基本信息
	客服中心	描述客服中心的基本信息
	支行	描述支行的基本信息
事件	存款	略
	取款	略
	查询	略
	付款	略
活动	营销策略	略
	营销行为	略

2.3　数据架构现状分析

2.3.1　数据分布现状分析

通过对现有系统的梳理，数据小类在现有数据库的分布状况见表 2-2。

表 2-2　数据小类

数据小类	分布的数据库	数据小类	分布的数据库
个人客户	A 库，B 库，C 库	贷款合同信息	A 库，B 库
企业客户	A 库，B 库	担保合同信息	A 库，B 库
雇员	A 库，C 库	分行	A 库，B 库
服务类产品	A 库，B 库	客服中心	A 库，B 库
统计类产品	A 库，C 库	支行	A 库，B 库
分析类产品	A 库，B 库	存款	A 库，B 库
ATM	A 库，B 库，C 库	取款	A 库，B 库
柜面	A 库，B 库	查询	A 库，B 库
POS 终端	A 库，B 库	付款	A 库，B 库
银行的总账科目余额	A 库，B 库	营销策略	A 库，B 库
银行财务预算	A 库，C 库	营销行为	A 库，B 库

通过表 2-2 所示的分布可以看出，主要的分类数据有多个副本，数据的冗余度较高。

2.3.2 数据流转现状分析

通过对业务流程现状的分析，在处理流程环节中可能存在以下问题：

1. 数据处理各环节是否清晰

数据处理环节包括数据采集、产品加工和对外服务。我们需要从以下几个方面分析数据的处理环节。

1）是否在加载和数据迁移过程中进行了产品加工，加工方式是否统一。

2）加工生成的产品是否单一。

3）是否可以快速向用户提供丰富和个性化的产品。

2. 是否对数据流转进行了统一管理

数据处理的关键在于数据加载、清洗、整合、加工、迁移的各个环节。我们需要分析数据加载、整合和数据迁移的运行方式是否缺少统一的运行监控手段。

我们按照分类对数据流转现状进行描述，如图 2-4 所示。可以看出，数据出现反复抽取的过程，同一类的数据在多个数据库之间进行流动和复制，导致数据链条过长，严重影响系统执行的效率。

这种现象的原因是缺乏完整、良好的数据架构规划，导致“因事设库”现象的增多，缺乏数据的一致性。主要数据重复分布在不同的数据库中，造成冗余度较高，因为数据反复抽取，严重影响系统的效率。

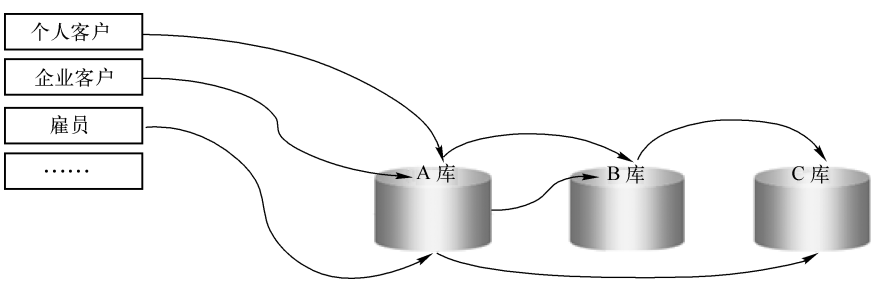


图 2-4　数据流转现状

2.3.3 数据处理架构现状总结

我们参考数据处理架构的先进经验，对现状进行抽象和归纳，如图 2-5 所示。数据处理架构可以分成数据源层、数据交换层、数据基础层、数据加工层和应用层等几个部分。

（1）数据源层

数据源层是通过各种方式从业务系统中抽取数据。

（2）数据交换层

数据交换层是对数据进行校验，最后再加载到目标库中。

（3）数据基础层

数据基础层是保存校验通过的数据，作为后续加工的唯一可信数据源。

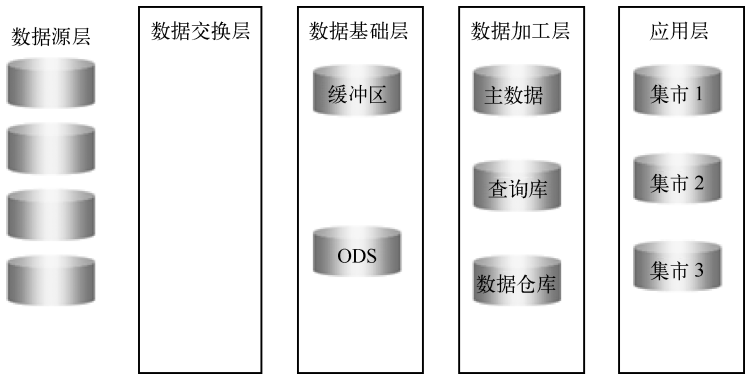


图 2-5　数据处理架构先进经验

（4）数据加工层

数据加工层是保存核心业务数据、当前的数据和历史数据，并且进行加工，以供应用层使用。

（5）应用层

主要进行产品加工，包括对基础产品的加工和增值产品的加工。

参考数据处理架构与系统现状的映射关系，从数据采集、数据加载、数据处理、数据加工和数据迁移等几个方面分析数据处理架构可能存在的问题。

一、数据采集现状分析

数据采集现状分析分为数据报送和上传的现状，包括采集的分类、数据的类型、文件的大小、采集的频率和传输的方式等内容。

通过对采集的分类、数据的类型、文件的大小、采集的频率和传输的方式的分析，可以得知，数据处理架构在数据采集和文件传输上有较大的提升空间。例如，增加自动上传、断点续传、传输监控等方式提高数据的采集和传输效率。

表 2-3 为某银行的数据采集现状分析。

表 2-3　某银行的数据采集现状分析

采集的分类	数据的类型	文件的大小	采集的频率	传输的方式
银行报送的数据文件	个人客户基本信息	10 MB	按天	通过数据库工具 export/Import 导出、导入
银行报送的数据文件	企业客户基本信息	12 MB	按周	通过数据库工具 export/Import 导出、导入
银行报送的数据文件	雇员基本信息	15 MB	按月	通过数据库工具 export/Import 导出、导入
银行报送的数据文件	银行财务预算	2 MB	按天	通过数据库工具 export/Import 导出、导入
银行报送的数据文件	贷款合同信息	5 MB	按月	通过数据库工具 export/Import 导出、导入
银行报送的数据文件	担保合同信息	12 MB	按月	通过数据库工具 export/Import 导出、导入

二、数据加载现状分析

对于数据加载的现状分析包括数据校验、数据加载入库等几个部分。

（1）数据校验的现状分析

数据校验包括对文件的格式校验和逻辑校验，一般来说，只有通过格式校验后，才能进入逻辑校验过程。当数据文件通过数据校验后，再直接加载到数据库中，如图 2-6 所示。

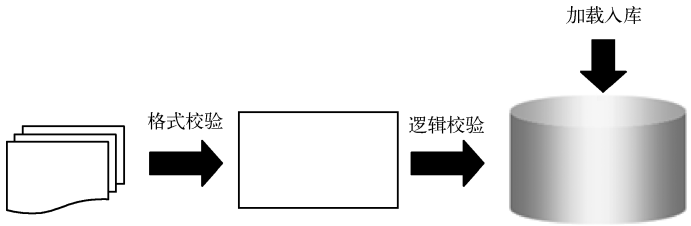


图 2-6　数据加载的现状分析

（2）数据加载的现状分析

基于数据加载的现状，可以从灵活性、扩展性和高效性上分析系统可能存在哪些问题。

1）灵活性。分析系统的现状是否可以支持对校验的灵活配置。

2）扩展性。分析系统现状的情况，包括数据的逻辑校验和入库处理方式是什么、是否具有可扩展性、是否是系统性能的瓶颈。

3）高效性。分析数据加载过程是串行处理方式还是并行处理方式、对于数据的校验是批量校验还是一条条校验，以及是否具有高效性。

三、数据处理现状分析

判断系统是否进行了身份信息类的加工和整合。例如，包括对身份信息的识别和归并，对各种规则进行有效匹配，列出疑似名单，然后通过技术手段或者人工确认的方式对身份信息进行确认。如图 2-7 所示，可以采取这种方式进行客户身份整合，获取唯一客户信息。

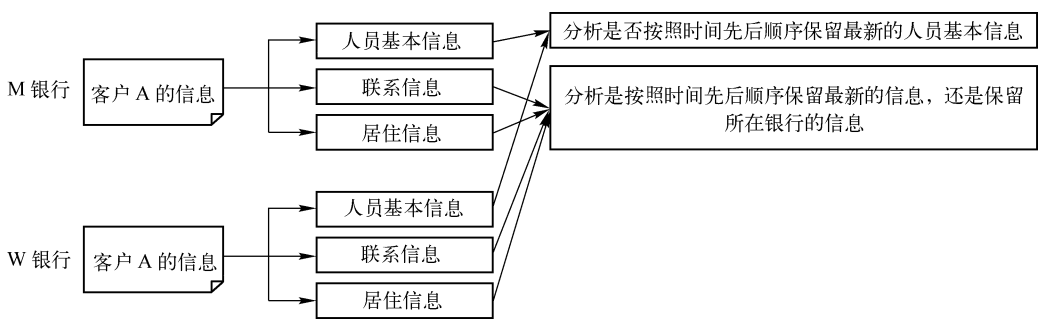


图 2-7　数据处理现状的分析

同时为了保证客户的完整性、准确性和反映客户当前信息，也可以参考如图 2-8 所示的这种方式，多个银行的同一客户信息，经过唯一码分配的过程，包括数据标准化、清洗、算法匹配和分配唯一码，再经过数据加工的过程形成唯一真实的客户信息。其中完整性是指包含业务所需的所有客户属性，准确性是指每个属性均反映客户的真实信息。

四、数据加工现状分析

考虑数据加工存在哪些问题：

1）数据加工是否进行了整体的规划和通盘的考虑，如将相同的数据加工抽象成公共数据加工。

2）判断相同的数据是否存在多次抽取的情况，是否存在数据不一致的风险。

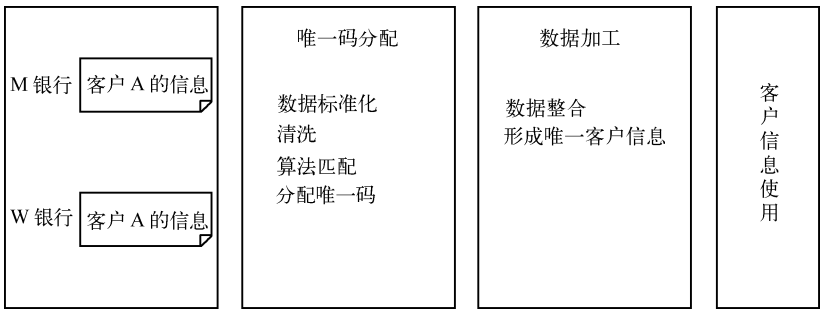


图 2-8　数据处理参考

五、数据迁移现状分析

考虑现状数据迁移可能存在哪些问题：

1）判断系统是否存在同一数据源反复抽取数据到多个目标库的情况，这种迁移方式会有数据不一致的风险。

2）判断系统是否对数据迁移进行了统一管理和维护，避免不必要的迁移过程。

通过对相关负责人员的访谈，以及对数据分布和流转现状的分析，我们归纳总结了数据架构规划的关键问题，判断数据架构总体架构原则是否缺失。

我们从数据采集、数据加载、数据处理、数据加工和数据迁移等几个方面对数据处理架构现状进行说明，说明现状系统中存在哪些问题和可以改进的地方。

例如，数据处理架构可能存在以下几种问题：

1）判断数据加载高效性、灵活性和可扩展性是否存在问题。

2）是否具有统一的数据加工规划，数据迁移是否有统一的调度。

如果存在上述问题，可以通过增加数据缓冲区，避免多个目标数据从同一数据源重复抽取数据，降低对数据源的影响和数据不一致性的风险。

例如，通过使用数据迁移工具，增强对数据转换和迁移的统一管理，避免重复的工作。当大量的数据从一个库迁移到另一个库，会影响数据的一致性，导致数据冗余度高，影响效率和导致时间窗口过长的问题，特别是如果某个数据没有明显的加工要求和应用要求，从一个库不停地流转到另一个库，会导致迁移的数据量很大，影响性能和数据的不一致性，所以尽量减少数据的全量迁移。

2.4　数据治理现状分析

数据治理现状分析框架，主要用于帮助系统对数据治理现状进行分析，一般包括数据治理机制和数据治理领域两个部分。数据治理领域可以包括数据质量、数据生命周期、数据标准和元数据管理，如图 2-9 所示。数据治理机制包括政策、组织、流程和技术工具等 4 个方面。

下面先谈一下数据治理领域：

（1）数据质量

对于数据质量来说，通过使用技术工具解决数据质量问题，通过改善和提高组织的管理

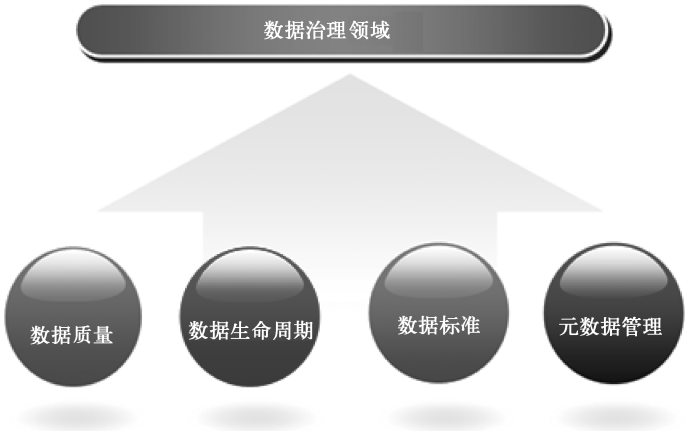


图 2-9　数据治理领域

水平，执行相关的政策和流程，使得数据质量得到进一步的提高。

（2）数据生命周期

对于数据生命周期来说，可以划分为 4 个阶段来描述数据的生命周期，包括数据创建、数据使用、数据归档和数据销毁。然后通过使用技术工具解决 4 个阶段的问题，通过改善和提高组织的管理水平，执行相关的政策，加强对数据生命周期的管理。

（3）数据标准

对于数据标准来说，它通过建立数据规范、政策体系、组织、管控流程和使用相应的技术工具来确保系统内重要核心的数据是一致和准确的。数据标准是企业级的数据定义，企业内所有的系统都应该遵守和执行数据标准。

（4）元数据管理

对于元数据管理来说，它通过建立数据规范、政策体系、组织、管控流程和使用相应的技术工具来满足对元数据的管理。通过元数据管理可以了解数据的变化过程，包括这些变化会给系统带来什么影响。

我们从政策、组织、流程、技术工具 4 个方面对数据质量、数据生命周期、数据标准和元数据管理进行分析，如图 2-10 所示。

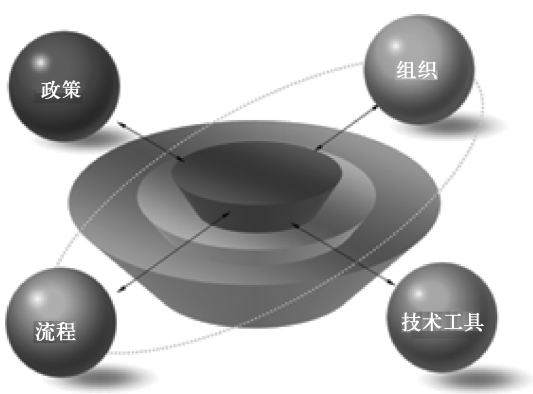


图 2-10　分析的维度

（1）政策

通过制定相应的政策明确相关部门的责任，明确数据治理各个领域的政策和规范，通过政策的制定去规范相关人员的行为。

（2）组织

通过建立明确的组织架构和人员角色，明确数据治理相关责任人，定义不同责任人的职责。

（3）流程

通过制定数据治理各个领域的工作方法和步骤，明确相关人员的分工和协作关系。

（4）技术工具

通过技术工具保证数据质量的管理，支持数据标准和元数据的发布和查询等流程。对数据生命周期进行管理。

2.4.1 数据质量管理现状分析

数据质量管理现状分析包括政策、组织、流程和技术工具现状分析。

（1）数据质量管理政策现状分析

判断是否建立了完整的数据质量管理政策体系。

（2）数据质量管理组织现状分析

判断是否建立了完整的数据质量管理组织，如数据质量管理的组织包括业务部门和客服部门。业务部门的职责是质量验收管理、数据质量量化考评、数据质量现场监测、数据质量量化考评、数据质量反馈管理和日常数据质量管理等内容；客服部门的职责是制定数据质量处理规范和负责客户关于数据质量问题的咨询和服务，并且对问题进行跟踪。

（3）数据质量管理流程现状分析

判断系统是否建立了完整的数据质量管理流程。例如，数据质量事前防范、加工处理质量监控和入库后事后治理。

1）数据质量事前防范。先对数据的接口程序进行测试和验收，例如按照某个测试标准，完成测试报告，对测试结果进行验证，根据验证结果判断验收是否通过。对于未通过验收的数据接口程序，将发现的问题反馈给相关机构或者人员，并指导其解决问题。

当修改完数据接口程序后，需要重新进行验证和测试，当完成测试后，重新申请验收流程。可以通过搭建测试环境，专门用于对数据的测试和验证工作，增强对数据质量的事前防范工作。

2）加工处理质量监控。在数据加工处理过程中，对数据进行预处理校验和入库校验，保证合格的数据能够入库，不合格的数据反馈给相应的机构，然后根据数据质量检查规则，检验入库的数据是否正确。可以通过提高数据自动化的程度，优化数据加载功能，实现自动调度加载；优化原有反馈渠道，提高数据报送自动化程度。尽量减少未知错误的反馈，降低错误数据的更正难度。

3）入库后事后治理。入库后事后治理可以包括两端数据核对，对数据质量进行现场监测，对有异议的数据进行分析，目的是不断提高数据的质量，减少异议情况的发生。我们建议构建数据管理平台完成对系统数据质量的统计分析工作，清楚掌握数据质量状况，从而提高工作效率，更好推进数据质量工作。例如增加以下几个功能：两端数据明细核对功能、定

点监测功能、历史处理情况查询功能、数据统计与分析功能、数据提取与反馈功能、数据质量档案管理功能、异常数据核实工作管理功能、数据质量统计报表功能、文档查阅功能、问题在线解答功能。

（4）数据质量管理技术工具现状分析

数据质量管理技术工具不作为本书重点。

综上所述，我们可以参考先进实践经验，判断系统的数据质量管理还存在哪些问题和差距。数据质量的提升和检查过程不是一蹴而就的，而是一个不断提升和改进的过程，同时数据质量管理不仅仅是一个技术问题，它更是一个管理问题，需要技术人员和业务人员互相配合，制定规则和管理流程。

2.4.2 数据生命周期管理

完整的数据生命周期管理涵盖数据从产生到销毁的全过程。

（1）数据生命周期管理政策现状分析

判断该系统是否建立了完整的数据生命周期政策体系，如在数据创建、数据使用过程中是否建立了相应的接口规范。在数据归档和数据销毁过程中是否有相应的数据生命周期管理方法和实施细则等内容。

（2）数据生命周期管理组织现状分析

判断系统是否建立完整的数据生命周期管理流程。例如，分析数据生命周期管理流程在数据创建、数据使用、数据归档和数据销毁过程中，有哪些组织架构和人员进行专项管理。

（3）数据生命周期管理流程现状分析

判断系统是否建立完整的数据生命周期管理流程。例如，分析数据生命周期管理是否具有数据的评估、管理手段设计和落地执行流程。

数据生命周期重点关注的部分主要包括数据创建、数据使用、数据归档、数据销毁。

- 数据创建

通过建立数据标准，保证数据的准确性。通过数据质量管理保证数据创建的准确性。

- 数据使用

在数据使用过程中，可以利用元数据管理监控数据的使用过程，利用数据标准保证数据的准确性。利用数据质量管理保证数据加工的准确性。

- 数据归档

通过数据生命周期评估手段，评估数据什么时候归档。

- 数据销毁

通过数据生命周期评估手段，评估数据什么时候销毁。

数据生命周期可以满足审计管理的需求，减少数据的冗余度，提高数据的一致性，同时减少数据的存储，提升系统的性能。

2.4.3 数据标准管理

数据标准管理现状分析主要包括数据标准管理政策现状分析、数据标准管理组织现状分析、数据标准管理流程现状分析、数据标准管理技术工具现状分析。数据标准是企业级的数据定义，企业所有的系统都应遵守和执行数据标准。

（1）数据标准管理政策现状分析

判断该系统是否建立了完整的数据标准政策体系。例如，在数据标准的建设过程中是否建立了相应的管理政策，数据是否得到了统一的定义。

（2）数据标准管理组织现状分析

判断系统是否建立完整的数据标准管理流程。例如，分析数据标准管理流程中有哪些组织架构和人员进行专项管理。

（3）数据标准管理流程现状分析

判断系统是否建立了完整的数据标准管理流程。

（4）数据标准管理技术工具现状分析

数据标准管理技术工具现状分析不是本书重点。

2.4.4 元数据管理

元数据管理现状分析主要包括：元数据管理政策现状分析、元数据管理组织现状分析、元数据管理流程现状分析、元数据管理技术工具现状分析。

（1）元数据管理政策体系现状分析

判断企业是否建立了完整的元数据管理政策。

（2）元数据管理组织现状分析

判断企业是否建立了完整的组织架构。例如，分析元数据管理流程中有哪些组织架构和人员进行专项管理。

（3）元数据管理流程现状分析

判断企业是否建立了完整的元数据管理流程。

（4）元数据管理技术工具现状分析

元数据管理工具现状分析不是本书重点。

元数据是“描述数据的数据”。一般来说，元数据就是用来描述上下文的信息，帮助人们更好地理解和使用数据。

元数据的分类包括：业务元数据、技术元数据和管理元数据。

（1）业务元数据

业务元数据是指从业务角度描述业务领域相关的概念、关系和规则的数据，主要包括业务术语和业务规则等信息。

（2）技术元数据

技术元数据是指描述系统中技术细节相关的概念、关系和规则的数据，主要包括对数据结构、数据处理方面的描述，以及数据仓库、ETL、前端展现等技术细节方面的信息。

（3）管理元数据

管理元数据是指描述管理领域相关的概念、关系和规则的数据，主要包括管理流程、人员组织和角色职责等信息。

2.5 数据架构现状要点分析总结

我们从几个方面分析数据架构是否存在问题：数据架构的合理性、数据模型的合理性、

数据的交互和加工环节是否畅通、数据的处理效率、是否满足数据源采集的灵活性、是否具有完善的数据治理框架等。

（1）数据架构的合理性

主要判断数据架构的设计能否适用于系统的使用，可以采集需要的信息，并加工成不同的产品。

（2）数据模型的合理性

判断数据模型是否适应功能的扩展性和对新业务的支持。

（3）数据加工环节是否畅通

判断系统之间的信息能否互相沟通，针对数据加工和处理的要求，能否在最短时间内，把需要的数据汇总和加工。同时需要考虑数据分析的维度和粒度问题。

（4）数据的处理效率

需要考虑数据加载方面，包括数据量的大小和数据的运算能力。还需考虑数据是否可以快速入库。在提高效率和处理模式上，是否使用多个进程并行处理的方式。

（5）是否满足数据源采集的灵活性

判断系统是否可以根据业务的需求采集结构化、半结构化和非结构化的数据。在数据采集的深度上，是否可以扩大采集范围，能够覆盖整个业务，进而满足数据采集的灵活性。

（6）是否具有完善的数据治理框架

对于数据标准的建设，是否形成统一、有效的数据标准，以保证参与信息的稳定性和完整性，是否保证历史数据变更的可追溯性。对于数据质量的检查，要求全面性、及时性和准确性等内容。

小结

- 数据架构现状的分析主要通过现状调研、资料的分析、高层领导访谈或者是对业务部门的访谈，了解数据架构的现状。现状分析，主要以发现问题、分析问题为主，在理解现状的基础上，借鉴行业内先进的经验，从 4 个方面对数据现状进行对比，从而发现数据架构存在哪些问题，同时提出改进的方向。把发现的问题作为未来架构规划的依据。
- 数据分类是按照选定的属性（或特征）区分分类对象，将具有某种共同属性（或特征）的分类对象集合在一起的过程。
- 数据分类最终可以形成数据大类和数据小类，数据大类是从全局角度理解业务，数据小类是从微观角度对同一大类的进一步细分。
- 参考最佳行业实践和 Teradata 金融业逻辑数据模型，可以将数据大类分成 8 个部分：当事人、产品、渠道、合约、财务、机构、事件、活动。
- 数据处理架构可以分成数据源层、数据交换层、数据基础层、数据加工层和应用层等几个部分。
- 数据采集现状分析包括数据报送和上传的现状，包括采集的分类、数据的类型、文件的大小、采集的频率和传输的方式等内容。
- 对于数据加载的现状分析，包括数据校验、数据加载入库等几个部分。

- 数据治理领域可以包括数据质量、数据生命周期、数据标准和元数据管理。数据治理机制包括政策、组织、流程和技术工具等 4 个方面。
- 数据治理现状分析框架，主要用于帮助系统对数据治理现状进行分析，一般包括数据治理机制和数据治理领域两个部分。
- 数据质量管理现状分析包括数据质量管理政策现状分析、数据质量管理组织现状分析、数据质量管理流程现状分析和数据质量管理技术工具现状分析。
- 对于数据质量来说，通过使用技术工具解决数据质量问题，通过改善和提高组织的管理水平，执行相关的政策和流程，使得数据质量得到进一步的提高。
- 数据生命周期管理现状分析主要包括数据生命周期管理政策现状分析、数据生命周期管理组织现状分析、数据生命周期管理流程现状分析、数据生命周期管理技术工具现状分析。
- 对于数据生命周期来说，可以划分为 4 个阶段来描述数据的生命周期，包括数据创建、数据使用、数据归档和数据销毁。然后通过使用技术工具解决 4 个阶段的问题，通过改善和提高组织的管理水平，执行相关的政策，加强对数据生命周期的管理。
- 数据标准管理现状分析主要包括数据标准管理政策现状分析、数据标准管理组织现状分析、数据标准管理流程现状分析、数据标准管理技术工具现状分析。
- 对于数据标准来说，它通过建立数据规范、政策体系、组织、管控流程和使用相应的技术工具来确保系统内重要核心的数据是一致和准确的。数据标准是企业级的数据定义，企业内所有的系统都应该遵守和执行数据标准。
- 元数据管理现状分析主要包括元数据管理政策现状分析、元数据管理组织现状分析、元数据管理流程现状分析、元数据管理技术工具现状分析。
- 对于元数据管理来说，它通过建立数据规范、政策体系、组织、管控流程和使用相应的技术工具来满足对元数据的管理。通过元数据管理可以了解数据的变化过程，包括这些变化会给系统带来什么影响。
- 元数据是“描述数据的数据”。一般来说，元数据就是用来描述上下文的信息，帮助人们更好地理解和使用数据。
- 元数据的分类包括业务元数据、技术元数据和管理元数据。
- 我们从几个方面分析数据架构是否存在问题：数据架构的合理性、数据模型的合理性、数据的交互和加工环节是否畅通、数据的处理效率、是否满足数据源采集的灵活性、是否具有完善的数据治理框架、是否建立数据标准体系、是否有完整的数据生命周期体系和数据质量管理体系是否完善等。

第 3 章　数据架构目标规划

本章目标

通过前一章的学习，我们已经理解了数据架构现状分析的工作方法，以及数据架构现状分析的相关案例。本章将重点介绍如何在现状分析的基础上，对目标数据架构的建设，包括数据模型的建设、目标数据架构分布和流转的规划等内容。

学习本章后，读者将掌握：

- 数据架构的工作方法和指导原则
- 针对数据架构现状的总结
- 提出数据架构的改进方向
- 概念模型的建设
- 数据分类的规划
- 逻辑模型的建设
- 物理模型的建设
- 未来数据架构的分布
- 目标数据架构的流转
- 数据归档
- 对数据架构的验证

3.1　数据架构理论体系概述

数据架构理论体系是把业务和技术融合到一起的一套体系。它包括技术、方法和相应的管理过程。经过几十年的发展，数据架构已经形成了完整的理论体系。

什么是数据架构呢？

数据架构是企业架构的重要组成部分，实现数据的合理组织和共享，保证数据在系统之间的一致性、完整性、安全性和正确性。一般来说，数据架构包含数据模型和分类、数据分布和流转等内容。

对于数据治理来说，它是为了提升数据架构各个层次的管控和协作能力。同时数据架构为数据治理提供基础能力支撑，数据治理与数据架构是相辅相成的。数据治理包含数据质量管理、数据生命周期管理、数据标准、元数据管理等多个管控专项。数据治理会在下一章节详细介绍。

（1）数据模型

数据模型是指用实体、属性及其关系对企业运营和管理过程中涉及的业务概念和逻辑规则进行统一定义、命名和编码。

（2）数据分类

数据分类是根据业务特征对数据进行归类和划分，并用层级列表的方式展示数据内容，

数据分类的规范需要满足各种业务需求对数据组织的要求。

（3）数据分布

数据分布主要包括业务分布与系统分布。数据分布主要分析数据在各个环节中的创建、引用、更新和删除，并根据业务对数据的处理特点，合理规划数据的分布。在规划数据分布的时候，需要考虑如图 3-1 所示的几个方面。

图 3-1　规划数据分布需要考虑的内容

3.1.1　数据架构的工作方法和指导原则

在第 2 章中，我们了解了数据架构现状分析的方法，那么如何与需求结合起来，并且对目标数据架构进行规划呢？

数据架构的工作方法就是参考数据架构的原则，在理解现状问题和改进方向之后，在需求要点的基础上结合最佳实践进行目标数据架构的规划，如图 3-2 所示。

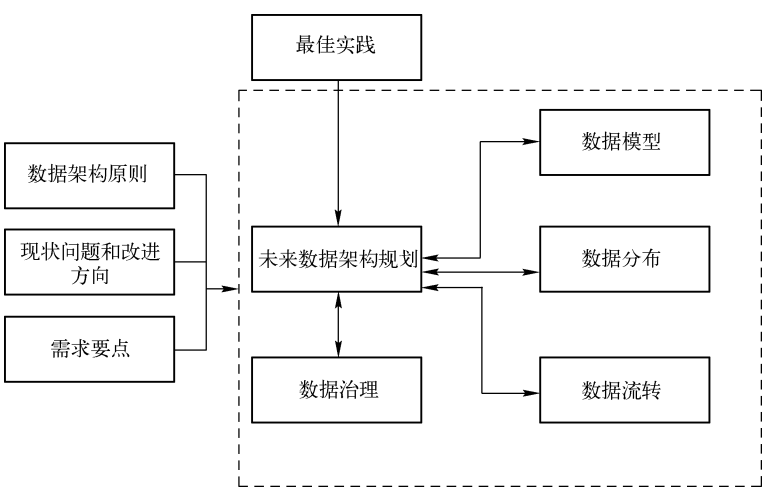


图 3-2　数据架构的工作方法

在数据架构规划中需要保证数据的安全性、可用性、完整性、真实性和抗抵赖性。

（1）安全性

安全性是指在数据处理中保密，传输及存储中加密。

（2）可用性

可用性是指提供数据备份和恢复功能。

（3）完整性

完整性是指在处理、传输和存储过程中校验完整性。

（4）真实性

真实性是指在传输和处理前识别数据的真实性。

（5）抗抵赖性

抗抵赖性是指传输和处理中保证不可否认性。

因为数据架构和系统程序设计上本身就有妥协的成分，例如为了某种优化，是放在整体架构上去解决还是从程序上去调整和完善呢？

这并没有一个确定的答案，这需要根据不同的场景去考虑，因此，需要一个策略和标准去指导什么问题需要在架构上考虑，什么问题需要在系统详细设计上考虑。以下是数据架构的指导原则。

（1）灵活性原则

数据架构要充分考虑灵活性，满足不同的业务需求，以适应业务的变更。

（2）高效性原则

保证数据校验、加载、迁移、加工的高效性，支持产品的快速生成。

（3）可扩展性原则

数据架构需要考虑未来的可扩展性，当需求发生变化的时候，尽量减少对数据架构的变更。

（4）数据共享原则

提高数据公共加工的功能，保证相同指标加工的唯一性，最大程度地共享公共加工的结果。

（5）数据可用性原则

对数据的采集应该满足业务的需求。

（6）数据安全性原则

数据按照非功能性属性制定不同的安全级别，并区分敏感数据和非敏感数据。

3.1.2 针对数据架构现状的总结

对于数据架构来说，可以从几个方面去了解现状存在的问题是什么。例如，判断数据架构的原则是否清晰、架构层次的划分是否合理等内容。

（1）数据架构的原则是否清晰

判断现状中作为数据架构设计的指导原则是否清晰，是否能成为数据架构和数据治理可以遵循的依据。

（2）架构层次的划分是否合理

从数据分布、数据流转的角度判断当前的数据架构是否合理。

例如，对于数据分布来说，是否有缺失的层级，数据的分布是否混乱，该分布是否引起效率的问题。对于数据流转来说，是否过于重叠、复杂，是否有数据不一致的风险。

（3）数据采集方式

对于数据采集来说，我们需要了解采集的方式是什么，例如是采用中间件的方式还是HTTP 的方式，采集的对象包括什么，以哪类信息为主，数据采集的时间周期是什么，数据

的采集能否满足扩展性、灵活性和高效性等特点。

同时需要考虑在安全上是否有提升的空间，是否有自动上传、断点续传和在数据传输过程中能够监控等内容。

（4）数据的校验、加载方式

数据校验一般分为格式校验和逻辑校验，我们需要了解格式校验的方式是什么，逻辑校验的规则有哪些等。对于数据加载来说，是否可以处理批量的加载和校验，是否能够在灵活性、扩展性和高效性上有提升的空间。

（5）数据、产品的整合和加工

我们需要了解数据整合、加工的粒度是多少，是否可以进行身份识别、疑似归并和对主数据的加工等方面。

例如，人员身份信息是以什么方式进行整合的，是否能满足对于同一个人、不同证件信息的整合和加工。对于产品加工来说，我们需要了解产品类型有哪些，是否存在“因事设库”的情况，对于相同的业务需求，是否存在重复抽取、重复加工的过程。在公共加工方面，是否有统一的规划、是否有提升的空间等内容。

3.1.3 需求要点

对采集的数据项进行分析，判断是否能满足对产品的加工需求，效率问题是否存在改善的空间，是否能够支持数据的快速入库，不同系统之间的数据是否可以共享，是否可以规划数据交换平台，提高数据加工的效率，保证数据架构满足灵活性、高效性和可扩展性。

3.1.4 数据架构的改进方向

可以参考数据架构的现状问题，提出对数据架构的改进方向。例如，首先应该明确数据架构总体指导原则和现存问题是什么，以此原则指导未来数据架构的建设，同时提出未来数据架构的改进方向是什么。最后明确数据架构的各个层级，以及对每个层级进行数据治理和管控。

3.2 数据模型

数据模型是对数据特征的抽象，它一般分为概念模型、逻辑模型和物理模型。概念模型是以数据分类的形式体现，而逻辑模型以 ER 图的形式展示。

3.2.1 概念模型

什么是概念模型？

概念模型是从业务的角度对数据进行抽象，包括业务层面上主题域的划分，以及各个主题域下的数据分类，和基于分类的非功能属性。

3.2.2 数据分类

什么是数据分类？

数据分类是以业务特征对数据进行归类和划分，一般用层级列表的方式展现数据内容，数据分类是概念模型的体现。数据分类可以促进业务人员和技术人员之间的沟通，指导数据

的分布和流转。

什么是主题域？

主题域是从较高层级上对业务的抽象和归纳，从概念层面对系统的全面描述，主题域主要考虑业务扩展性，主题域划定后，较少变更。

主题域下的数据分类是什么？

分析数据的非功能特性，未来架构的数据分类从较细维度进行划分，保证已有的数据分类较少变化。当有新业务扩展时，可以增加新的数据分类。

一、数据分类的指导原则和非功能属性

1. 数据分类的指导原则

对业务数据进行主题域及主题域下的划分，需要遵循如下几个原则：业务驱动性、完整性原则，分类通用性、互斥性原则，非功能属性一致性原则，排除衍生数据原则，分类关联性、可理解性原则等内容，如图 3–3 所示。

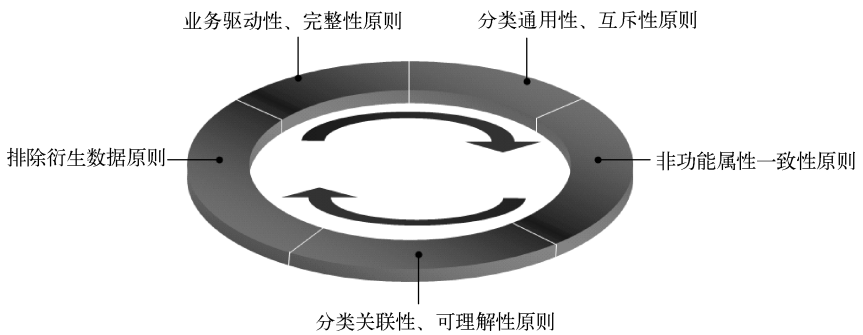


图 3–3　数据分类需要遵循的原则

（1）业务驱动性、完整性原则

信息项的设立从业务特性出发，不考虑技术及落地实现。数据分类做到全面、完整，保证对业务的完整覆盖。

（2）分类通用性、互斥性原则

数据分类尽可能支持业务多变性，力求以最少改动支持业务变更，数据分类相互之间不能包含相同数据内容。

（3）非功能属性一致性原则

数据分类包含的所有信息项对应的非功能属性应该一致。

（4）排除衍生数据原则

分类信息不包括衍生数据。

（5）分类关联性、可理解性原则

数据分类，同一类下数据项应有关联性。分类应做到定义清晰、无二义性。

2. 数据分类的非功能性属性

针对主题域下的数据分类，需要从变动频率、变动量、变动模式、数据量大小、格式、共享性等各个维度进行分析。数据分类的非功能属性对于数据分布的设计具有重要的参考意义。如图 3–4 所示，数据分类的非功能性属性主要包括数据量大小、格式、共享性、变动频率、变动量、变动模式等内容。

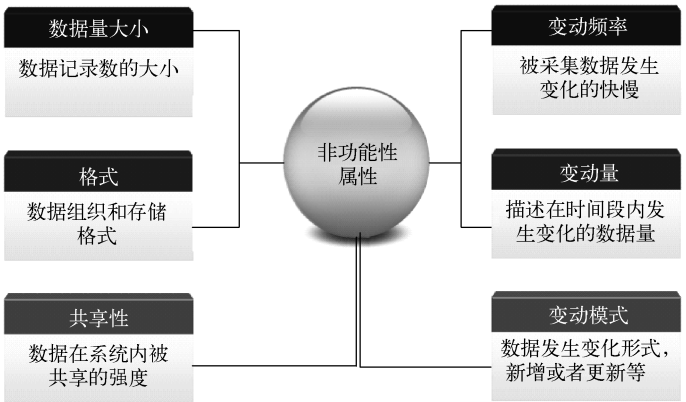


图 3-4　数据分类的非功能性属性

（1）数据量大小

数据量大小可以分为大、中、小三类。划分的方法根据实际需求不断调整。例如，10 亿条记录以上的，数据量为大；1 亿 ~10 亿条之间的，数据量为中；1 亿条记录以下的，数据量为小。

（2）格式

数据的格式有结构化数据、半结构化数据和非结构化数据。所谓结构化数据是以二维表格形式进行逻辑表达存储的数据。半结构化数据包括一些文本文件、文档。非结构化数据包括图片、图像和音频/视频信息等。

（3）共享性

数据共享性可以分为较高、一般、较低。例如，一些主体信息在各个业务模块共享的需求较高。对于一些特定业务领域的数据，共享性要求较低。

（4）变动频率

变动频率可以分成极少、偶尔和固定周期。例如，我们可以把固定不变的或者年变动率非常低的，如姓名、身份证信息和组织机构号等信息归为变动频率极少发生变化的一类。

从业务角度出发，数据存在变动的可能，而且变动时间不可预知。例如，地址信息和电话信息等内容，这些信息归到变动频率偶尔发生变化的一类。对于一些数据按照固定周期变更，如还款、扣收等内容，可以归为变动频率在固定周期内发生变化的一类。

（5）变动量

以年或者月为基础对数据的变动量进行估值。

（6）变动模式

变动模式分成增加、更新和删除模式。增加是以新增方式产生数据，如业务交易类信息。更新是数据存在更新的可能，如企业规模、联系方式等。

二、数据分类举例

对于数据分类，我们以金融逻辑模型为例进行说明：

参考 Teradata 金融业逻辑数据模型，分成当事人和当事人角色、产品、协议、事件、地域、金融资产。

（1）当事人和当事人角色

银行所服务的对象和感兴趣进行分析的对象，如个人或公司客户、雇员等信息。

（2）产品

产品是金融机构向用户销售的或者提供给客户的服务。

（3）协议

金融机构与当事人之间针对某种特定产品或者服务而签订的合约关系，如客户和银行签订的合同等内容。

（4）事件

记录与银行相关的活动的详细情况。可以由客户发起，也可以由银行发起。

（5）地域

观察和分析的区域，包括传统的地址信息。

（6）金融资产

可以包括客户的资产（负债）信息。

金融数据模型如图 3-5 所示，是指当事人之间针对某种特定产品或者服务而签订的协议关系，协议内容被加工成产品，事件的发生基于协议内容，协议自动触发事件。

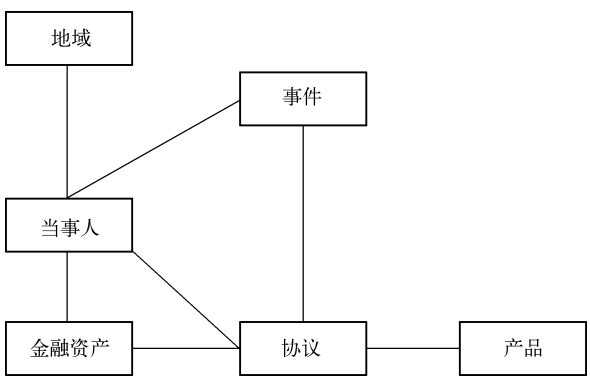


图 3-5　金融数据模型

3.2.3　逻辑模型

什么是逻辑模型呢？

逻辑模型是用来发现、记录和沟通业务的详细“蓝图”，由一系列表和实体详细描述组成，是通用的业务语言，便于业务与业务之间的功能理解，遵循第三范式。它包括主题域的设计、基本实体的设计和主要属性的设计，是 IT 人员和业务人员沟通的工具和桥梁。逻辑模型建设的一般步骤，如图 3-6 所示，首先分析需求，选择感兴趣的数据，然后在实体中增加属性，进行粒度层次的划分，最后进行关系模式的定义。

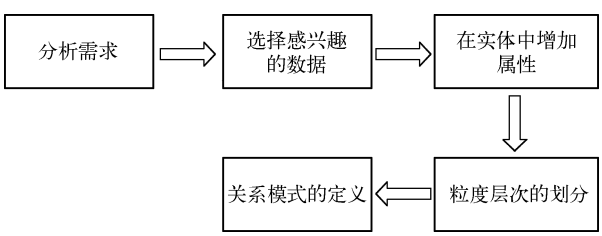


图 3-6　逻辑模型建设的一般步骤

3.2.4 物理模型

物理模型是逻辑模型针对具体实现环境的物理化，可以不遵循第三范式，主要包括实体属性的物理化，属性的长度、类型、主键、外键、索引等详细设计。物理模型主要是描述模型实体的细节，对列的属性进行明确的定义。物理模型的建设过程是在逻辑模型的基础上，为应用生产环境选取一个合适的物理结构的过程，包括存储结构和存储方法。

主要步骤如下：

1）实体名转变为表名。

2）属性名转换为列名，确定列的属性。

3.3 目标数据架构规划

3.3.1 目标数据架构的分析重点

一、非功能性指标

未来数据架构的建设需要考虑系统的非功能性指标，见表 3–1。

表 3–1 非功能性指标

指　　标	要　　求
数据加载	数据加载的效率从 XX 条/小时可以提高到多少
服务查询	系统最多的并发用户数是多少，响应时间是多少秒
数据加工	加工时间窗口是多少小时。处理能力是每小时能处理多少条记录
可用性	例如，系统可以达到 24 小时不停机

具体实现上述指标的做法可以有以下几种，如图 3–7 所示。

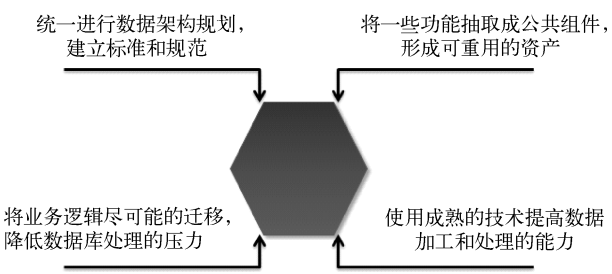


图 3–7 实现指标的方法

1）统一进行数据架构规划，建立标准和规范，在统一的平台进行设计、开发、测试和部署。

2）将一些功能抽取成公共组件，形成可重用的资产。

3）使用成熟的技术提高数据加工和处理的能力，支持对关键环节的并行处理能力，扩大数据处理和对外发布的可用时间。

4）将业务逻辑尽可能迁移，减小数据库处理的压力，提高系统并发处理的能力和可扩展的能力。

二、数据架构现存问题影响及分析

问题影响及分析主要包括对现状问题的描述，并且提出对系统的改进点。举例如下所示。

问题 1

系统的并行处理能力差，数据迁移期间无法对外发布服务，对外服务窗口时间缩短，资源使用不均衡，数据库服务器的压力过大。同时硬件资源使用情况不均衡，出现资源浪费的情况。

改进点

将业务处理逻辑拆分，减少数据库服务器的压力，提高应用的并行处理能力，增强对关键环节的并行处理能力，选择成熟的数据处理和加工技术，尽量做到数据采集、加工和对外服务的并行处理，减少数据处理环节间的技术依赖和约束。

问题 2

没有统一的技术开发框架平台，每个模块都有自己的开发框架，代码的可重用性降低，维护难度高。

改进点

制定统一的架构原则和方法，抽取公共组件。完善设计开发规范，形成统一的、完整的技术体系框架。

问题 3

未形成统一的数据采集技术支撑体系，特别是多渠道的、零散的对外采集子系统，增加了数据采集质量的管理难度。

改进点

形成统一的数据采集技术支撑体系，整合数据采集技术，实现自动化的数据采集功能，增加断点续传能力和数据传输监控能力。

问题 4

对数据校验、入库、加工处理和统计分析能力的不足。

改进点

可以引入 ETL 技术，满足数据处理和加工的工作要求。同时引入数据仓库的技术，提高对海量数据的统计分析能力。

三、未来数据架构的参考点

对于未来数据架构，可以参考以下思想内容：

首先强调数据的存储与流转，支持层次化的处理，包括对结构化数据与非结构化数据的处理能力。例如，数据架构的层次可以包括源数据、内容管理、数据交换、数据存储区、数据加工区和应用。下面对这几个层次进行说明。

（1）源数据

源数据可以包括如来自互联网、政府部门、同业、手工录入的信息。对于数据源来说，主要定义数据采集的来源、格式和采集方法等内容。

（2）内容管理

内容管理提供对各种非结构化数据的存储、访问和管理的能力。例如，对图像、音频信息和办公文档等信息的处理能力。为半结构化和非结构化数据提供捕获、管理、存储、保护和交付等方面的功能。

（3）数据交换

数据交换包括数据的抽取、订阅，以及 ETL 过程等内容。为系统与外部数据交换提供支持。

（4）数据存储区

例如 ODS、基础数据存储、非结构化数据存储。数据存储是保存从各个数据源采集的、贴数据源的、近期的数据和全量的基础数据，全量的基础数据将作为后续数据加工的唯一可信数据源。

（5）数据加工区

数据加工区包括数据仓库、主数据和查询库等。例如在主数据中进行身份信息整合加工。

（6）应用

应用包括查询类应用、分析类应用和管理类分析。根据参考架构进行目标架构的设计，未来数据架构是在参考架构的基础上，结合业务特点进行一系列的调整而成的。

四、对未来架构的解读

未来架构的重点在于对源数据层、内容管理、数据交换层、数据存储区、数据加工区、应用的分析和解读。

1. 源数据层

数据源层需要描述采集数据的类型，例如采集的数据一般分为结构化数据和非结构化数据，其中非结构化数据可以包括各种音频、图像、视频等信息。我们从以下几个角度对数据源层进行分析：数据来源、格式特征、数据量和频率等内容。如图 3-8 所示。

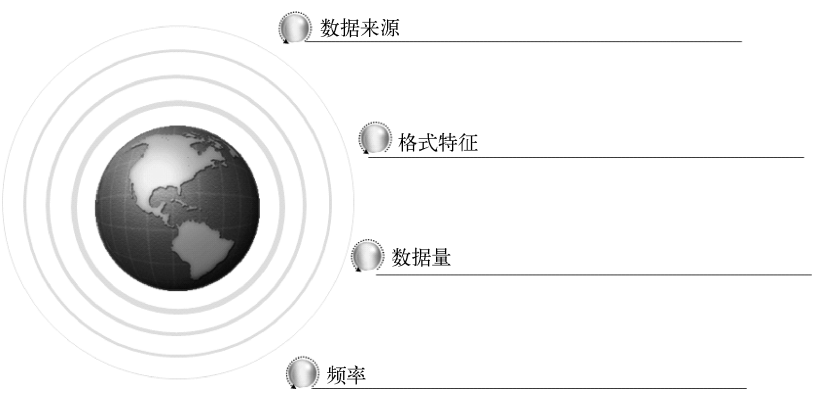


图 3-8　对源数据层进行分析

（1）数据来源

例如，从外部数据库，或以手工录入、网络爬虫等多种形式抽取数据。

（2）格式特征

判断数据的采集是以结构化数据为主还是以非结构化数据为主。

（3）频率

考虑多长时间生成数据。

（4）数据量

考虑采集的数据量是多少，是新增的数据量大还是更新的数据量大。

对源数据的采集，需要考虑对采集数据的唯一定位。举例来说，个人信息的采集需要考虑是否可以用姓名、证件类型、证件号码的方式对个人进行定位，因为同一个人的不同证件可能会定义成不同的实体。

2. 内容管理

内容管理主要提供对非结构化数据的存储、访问和管理功能。例如，系统可以从其他渠道采集非结构化数据，然后再通过标注或者文本挖掘技术，建立非结构化数据的元数据，在此基础上与结构化数据整合，再存储到数据仓库中，以供分析使用，或者对非结构化数据建立单独的分析应用。

具体做法是先将非结构化数据存储在库中，然后通过建立标签和摘要等方式获取结构化的信息，再利用数据交换层加载到数据缓存区中，最后加载到数据仓库中，以供分析使用。

3. 数据交换层

数据交换层承载着数据库之间的数据交换功能，交换层可以包括外部交换层和内部交换层。

一般来说，数据交换层包含 ETL 过程，数据的抽取、订阅，质量检查等功能，如图 3-9 所示。

图 3-9　数据交换层的功能

（1）ETL 过程

ETL 过程包括数据的抽取、清洗、转换和加载。在清洗过程中还包括数据的预处理校验、入库校验 、数据关联校验等内容，经过去重、合并、拆分、标准化和整合等过程，将转换后的数据加载到目标库中。

（2）数据的抽取、订阅功能

数据的抽取、订阅是为了从数据源层中获取原始数据，并且实现一源多目标的数据更新方式。如图 3-10 所示，抽取、订阅是可以实时或准实时、批量获取源系统的增量或全量数据，然后再根据不同的需求和业务规则将数据分发到不同的目标库中。

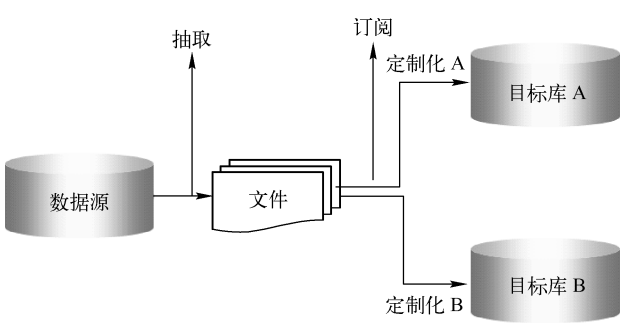


图 3-10　数据的抽取、订阅

（3）质量检查功能

数据质量检查是数据交换层的重要工作，包括验证数据的类型、格式、长度等内容，确保经过数据质量检查后，数据能够满足业务和技术对于数据的基本质量要求。经过数据交换层的质量检查后，可以生成一系列的文件，例如清洗的结果文件、记录清洗结果的报表文件和不合格文件等。

清洗结果文件是经过数据质量检查后，符合一致性、准确性和完整性的合格的文件，可以当做后续加工处理的唯一可信的输入文件。

记录清洗结果的报表文件包含了数据最原始的信息和清洗过程中的相关信息，例如数据不合格的原因、对数据不合格的标识等内容。

不合格文件是经过数据质量检查后，不符合数据一致性、准确性和完整性要求的数据，是没有通过质量检查的数据。

数据交换层关键设计：

数据交换层，包括外部数据交换和内部数据交换，支持系统内部和系统之间的数据在各个数据库之间的流转，如图 3-11 所示。

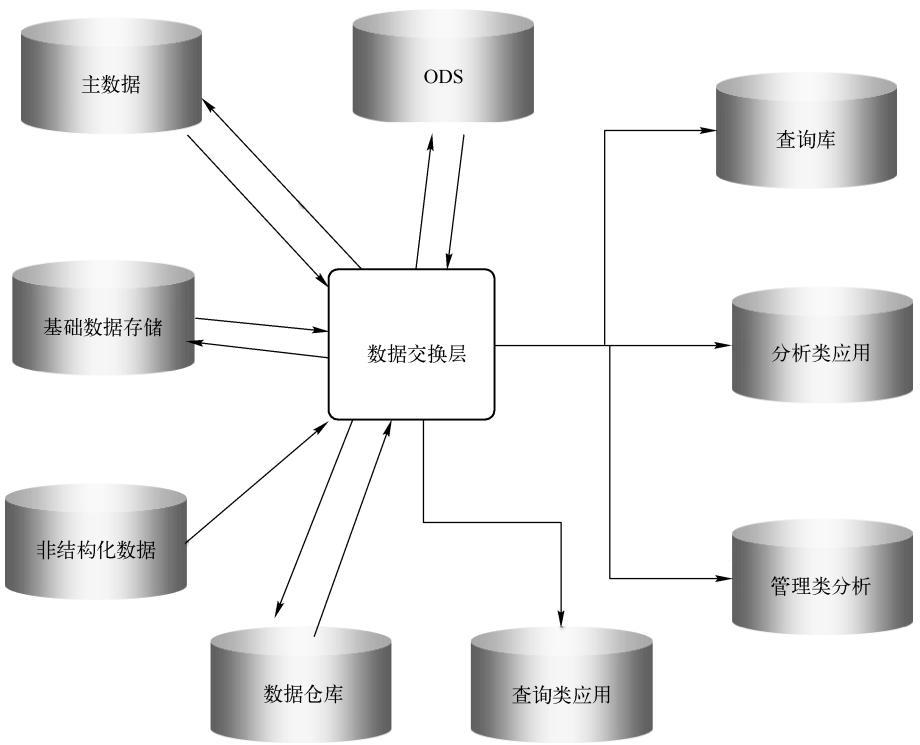


图 3-11　数据交换层

● 以增量的方式捕获数据

将源系统导出为增量文件，供后续加工和并行加载，用来提升效率。增量捕获的方式包括：触发器、时间戳、全表对比和系统日志分析的方式等。

● 提高数据交换的效率

通过细化作业任务，保证数据在传输过程中不执行加工操作，使传输和加工以并行的方

式进行，同时分析任务之间的关联关系，确定任务的调度机制。这些方式都有效地提高了数据交换的效率。

4. 数据存储区

数据存储区是对采集来的数据进行校验和存储，最后形成系统后续加工唯一可信的数据源。数据存储层包括 ODS、基础数据存储和非结构化数据存储，如图 3-12 所示。

（1）ODS

ODS 可以分成两部分内容，一个是临时缓冲区，另一个是加载区，如图 3-13 所示。

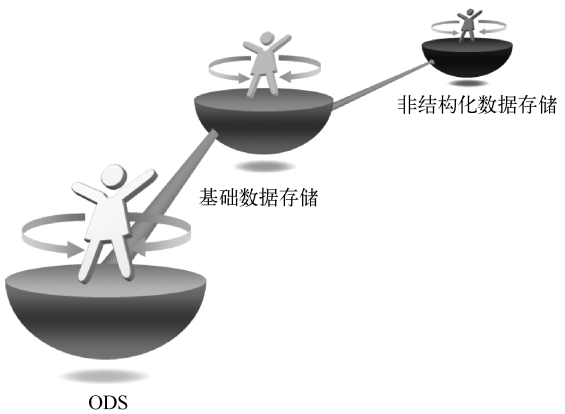


图 3-12　数据存储区

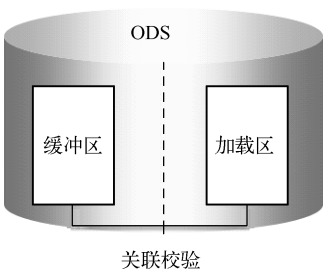


图 3-13　ODS

临时缓冲区是经过格式校验的数据缓冲区，它是贴数据源的数据存储。临时缓冲区的数据和加载区数据可以进行关联校验，如果满足逻辑校验的要求，则该新增数据直接插入到数据加载区，并且替换掉加载区上期的数据。

下面介绍一下 ODS 具有的特性：

首先，对于传统的 ODS 来说，它是面向主题的、即时的，也可以是贴数据源的，反映当前数据变化的内容。

ODS 保存最近一期的数据，为了快速生成查询报告，同时校验数据和对基础数据存储进行供数，提高对海量数据的快速加载和校验能力。

其次，对于数据的校验来说，加载区保存了上期的数据，根据逻辑校验的需求，可以包含贴数据源的数据，也可以对某些指标进行累计汇总。校验规则可以有以下几种：

1）新增数据和最近上期数据的关联校验。例如，对于本月还房贷的累计次数一定大于上一期的累计次数。

2）新增数据和累计汇总指标的关联校验。例如，贷款金额 - 贷款余额 ≤ 累计还款金额。

但是对于漏报补报的数据，一般来说，不具备关联校验的条件。

（2）基础数据存储

基础数据存储作为系统唯一可信的数据源，存储校验通过的数据，也存储非结构化数据结构化后的信息。基础数据存储可以实时批量地导出增量文件，以供后续加工使用，如图 3-14 所示。

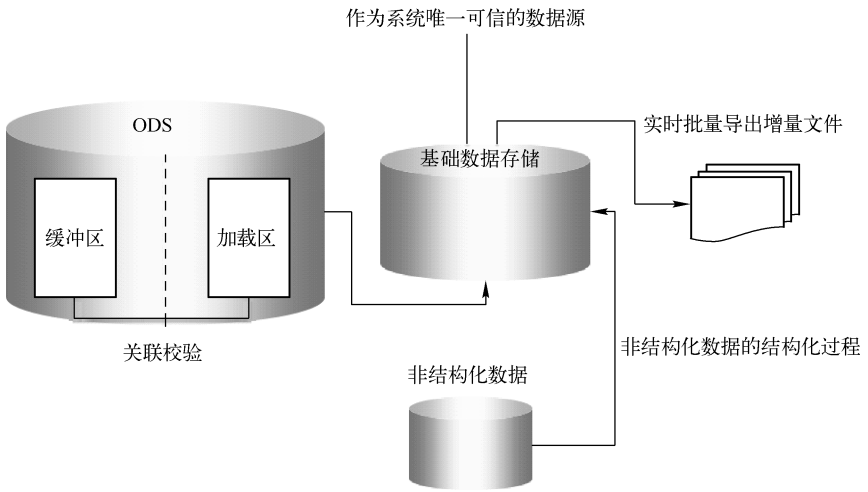


图 3-14　基础数据存储

（3）非结构化数据

非结构化数据是指存储经过处理后的非结构化数据。

5. 数据加工区

数据加工区的数据来源于基础数据存储，并将加工后的数据提供给应用层。数据加工区包括查询库、主数据和数据仓库，如图 3-15 所示。

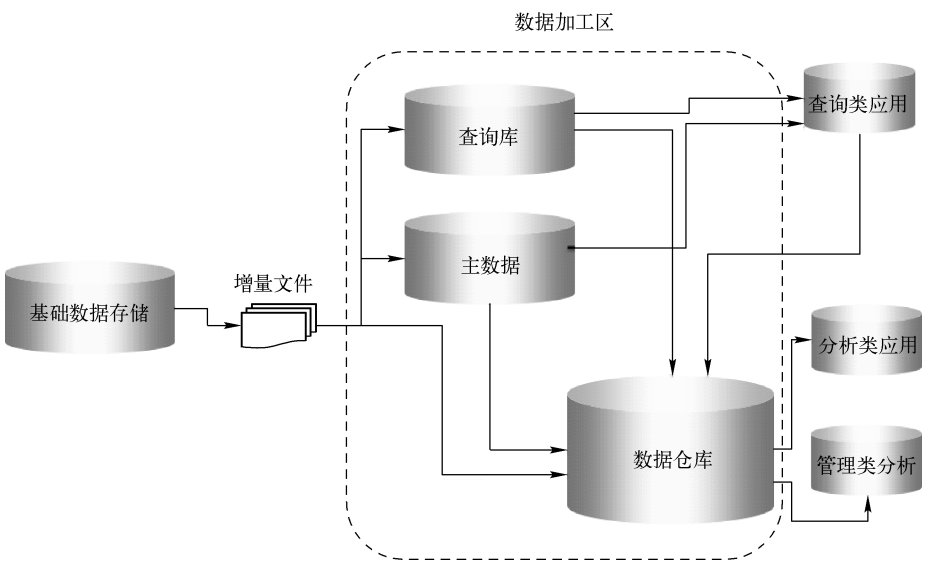


图 3-15　数据加工区

（1）查询库概述

对于查询库的产品，可以批量地将基础数据存储导出的增量文件加载到查询库中，然后再进行产品的加工。

（2）主数据概述

主数据是描述核心业务实体及其关系的数据，但是不是交易流水类的数据，主数据具备

共享价值、相对静态稳定的特点。在主数据中，包括对主体的识别和归并，也就是利用规则的识别、合并和覆盖处理，实现主体的唯一性，提高主体数据的可信度，并且使用唯一主体标识进行标识。

例如，对于个人的基本信息可以使用证件类型、证件号码和姓名作为唯一标识。其他信息均反映个人的真实信息。个人信息的识别过程如图 3-16 所示，证件类型、证件号码和姓名可以作为客户 A 的唯一标识。

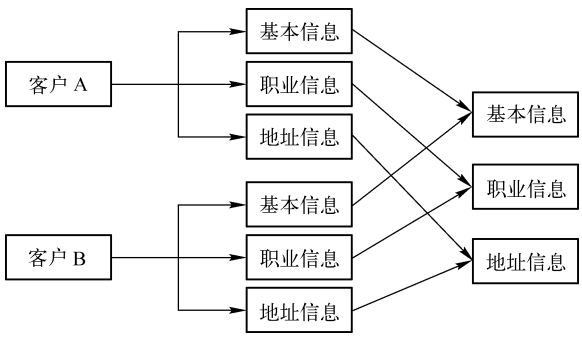


图 3-16　主数据

主体识别的详细过程：

因为国情的不同，有些国家的身份信息整合方式是以自然人为整合对象，主要利用姓名、证件号码、地址等信息进行整合，然后采用自主研发的数据匹配和整合技术，并且通过疑似查询、模糊匹配等先进手段，对信息进行整合。而国内可以采用证件类型、证件号码和姓名进行身份识别，对于识别出来的职业、地址等信息可以按照时间排序等手段来取舍。

例如，对于一些疑似身份信息的整合过程，包括：明确身份信息整合的规则定义、疑似身份信息清单的生成、疑似身份信息的整合及归并等内容。

举例来说，身份信息的疑似规则可以包括：姓名 + 手机号、姓名 + 出生日期等。疑似身份信息整合可以将疑似清单发送给相关人员进行确认。归并过程是将疑似身份信息清单进行合并和整合，如将地址信息、联系信息合并。

数据整合技术如图 3-17 所示。

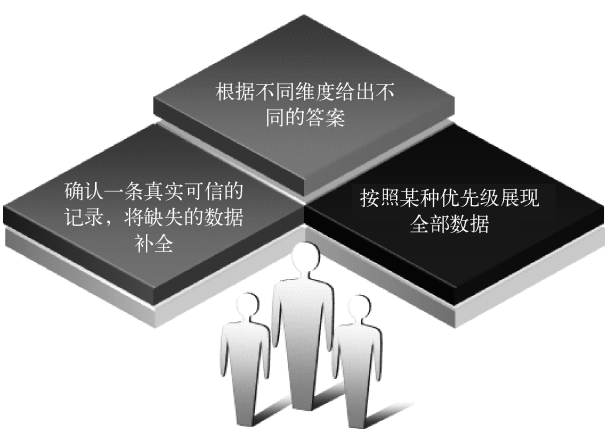


图 3-17　数据整合技术

1）对于多条信息的识别，首先根据规则确认一条真实可信的记录，然后将缺失的数据补全。

2）保存所有的信息，根据不同维度给出不同的答案。

3）保存所有的信息，按照某种优先级展现全部数据。

（3）数据仓库概述

数据仓库主要存储全局的信息。我们可以把数据仓库分成基础数据、汇总加工和库内集市，如图 3-18 所示。其中基础数据和汇总加工主要为库内集市提供数据。对于简单加工和以查询为主的数据服务，尽量不使用数据仓库。对于需要大量历史数据和复杂计算的，可以使用数据仓库。

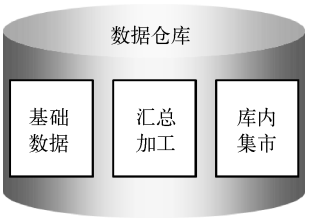


图 3-18　数据仓库

因为数据仓库通常包含历史数据，记录了各个阶段的历史信息，所以对查询的时效性要求不高。一般来说，数据仓库是不进行删除的。

（1）基础数据

基础数据存储了数据仓库最具细节性的数据，它可以来源于基础数据存储、主数据中的身份信息整合和查询类相关产品的信息等内容。一般来说，基础数据按照数据仓库模型进行组织，同时作为汇总加工层的数据源。

数据仓库中的基础数据和基础数据存储是有区别的。

1）首先，它们的目的不同，基础数据存储作为系统唯一可信的数据源，而数据仓库中的基础数据是为数据仓库后续加工考虑的。

2）然后，基础数据存储是贴数据源的，支持对各种产品的加工，时效性较高，并且对数据仓库供数。而数据仓库中的基础数据一般来说是按照第三范式进行存储的，它强调对各种数据的集成，时效性较低。

3）最后，数据仓库中的基础数据除了存储基础数据存储的数据外，还存储主数据的身份整合信息和产品信息等内容，目的是支持高级的决策分析。

（2）汇总加工

汇总加工是对基础数据的明细数据进行轻度汇总，通过对常用数据的汇总，可以降低后续 ETL 的复杂性。

（3）库内集市

库内集市可以分成分析类集市和管理类集市。它们都是根据业务需求形成的数据集合。

分析类集市是通过数据挖掘、文本分析、预测分析等手段，帮助企业挖掘有用的信息，以提高企业决策分析的能力。

管理类集市是指为了企业管理的需求而进行的数据分析，可以包括管理驾驶舱、固定的报表、OLAP 多维分析等内容。

对于数据仓库质量的管理，可以包含以下几种方式：

1）采用抽样统计分析的方法监测数据仓库的质量。

通过抽样的统计分析方法来提高数据的加载效率和快速发现数据的错误。首先判定该批次数据的质量等级，然后根据不同的质量等级，采用不同级别的校验规则。对于质量等级非常好的一批数据，可以采用较为宽松的校验规则对每条记录逐条检查。反之，则采用较为严格的校验规则逐条检查。这种方式可以大大提高数据的加载入库效率和数

据质量检查效率。

换句话说，保证数据的质量检查尽量在入库前完成。如果发现入库后的数据质量有问题，那么可以采用异议处理或者其他方式进行改进。如果在入库前发现系统级别的错误，则将错误结果反馈给源系统，如图 3-19 所示。

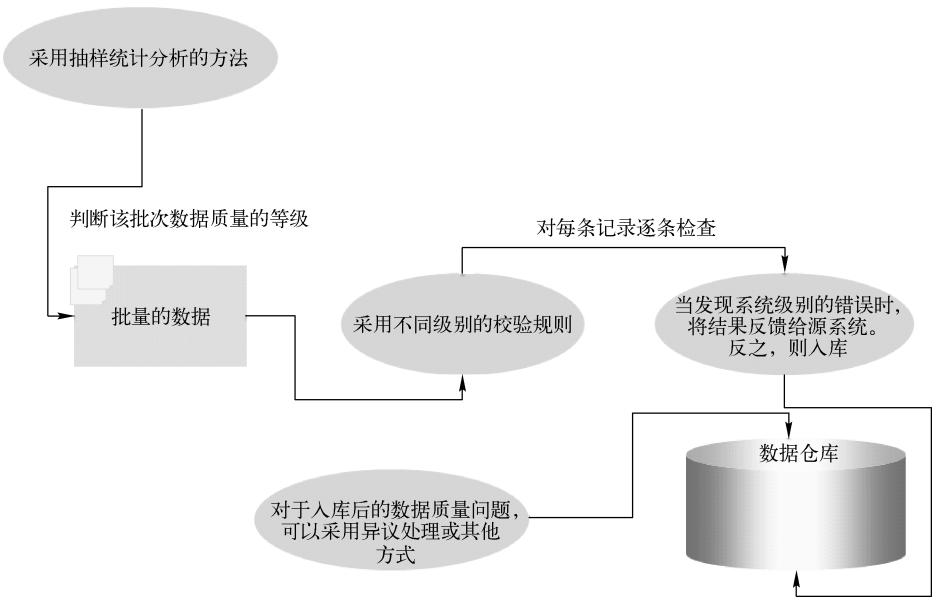


图 3-19　抽样统计分析的方法

2）对于数据仓库质量，可以采用格式校验、逻辑校验的方式和两端数据对比的方式进行验证。如果发现两端数据不一致，则将结果反馈给源系统进行核查，如图 3-20 所示。

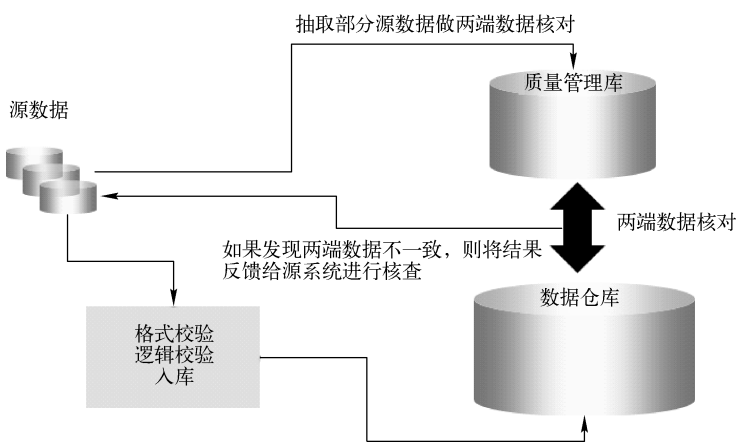


图 3-20　数据仓库质量验证

3）在数据仓库部署业务检查规则和技术检查规则，周期性地对数据仓库质量进行检查，并且将检查结果提交给质量管理平台，由质量管理平台对提交的检查结果进行识别和分析，最后再提交给源系统去治理和改进，如图 3-21 所示。

需要理解的是，数据仓库不仅仅是技术，它更是一个管理课题。从内部管理上来说，它

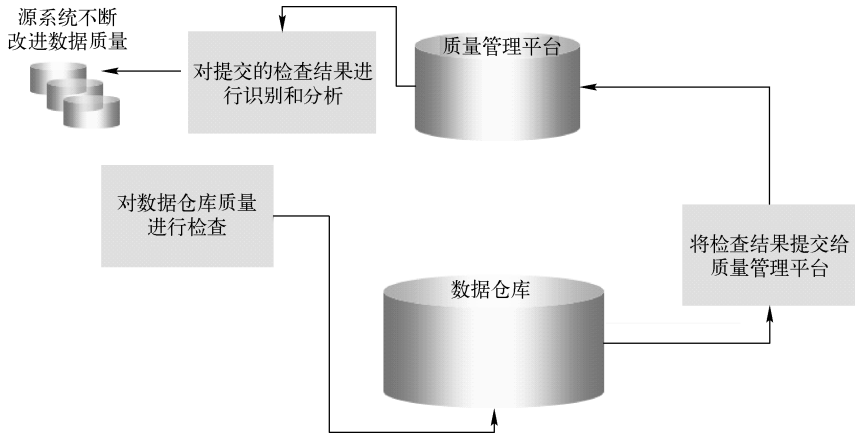


图 3-21　数据仓库部署检查规则

可以真实地反映企业的经营状况和未来的发展趋势，为领导决策和经营管理提供准确和全面的分析。同时可以利用数据挖掘，更好地为客户服务。有关数据仓库的介绍，将在第 9 章详细介绍。

6. 应用

应用可以包含各种查询类应用、分析类应用和管理类应用。它们的数据源来自于数据加工区的数据，同时可以将数据查询记录返回给数据仓库作为分析数据使用。

3.3.2　目标数据架构的分布和流转

下面将从数据分类的角度，分析数据在未来数据架构各个逻辑库上的分布及流转。

对于逻辑库的设计原则，可以包含以下几个方面，如图 3-22 所示。

图 3-22　逻辑库的设计原则

（1）数据的共享性

减少数据复制并降低数据的冗余度，提高数据的共享性。

（2）数据的管理性

考虑系统对于数据管理方面的要求，特别是数据质量的管理。

（3）数据的高性能

基于性能的考虑，可以将加工和查询分开。

（4）数据的可用性

确保系统对外服务的时间窗口尽可能延长，减少停机的时间。

对于数据架构的分布和流转，需要先了解逻辑库包含哪些内容，如图 3-23 所示。

- ODS

主要存储贴数据源的最近一期的数据。

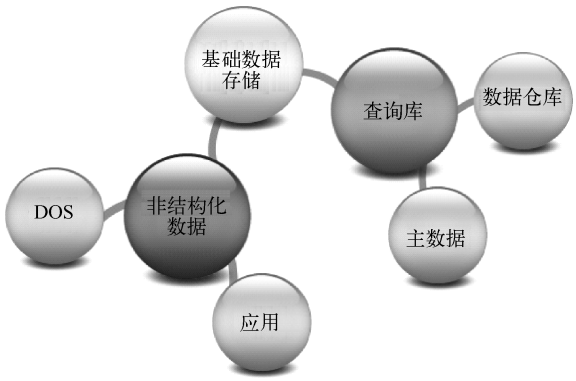


图 3-23　逻辑库相关内容

- 基础数据存储

主要存储校验过的明细的基础数据，存储的期限根据业务需求制定。

- 非结构化数据

主要存储互联网或者其他渠道获得的经过处理的非结构化数据。

- 查询库

主要进行数据加工或者产品加工，保存过程数据。

- 数据仓库

主要保存基础的历史数据，或者主数据、产品的信息，供后续加工和使用。

- 主数据

主要存储核心业务实体和实体之间关系的数据，如唯一身份识别信息。

- 应用

存储复制的数据并提供对外服务。

一、数据架构的分布

数据分布主要分析业务数据在多个系统之间和多个环节之间的分布情况。下面主要分析业务数据在各个逻辑库之间的分布状况，举例见表 3-2。

表 3-2　业务数据在各个逻辑库之间的分布状况

逻　辑　库	业 务 数 据
ODS	个人基本信息、企业信息、交易信息、财务信息等基础信息
基础数据存储	个人基本信息、企业信息、交易信息、财务信息等基础信息
非结构化数据	互联网信息
查询库	查询服务类的信息
数据仓库	个人基本信息、企业信息、交易信息、财务信息等基础信息，以及主数据信息、查询服务类信息等内容
主数据	个人身份信息、企业身份信息等内容
应用	查询服务类的信息

二、数据架构的流转规划

数据流转是描述业务分类在各个逻辑库之间的流转情况，如图 3-24 所示。

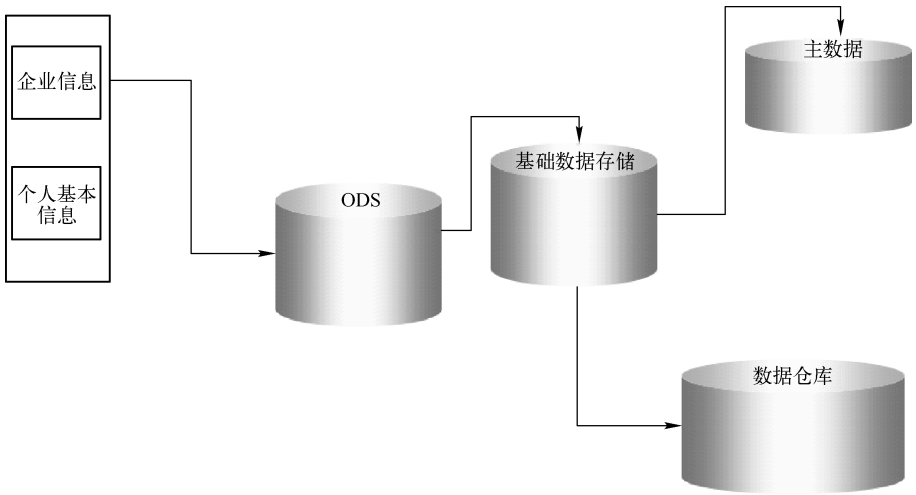


图 3-24　数据流转

首先，企业信息和个人基本信息在 ODS 中临时存储并且进行校验，当校验通过后存放到基础数据存储中，然后，将这些信息加载到主数据中进行企业身份信息整合和个人身份信息整合，最后，将个人基本信息和企业信息加载到数据仓库中。

合理的数据分布和流转可以提高数据的一致性，减少数据冗余，从而提高数据的灵活性和可扩展性。

首先，核心的数据尽量不要反复地分布在不同的数据库中，这样可以降低数据不一致性的风险，但是有时候基于系统性能的考虑，有些合理的冗余是可以存在的。

其次，在数据分布中需要建设一个唯一可信的数据源，这样保证在后续的加工过程中有依据可查，同时提高了数据的一致性。

再次，尽量缩短数据加工链条。例如，身份信息在主数据中加工，然后对应用和数据仓库供数，基础数据存储为数据仓库、主数据和查询库提供增量数据，这几条链路单独加工，并行处理，提高了效率。

三、数据归档

数据归档是指定期将基础数据存储、应用的数据进行归档保存，它的目的是为了保存原始数据。原则上数据归档对中间数据或者临时数据不进行归档操作。

数据归档可以帮助数据再次核对和备查。数据归档包括在线存储、近线存储和离线存储，如图 3-25 所示。

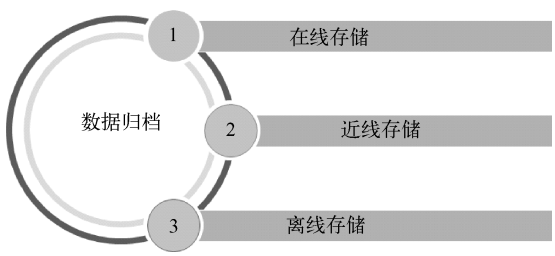


图 3-25　数据归档

（1）在线存储

在线存储主要保存近期业务数据，对在线存储的访问频率相对较高。可以使用高速磁盘对数据进行保存。

（2）近线存储

近线存储主要保存访问频率相对较低的数据，一般使用低速磁盘进行存储。

（3）离线存储

离线存储主要保存数据访问频率低，很少存在加工需求的数据，可以使用光盘，磁带等价格低廉的介质保存。

3.3.3 对数据架构的验证和总结

一、总体数据流转方案验证

首先，数据通过数据交换层进入到 ODS 中的缓冲区，缓冲区是贴数据源的。缓冲区的数据与加载区的数据进行关联逻辑校验，校验通过后再替换掉加载区的数据。

然后基于实时批量的方式，将校验通过的加载区的数据统一存储在基础数据存储中。

最后基于实时批量的方式将基础数据存储的数据导出成增量文件，为后续加工供数。

数据流转方案验证如图 3-26 所示。

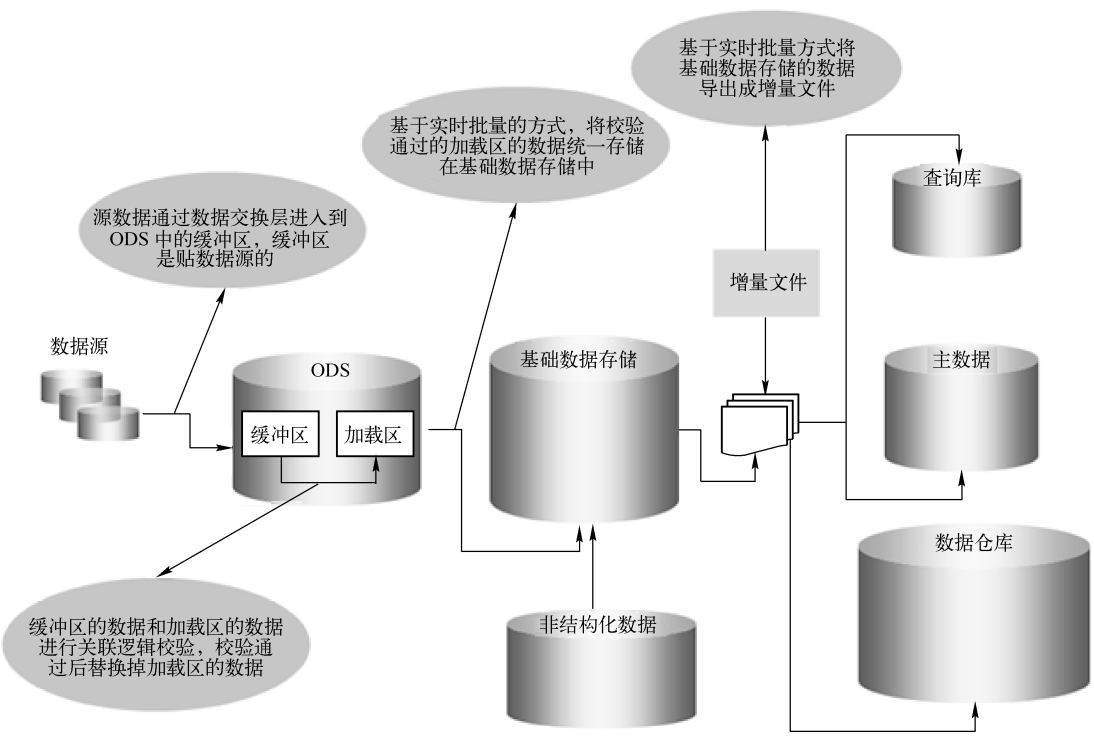


图 3-26　数据流转方案验证

二、产品加工方案场景验证

简单查询类产品在查询库中加工，然后按日统一将加工后的数据复制到应用中，统一对外提供查询服务。

对于挖掘分析类的需求，应该在数据仓库中加工，有时为了性能考虑，可以将数据仓库中的数据迁移到库外集市加工。

如图 3-27 所示，一些基础查询类的产品在查询库中加工获取，一些身份加工整合的数据从主数据中获取，然后通过查询类应用统一对外提供服务。

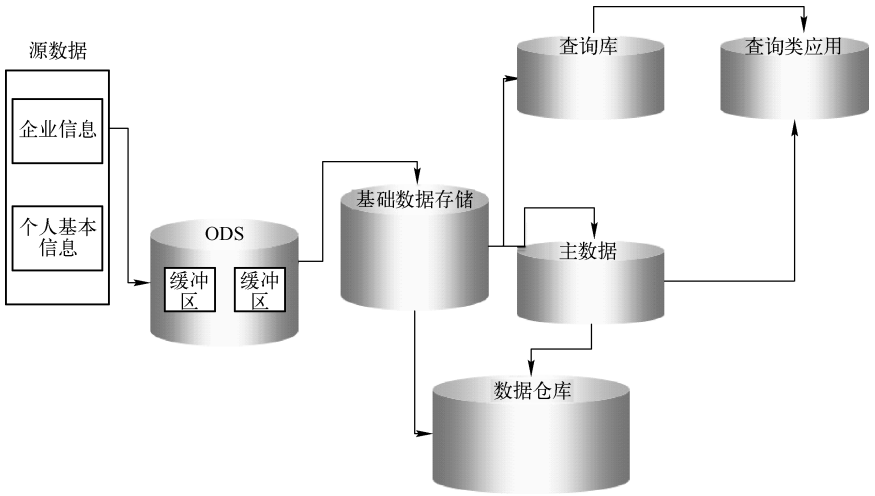


图 3-27　产品加工方案场景验证

三、数据删除场景验证

首先，对应的删除数据通过数据交换层进入到 ODS 的缓冲区，如果要删除的数据仅仅包含历史数据，加载区的数据不需要删除。如果删除的数据包含最近上一期的数据，则需要删除加载区的数据。

然后，删除基础数据存储中对应的数据，同时为了逻辑校验，加载区中最近一期的数据被删掉后，需要把基础数据存储中最近一期的数据回写到加载区。

最后，把查询库和数据仓库中对应的数据删除。当数据删除后再重新加工，如图 3-28 所示。

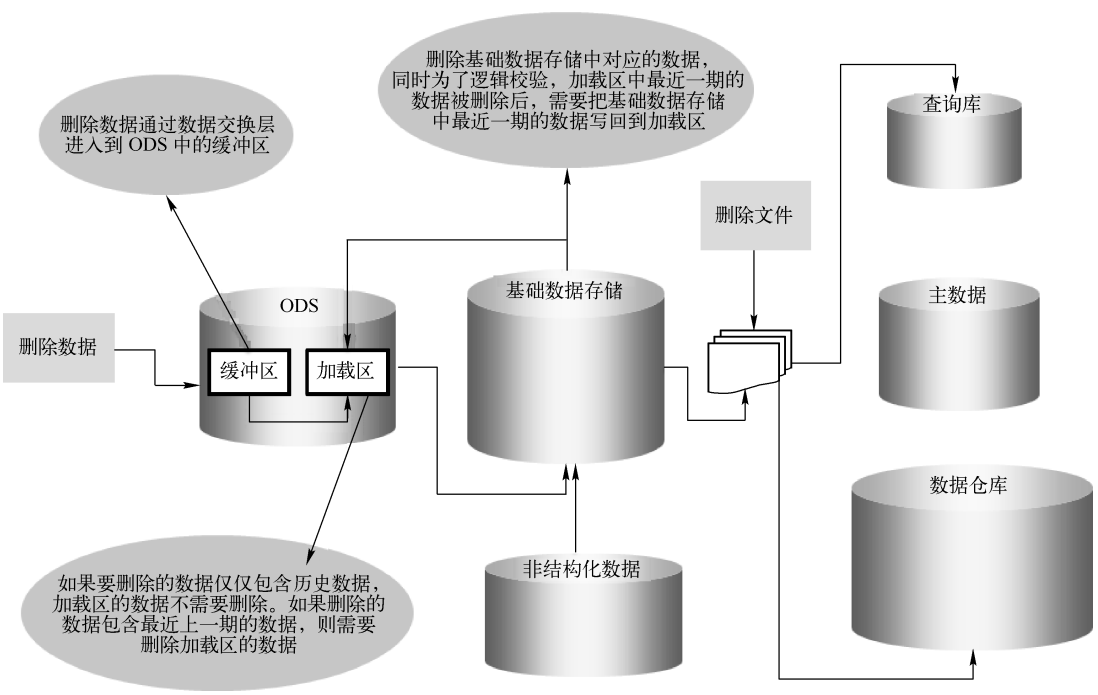


图 3-28　数据删除场景验证

基于调度机制确保数据一致性：

1）到达当天截至的时间点，例如凌晨 12 点，加载完当天上传的所有源数据，如图 3-29 所示。

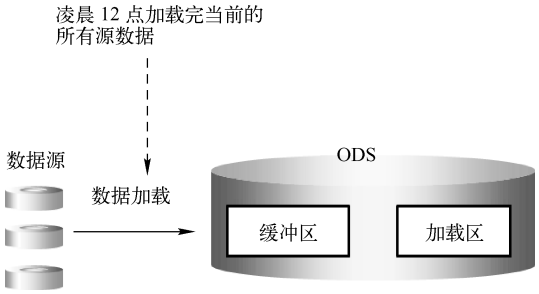


图 3-29　加载当天数据

2）当最后一个加载任务完成之后，再增加最后一个传输任务，因为传输的是最后一个新增数据，所以花费的时间不会太多，如图 3-30 所示。

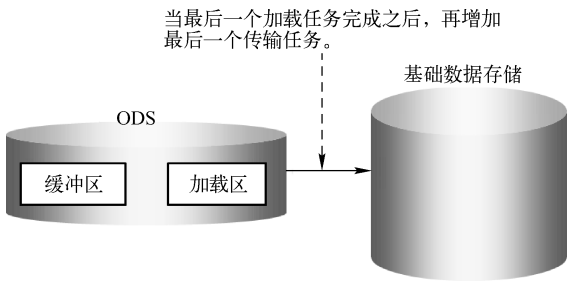


图 3-30　增加传输任务

3）查询库数据、主数据、数据仓库的数据由于加工节奏不一样，因此数据可能存在不一致的情况。在数据加工过程中，因为加工流水线的顺序执行原因，在某一个时刻点，不同库之间数据可能不一致，需要分析业务是否能够接收数据的不一致性，如图 3-31 所示。

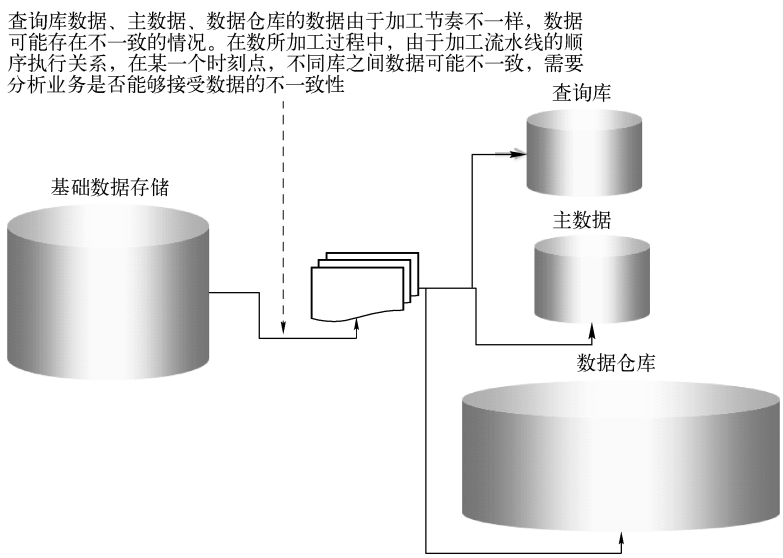


图 3-31　基于调度机制确保数据一致性

数据架构在效率、灵活性以及扩展性方面可以满足业务的需求，因为加载区仅仅存储最近上一期的历史数据，目的是为了支持快速加载和校验。

加载区通过导出文件的方式，同时对查询库、主数据和数据仓库供数。基础查询类产品在查询库中单独加工，分析类产品和对实时性要求不高的产品在数据仓库中加工。

对于基础数据存储来说，它是唯一可信的数据源，对于查询类产品的加工，它直接从基础数据存储中增量获取。对于数据仓库来说，它按主题存储基础数据，用于实时性要求不高的统计分析或者挖掘分析。

小结

- 数据架构理论体系是把业务和技术融合到一起的一套体系。它包括技术、方法和相应的管理过程。经过几十年的发展，数据架构已经形成了完整的理论体系。
- 数据架构是企业架构的重要组成部分，实现数据的合理组织和共享，保证数据在系统之间的一致性、完整性、安全性和正确性。
- 数据架构规划中需要保证数据的安全性、可用性、完整性、真实性和抗抵赖性。
- 数据架构的指导原则包括灵活性原则、高效性原则、可扩展性原则、数据共享原则、数据可用性原则、数据安全性原则。
- 一般来说，数据架构包含数据模型和分类、数据分布和流转等内容。对于数据治理来说，它是为了提升数据架构各个层次的管控和协作能力。同时数据架构为数据治理提供基础能力支撑，因此，数据治理与数据架构可以说是相辅相成的。
- 对于数据架构来说，我们可以从几个方面去了解现状存在的问题是什么。例如，判断数据架构的原则是否清晰、架构层次的划分是否合理等内容。
- 对采集的数据项进行分析，判断是否能满足对产品的加工需求，效率问题是否存在改善的空间，是否能够支持数据的快速入库，不同系统之间的数据是否可以共享，是否可以规划数据交换平台，提高数据加工的效率，保证数据架构满足灵活性、高效性和可扩展性。
- 数据模型是指用实体、属性及其关系对企业运营和管理过程中涉及的业务概念和逻辑规则进行统一定义、命名和编码。
- 数据模型是对数据特征的抽象，它一般分为概念模型、逻辑模型和物理模型。概念模型是以数据分类的形式体现，而逻辑模型以 ER 图的形式体现。
- 概念模型是从业务的角度对数据进行抽象，包括业务层面上主题域的划分，以及各个主题域下的数据分类和基于分类的非功能属性。
- 数据分类是根据业务特征对数据进行归类和划分，并用层级列表的方式展示数据内容。数据分类的规范需要满足各种业务需求对数据组织的要求。
- 数据分类是概念模型的体现。
- 数据分类的目标是可以促进业务人员和技术人员之间的沟通，指导技术人员对数据格式的制定，指导数据的分布和流转。
- 逻辑数据模型是用来发现、记录和沟通业务的详细“蓝图”，由一系列表和实体详细描述组成，是通用的业务语言，便于业务与业务之间的功能理解，遵循第三范式，包

括主题域的设计、基本实体的设计和主要属性的设计，是 IT 人员和业务人员沟通的工具和桥梁。

- 物理模型是对逻辑模型针对具体实现环境的物理化，可以不遵循第三范式，主要包括实体属性的物理化，属性的长度、类型、主键、外键、索引等详细设计。
- 针对主题域下数据分类，需要从变动频率、变动量、变动模式、数据量大小、格式、共享性等各个维度进行分析。数据分类的非功能属性对于数据分布设计具有重要的参考意义。
- 对于未来数据架构可以参考以下的思想：首先强调数据的存储与流转，支持层次化的处理，包括对结构化数据与非结构化数据的处理能力。
- 数据分布：数据分布主要包括业务分布与系统分布。数据分布主要分析数据在各个环节中的创建、引用、更新和删除，并根据业务对数据的处理特点，合理规划数据的分布。
- 数据架构不包含数据治理方面的内容，但是数据架构为数据治理提供基础能力支撑，而数据治理的目的是提升数据架构各个层次的管控及其协作能力。
- 数据架构的改进方向：首先应该明确数据架构总体指导原则是什么，以此原则指导未来数据架构。明确数据架构的各个层级，对每个层级进行数据治理。
- 数据流转是描述业务分类在各个逻辑库之间的流转情况。
- 数据归档是定期将基础数据存储、应用的数据进行归档保存，它的目的是为了保存原始数据。原则上数据归档对中间数据或者临时数据不进行归档操作。数据归档可以帮助数据再次核对和备查。数据归档包括在线存储、近线存储和离线存储。
- 数据架构在效率、灵活性以及扩展性方面可以满足业务的需求，因为加载区仅仅存储最近上一期的历史数据，目的是为了支持快速加载和校验。加载区通过导出文件的方式，同时对查询库、主数据和数据仓库同时供数。基础查询类产品在查询库中单独加工，分析类产品和对实时性要求不高的产品在数据仓库中加工。对于基础数据存储来说，它是唯一可信的数据源，对于查询类产品的加工，它直接从基础数据存储中增量获取。对于数据仓库来说，它按主题存储基础数据，用于实时性要求不高的统计分析或者挖掘分析。

第 4 章　数据架构案例

本章目标

通过前一章的学习，我们已经理解了数据架构的工作方法和指导原则，包括概念模型、逻辑模型、物理模型的建设，数据分类的规划，未来数据架构的分布和流转的建设，对数据架构的验证等内容。

本章在前一章的基础上，重点介绍项目总体规划的几个阶段、系统项目建设过程中可能面临的风险和对策、某金融行业数据架构的相关案例。包括数据架构的分布、流转、加工的处理时序、数据纠错方案介绍、数据架构的优化和数据架构实施规划等内容。

学习本章后，读者将掌握：

- 数据架构在项目阶段规划中的地位
- 项目总体规划的几个阶段
- 系统建设策略
- 项目阶段建设计划
- 系统项目建设过程中可能面临的风险和对策
- 任务分析规划
- 某金融行业数据架构的分布规划
- 某金融行业数据架构的流转规划
- 数据架构的纠错更正需求
- 数据加工处理时序规划
- 数据架构在线纠错更正方案设计
- 在线纠错更正的指导原则
- 非功能性需求
- 某金融行业数据架构优化
- 某金融行业数据架构案例描述
- 主数据规划
- 数据仓库规划
- 数据交换平台规划
- 产品加工流程概述
- 数据架构实施规划
- 系统切换规划案例

4.1　某金融行业数据架构的前期规划

4.1.1　理解数据架构在项目规划中的地位

数据架构在项目规划中占有非常重要的地位。

项目阶段分成以下几个部分：项目启动阶段，现状评估、高阶需求分析阶段、架构设计和规划阶段以及实施规划和运维阶段，如图 4-1 所示。其中现状评估和高阶需求分析阶段主要是理解企业发展战略和业务需求，对系统现状评估和高阶需求进行分析。在架构设计和规划阶段，主要包含应用架构、数据架构、技术架构和 IT 治理等内容。最后一个阶段就是实施规划和运维阶段。

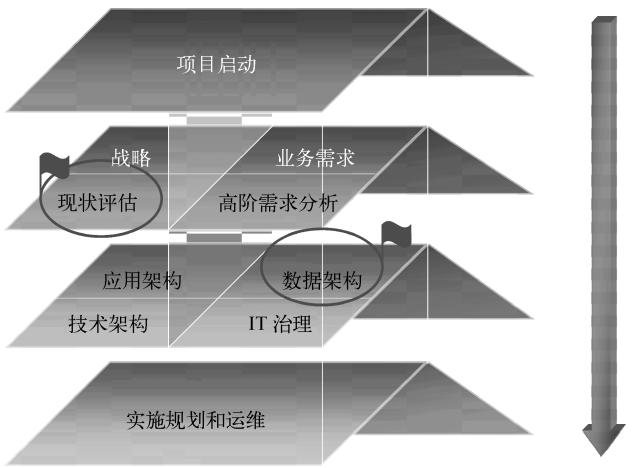


图 4-1　数据架构项目阶段划分

4.1.2　项目总体规划的几个阶段

在系统总体规划过程中，离不开下面 3 个阶段：现状分析和需求分析阶段、总体规划设计阶段和总体架构实施规划阶段，如图 4-2 所示。

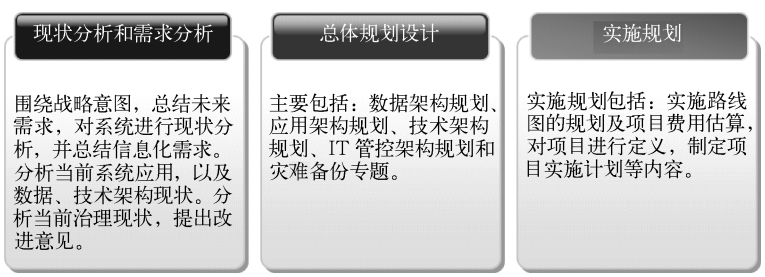


图 4-2　系统总体规划

其中现状分析和需求分析阶段主要是围绕战略意图，总结未来需求，对系统进行现状分析，并总结信息化需求。分析当前系统应用，以及数据和技术架构现状，提出改进建议。分析当前架构治理现状，提出改进建议。

总体规划设计主要包括数据架构规划、应用架构规划、技术架构规划、IT 管控架构规划和灾难备份专题。

实施规划包括实施路线图的规划及项目费用估算，对项目进行定义，制定项目实施计划等内容。

4.1.3　系统建设策略

系统建设策略主要包含以下两种方式：统一开发、统一推广和快速建设方式。

1. 统一开发、统一推广

这种建设策略是先建设系统的全部内容，然后再逐步推广。优点是阶段划分清晰，管理难度较小，但缺点是周期长，前期推广内容多，对业务变化的适应能力较弱。

2. 快速建设方式

这种建设策略是在统一规划的基础上，尽早完成基础平台的建设，然后按照业务重点需求，快速开发核心的系统，再逐步推广应用，最后按照优先级别的高低，完成系统的建设和优化，如图 4–3 所示。整个系统将涵盖所有的业务需求。这种方式的优点是可以迅速抓住重点，能够快速见效。可以把项目建设分成多个子项目，将核心系统的推广和其他子项目的建设结合起来，缩短时间周期，节约开发成本。

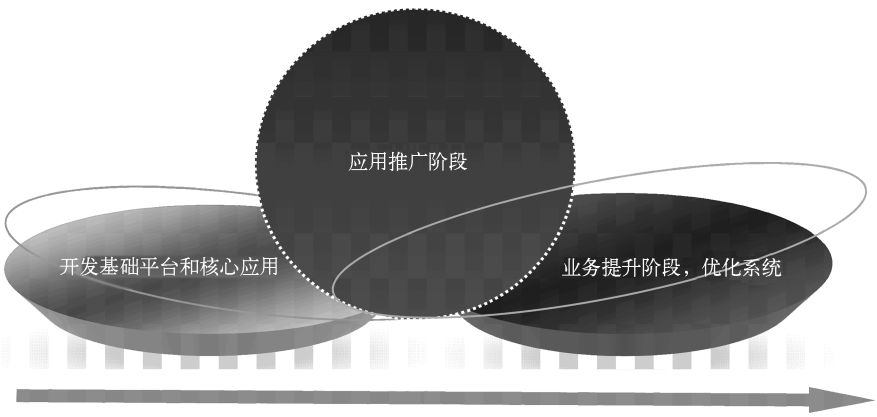


图 4–3　快速建设方式

对于快速建设方式的解读，可以分别从如何开发基础平台和核心应用、如何完成应用的推广、业务提升和优化系统等几个方面进行，如图 4–4 所示。

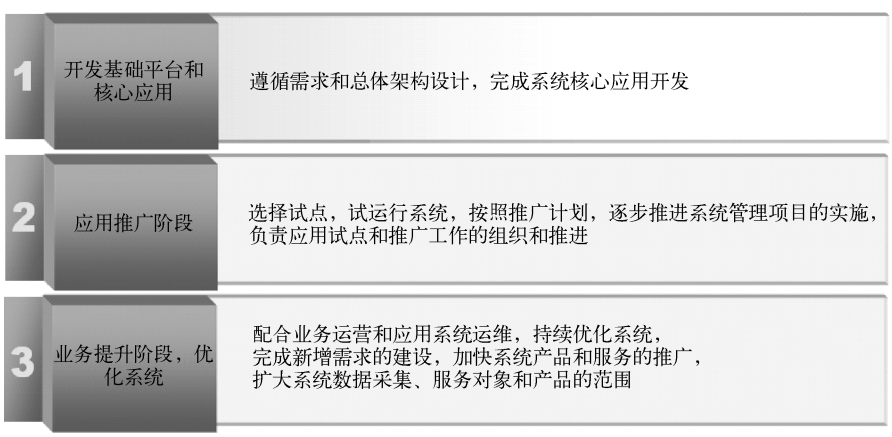


图 4–4　快速建设方式的解读

4.1.4　项目阶段建设计划

项目阶段的建设计划主要包含以下几个方面：项目启动、需求分析、系统设计、开发和

测试以及项目验收等内容，如图 4-5 所示。

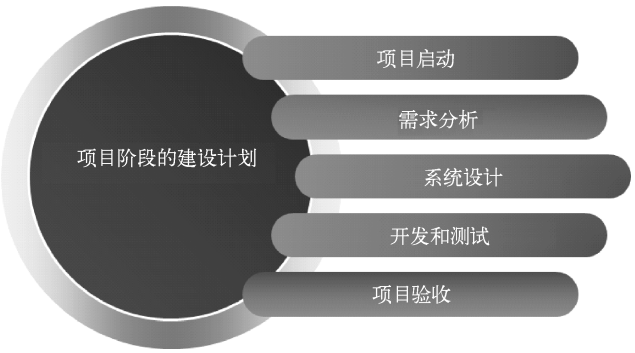


图 4-5　项目阶段的建设计划

其中项目启动包括制定项目计划、项目章程和制度等准备工作。

需求分析包括需求调研、原型开发、需求分析等工作。

系统设计包括架构设计、功能设计、数据库设计等工作。

开发和测试主要包括功能开发、系统对外接口开发、单元测试、功能测试、性能测试、用户测试集成测试等内容。

最后是项目的验收。

如图 4-6 所示，针对项目建设计划，可以对基础设施、容灾系统、产品加工、对外服务、数据加工和数据采集进行建设。

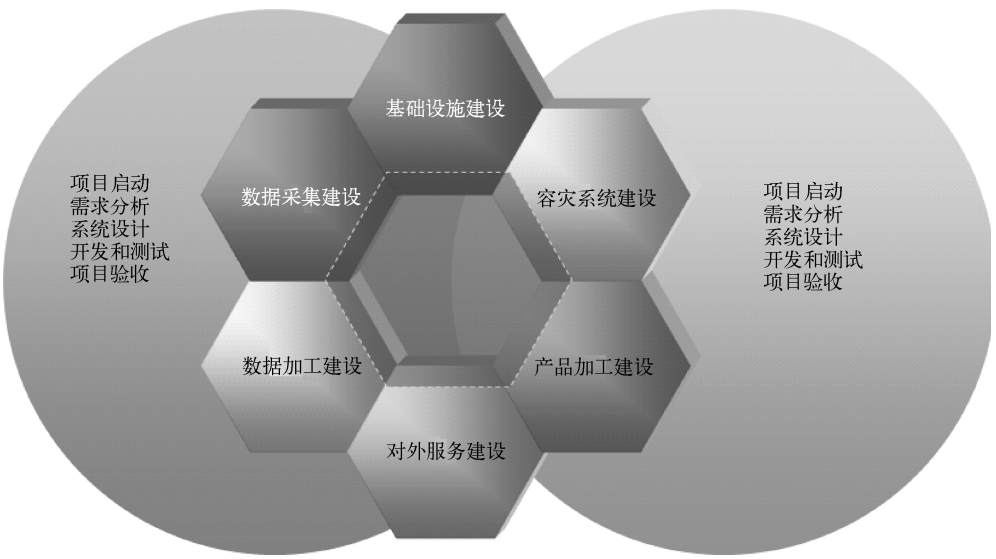


图 4-6　项目建设计划的主要内容

4.1.5　预算及风险效益分析

1. 预算

预算主要包含两个方面的内容：一是对硬件、软件平台、应用软件和各种服务的投资和

维护的费用估算，二是对人工服务费用的估算，如图 4-7 所示。

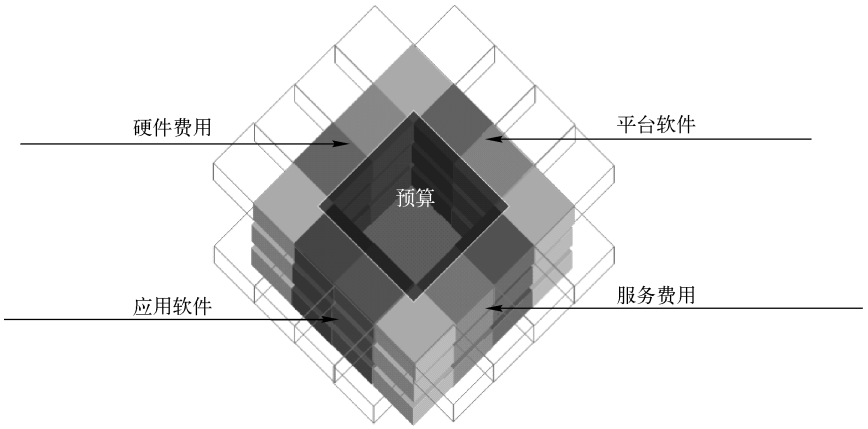


图 4-7　项目预算

（1）硬件费用

硬件费用主要包括各种服务器、存储、网络等配套设施的费用。

（2）平台软件

平台软件费用主要包括各种操作系统、数据库、中间件等基础平台软件的费用。

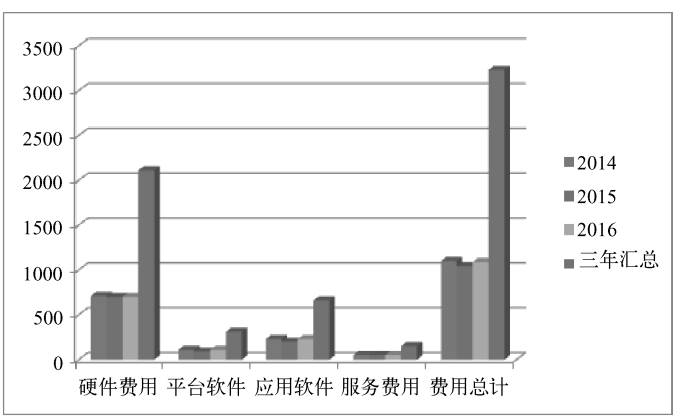
（3）应用软件

应用软件费用主要包括各种专业应用系统的实施费用，如人力资源管理系统、数据仓库、财务管理系统、IT 审计和日志管理平台、IT 运维管理平台、CRM 系统的建设和实施。

（4）服务费用

服务费用主要包括项目管理、系统架构设计、编码的费用。例如，某商业银行预算如图 4-8 所示。

单位：万元



	2014	2015	2016	三年汇总
硬件费用	710	700	700	2110
平台软件	110	90	110	310
应用软件	230	200	230	660
服务费用	50	50	50	150
费用总计	1100	1040	1090	3230

图 4-8　服务费用举例

2. 风险效益分析

在系统建设过程中，面临的风险包括：组织风险、业务变革风险、技术风险和项目管理风险，如图 4–9 所示。

图 4–9　系统建设过程中可能面临的风险

● 组织风险

可能出现的风险包括企业未来组织的不确定性，它会影响业务的流程和范围，从而影响系统的建设。应对这种风险的策略是明确业务策略和发展方向，合理规划组织机构。

● 业务变革风险

可能出现的风险是业务流程的调整有可能影响岗位职责的变化。应对这种风险的策略包括业务变革得到企业高层的支持，提前做好应对的准备。

● 技术风险

可能出现的风险是随着技术不断创新和发展，对技术的选择会带来相应的风险，从而造成技术先进性和成熟性难以平衡。为了避免技术风险，我们应该选择成熟度较高的产品。

● 项目管理风险

可能出现的风险是没有清晰的管理机制和组织，造成职责不清和进度延缓。应对这种风险的策略是采用成熟的项目管理办法。

建设项目效益分析主要是提高核心业务能力和流程执行效率，建立满足需求的系统架构体系等内容，如图 4–10 所示。主要表现在以下几个方面：

1）提高核心业务能力是为了满足业务需求，建立高效、灵活和可扩展的系统，提升产品加工能力和对外服务能力。

2）提高流程的执行效率是为了实现核心业务规范化管理和服务，通过对关键业务点的提示和控制，提升业务效率，防范各种风险。

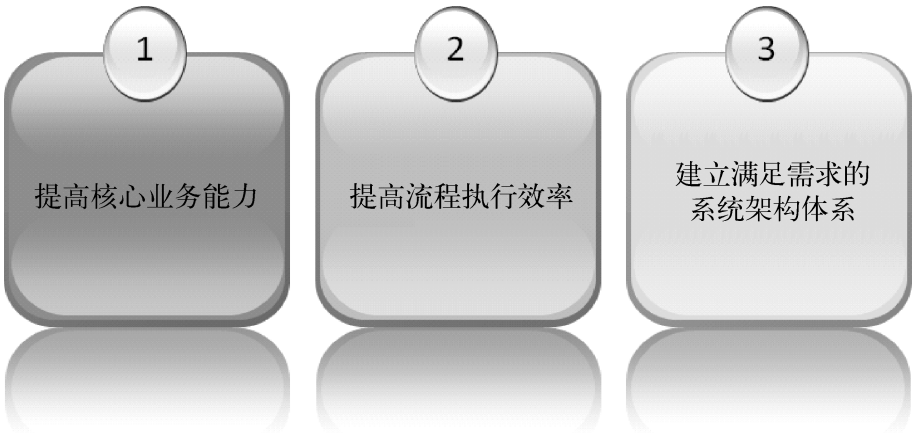


图 4-10　项目效益分析

3）建立满足需求的系统架构体系。例如，应用架构应该满足业务前瞻性和可落地要求，实现应用企业化。数据架构满足灵活、高效、可扩展、数据共享和数据安全等架构要求。技术架构采用成熟技术，符合信息化相关规范，保证系统平稳过渡，并复用现有资产。

4.1.6　任务分析

针对某金融行业信息化建设，可以分成以下几个任务，如图 4-11 所示。

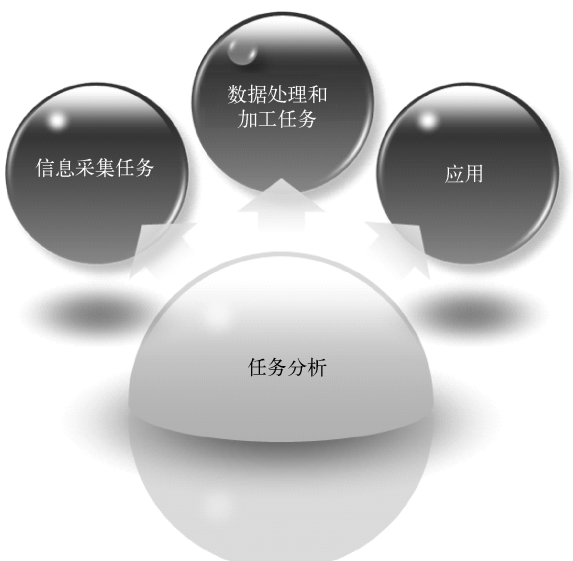


图 4-11　某金融行业信息化建设的任务分析

（1）信息采集任务

信息采集任务主要是充实采集内容，优化采集方式，根据业务需求，动态地增加采集信息。例如，在个人欠款信息中增加欠款发生的日期。同时需要扩大对公共信息的采集，包括各种的税务信息、司法信息和电信信息等。

（2）数据处理和加工任务

数据处理和加工任务是建立数据处理和快速加工响应机制，能够将各种新业务快速纳入

到系统中，提高数据的自动化处理能力和快速加载能力。例如，可以将客户的信用评分能力、身份验证、关联查询、风险预警和各种的数据统计功能快速接入到系统中。

（3）应用

应用任务是建立多样化的产品交付方式，如离线交付、专网交付等，尽量做到 7 × 24 对外服务。

随着大数据时代的到来，数据应用可以产生更大的机遇和挑战。只有更好地利用数据，才能在未来的竞争中获得更大的优势。一般来说，数据的应用主要包括报表功能、统计分析和数据挖掘三种方式，如图 4–12 所示。

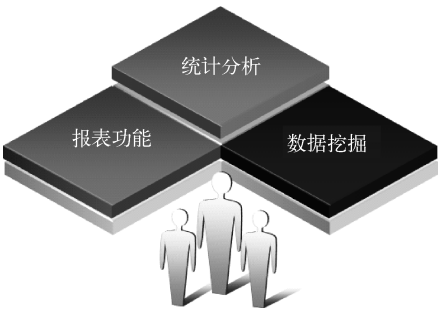


图 4–12　数据的应用

（1）报表功能

报表功能是数据应用的基础，也是最传统和常见的数据应用。报表是决策分析的基础。报表功能的完善、灵活程度能够影响工作的效率。

（2）统计分析

统计分析功能是常见的数据应用方式。随着统计分析工具的推广，统计分析在很多行业中得到了越来越广泛的应用。例如，通过假设检验或者方差分析帮助分析经济运行的规律。

（3）数据挖掘

数据挖掘是数据统计分析的进一步发展，是对数据的深度应用。

数据挖掘起源于 20 世纪 70 年代，但在最近 10 年内得到了广泛的应用和发展，特别是在金融行业、电信行业、互联网行业等。

数据挖掘的目的是为了发现数据背后隐藏的规律，它可以通过使用模型来表达复杂的事物和现象。例如，通过使用回归分析、聚类分析和分类分析等数据挖掘手段在银行业中发现事物的本质和规律。

总之，我们可以通过报表功能、统计分析、数据挖掘等技术手段利用数据和使用数据，为决策者提供决策依据和技术支持。

4.2　某金融行业数据架构的分布规划

数据分布主要包括业务分布和系统分布。数据分布可以分析业务和系统之间各个环节的创建、修改和删除关系，同时可以分析应用系统中数据结构和系统各个模块之间的关系。

其中业务对数据的处理主要包括数据的采集、加工和对外服务三种类型的业务处理。因此，在设计数据架构时，根据业务对数据的处理特点，规划设计合理的数据分布，以满足相关业务的需求。

在规划数据分布时，需要考虑合适的技术方案来满足以下需求：

1）明确不同位置之间的数据定位和数据流向。

2）保证对海量数据的快速加载和不同数据库之间数据的快速增量迁移。

3）保证海量数据的快速产品加工。

4）应该适应数据采集的多样化、产品加工的多样化和对外服务配置化等特点。

5）可以适应数据的纠错更新机制。

数据架构框架包含数据采集层、数据加工层和应用服务层，如图 4–13 所示。

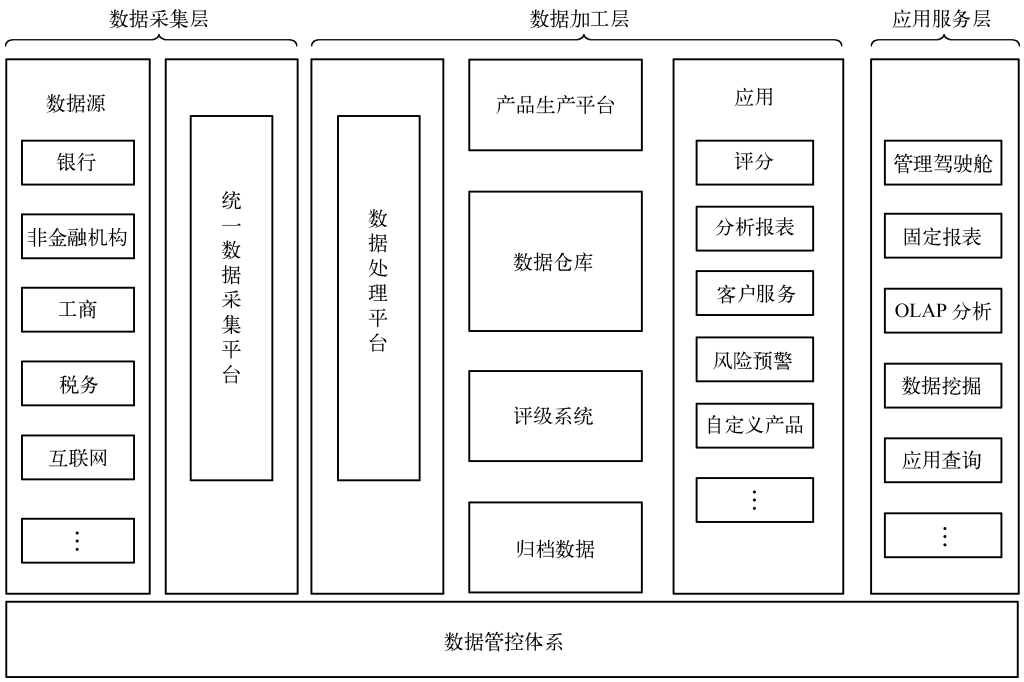


图 4–13　数据架构框架详细描述

下面对数据采集层、数据加工层和应用服务层进行详细描述。

（1）数据采集层

数据采集层主要包括数据源和统一数据采集平台。统一数据采集平台的目的是统一数据采集，包括定期全量、增量的采集。

（2）数据加工层

数据加工层包括数据处理平台、产品生产平台、数据仓库、评级系统、归档数据及应用。

数据处理平台一般是批量、实时地对增量数据或者全量数据进行处理，这种方式可以依赖一些主流的关系型数据库和大型平台来实现。

产品生产平台主要是针对数据类的产品进行生产，一般要求系统可以处理海量数据和复杂的数据，要求高并发和 7 ×24 小时不停机。这种方式可以依赖于大型的平台。

数据仓库以存储历史数据为主，用于对历史数据的分析，支持灵活分析和查询。数据仓库应该有海量数据处理能力、线性扩展能力和高可用性。

评级系统是金融行业的一个应用系统，主要用于对客户的评分。

归档数据是对归档数据的存储，原则上存储历史的原始数据。

应用主要包括评分、分析报表、客户服务、风险预警和一些自定义产品等内容。

（3）应用服务层

应用服务层包括管理驾驶舱、固定报表、OLAP 分析、数据挖掘、应用查询等内容。

综上所示，该数据架构框架基本满足了业务需求。统一数据采集平台从数据源中采集数

据，经过数据处理平台，可以实时、批量地将增量数据或者全量数据分发到产品生产平台、数据仓库、评级系统中，对于一些历史数据也可以放到归档数据中。最后在数据加工层对数据进行加工处理，满足应用的需求。

针对金融行业信息化总体建设的任务需要，可以对数据架构做进一步修改和优化，如图 4–14 所示。

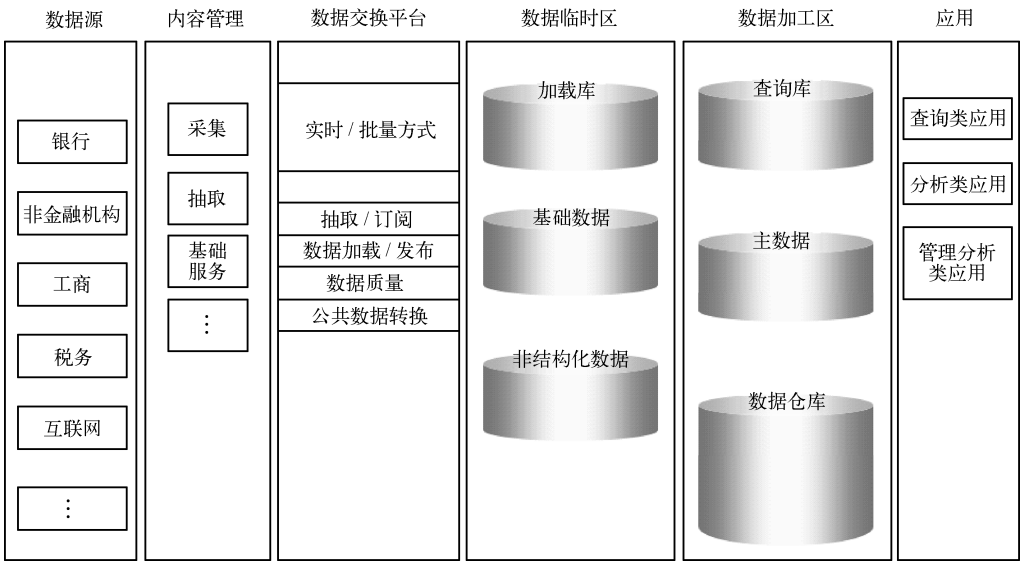


图 4–14　对数据架构的修改和优化

下面对各个层次进行详细说明。

（1）数据源

数据源主要以结构化数据和非结构化数据为主，定义数据采集的来源、内容、格式和采集方式等。

（2）内容管理

内容管理主要为半结构化和非结构化数据提供捕获、管理和存储等方面的服务，也就是非结构化数据的结构化处理。

（3）数据交换平台

数据交换平台主要为外部数据交换和内部数据交换提供支持。

（4）数据仓库

数据仓库是根据业务需求，对历史数据进行整合、轻度汇总和加工，提供分析的功能。

（5）主数据

主数据主要对身份信息进行识别和整合。

（6）加载库

加载库主要提供对源数据进行校验的功能。

（7）基础数据

基础数据主要获取校验通过的数据，作为后续加工的唯一可信数据源。

（8）查询库

查询库主要存储查询类应用的信息。

（9）应用

应用主要提供对外查询服务。

未来数据架构的主要内容包括数据源、内容管理、数据交换、数据仓库和应用，如图 4–15 所示。

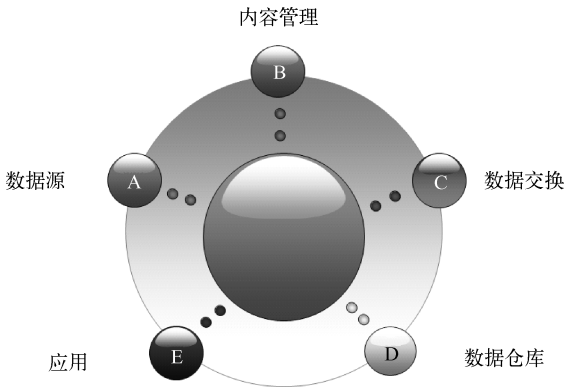


图 4–15　未来数据架构的主要内容

（1）数据源

结合业务特点和数据特征，对源数据层进行规划，同时需要充分考虑灵活性和可扩展性的要求。如图 4–16 所示，数据源层提供需要的源数据，可以描述从哪里、以什么样的方式和渠道加载到系统中。采集数据分为结构化数据和非结构化数据，非结构化数据主要来自互联网，结构化数据主要来自金融机构和公共部门。

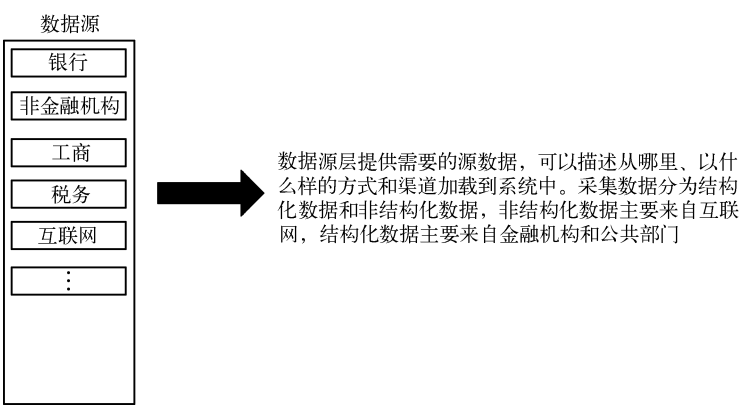


图 4–16　数据源的特点

（2）内容管理

内容管理是指对内部多种格式的信息资源进行组织、分类和管理的过程。内容管理作为一种应用软件，管理和访问各种非结构化数据，包括各种音频、视频、图像等信息。内容管理处理的信息对象比传统的关系型数据库管理系统处理的数据范围更加广泛，包括文字、多媒体、网页、广告和文档等。

内容管理重点解决非结构化数据和半结构化数据的采集和管理问题。然后将这些数据集成到信息系统中。

（3）数据交换

数据交换层满足数据架构各个层次之间的协作要求，承载着外部和内部的数据交换。一般来说，数据交换层包括数据抽取和订阅、质量检查、数据转换和数据加载等几个方面，如图 4–17 所示。

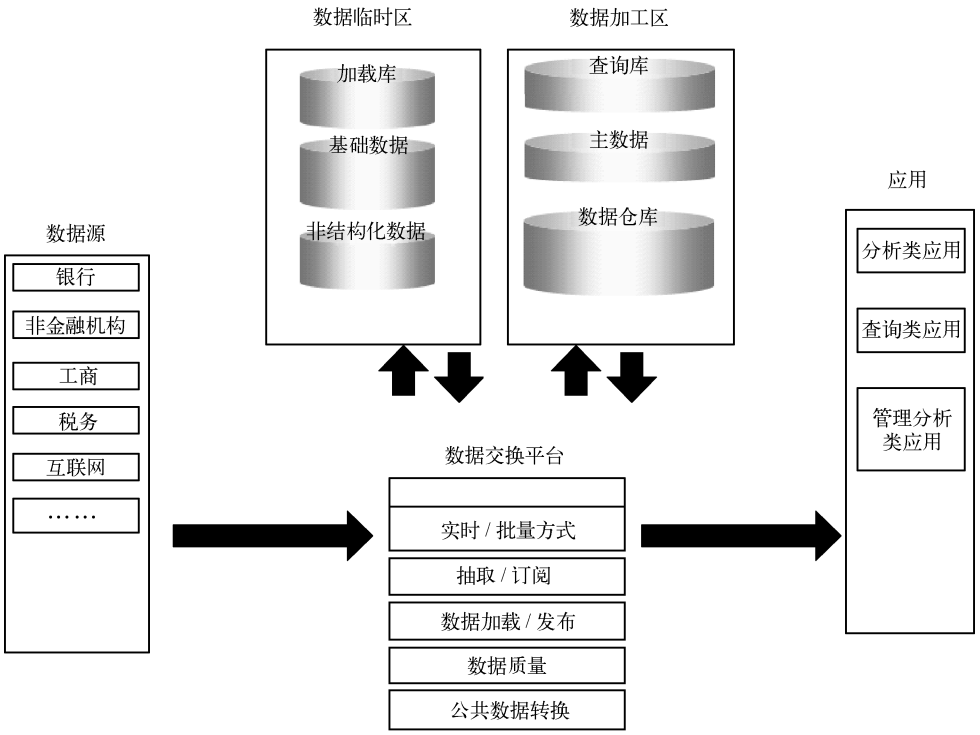


图 4–17　数据交换层

其中抽取/订阅是从数据源层、数据临时区中获取增量或者全量数据，然后分发到各个加工库或者应用库。

数据质量保证数据质量检查、清洗后，数据可以满足基本的质量要求。

公共数据转换是经过数据质量清洗后的数据进行业务和技术规则转换。

数据加载/发布将生成数据文件，然后加载到数据库中。

（4）数据仓库

数据仓库主要提供面向主题的、集成的、随时间变化的，但信息本身相对稳定的数据集合，它主要用于对决策分析的支持。

根据业务要求，在数据架构规划中设置数据加工层，同时在数据加工层中设置数据仓库。数据仓库一般以基础数据整合和汇总数据加工为主。

数据仓库整合全局的信息，包括基础数据层、汇总加工层和集市层。

数据仓库中的数据包含历史信息，记录了从过去某一时间点到目前各个阶段的信息。一般来说，数据仓库的数据不做删除和更新处理。通过这些信息，可以为企业的发展历程和未来趋势做出分析和预测。

数据仓库存储的粒度比较细，存储的历史周期长，可以在基于数据整合的基础上创建各

种应用。

（5）应用

主要存储产品数据，并对外提供查询服务。

4.3 某金融行业数据架构的流转规划

对于数据架构的流转来说，主要目的是降低数据冗余度、提高数据一致性，进而达到灵活、高效的目的。

例如，核心数据不反复分布在不同数据库中，同时允许合理的冗余存在，基础数据中的数据和数据仓库中基础数据层的数据存在冗余，但是在结构和功能上有较大不同。基础数据是作为唯一可信数据源对后续所有应用供数，而数据仓库中基础数据层的数据是为了库内汇总和加工做准备的，如图 4-18 所示。

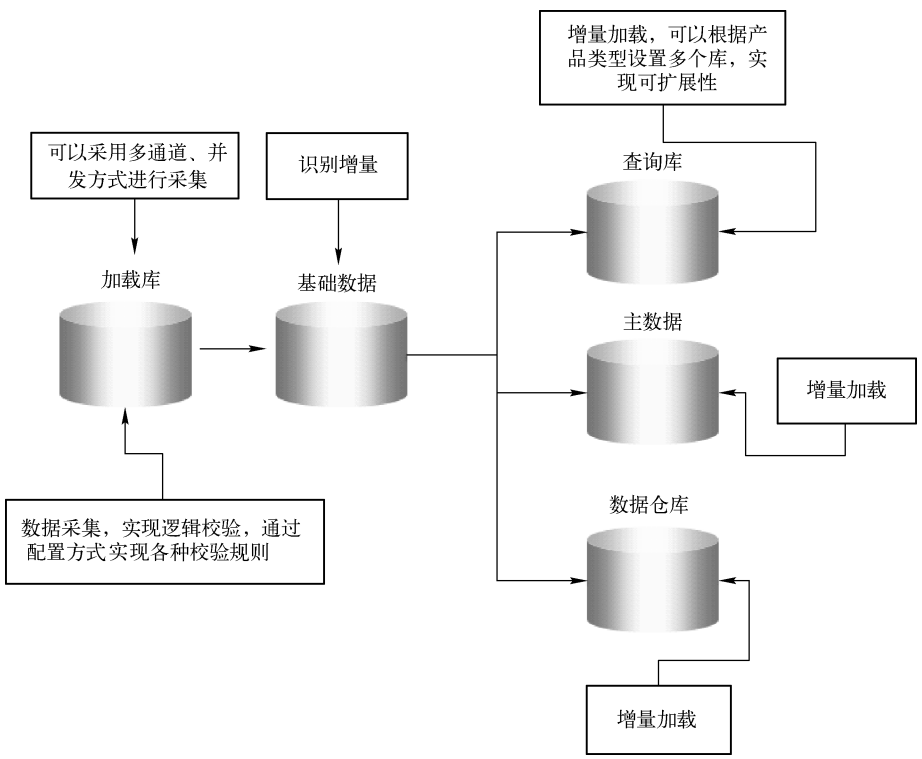


图 4-18　数据架构的流转规划

4.4 某金融行业数据加工处理时序规划

如图 4-19 所示，在数据临时存储区中，数据可以多路并行执行校验和加载，然后在基础数据中进行存储，最后按照某个时间周期往后进行增量数据迁移。

在数据加工区中，例如凌晨 1 点，可以在数据加工区中对前一天的数据进行加工和计算，其中查询库数据加工、主数据加工和数据仓库数据加工可以并行执行，最后在早晨 8 点

左右，加工完成后对外提供服务。当然，我们也可以考虑利用双机备份机制来对外提供不间断服务。

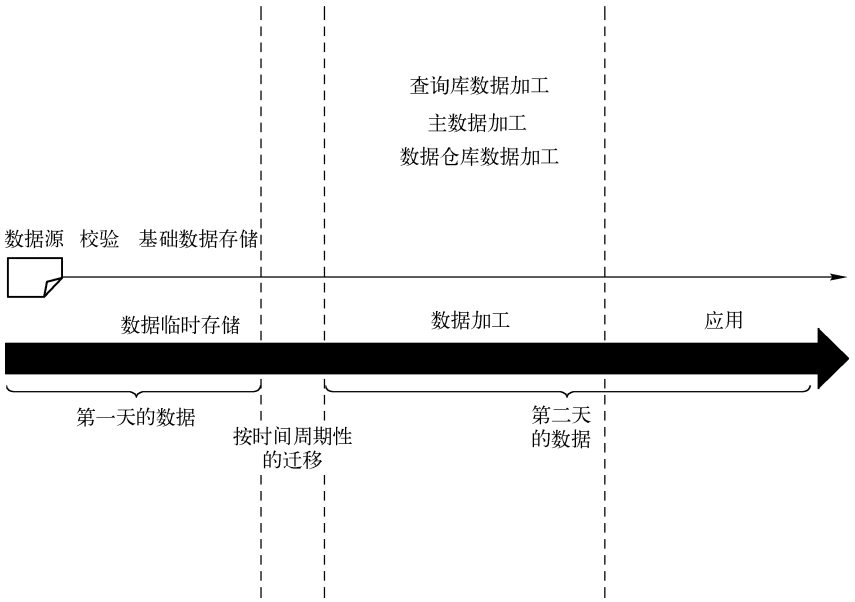


图 4-19　金融行业数据加工处理时序规划

4.5　某金融行业数据架构的纠错更正需求

4.5.1　数据架构纠错更正的功能性需求

某金融行业系统的纠错更正需求主要包括基础数据的数据纠错更正需求、查询库的数据纠错更正需求、主数据的数据纠错更正需求和数据仓库的数据纠错更正需求。下面分别讲述：

1. 基础数据的数据纠错更正需求

基础数据可以作为唯一可信的数据源，在基础数据做的任何修改也都会通过增量的方式同步到数据加工区中进行加工，然后在应用层得到体现，因此，尽量在基础数据中进行纠错更正，这样有利于数据的一致性。但是为了更好地控制数据，应该严格管理数据纠错更正的权限，所有的动作都应该被记录，以备后续查询使用。

2. 查询库的数据纠错更正需求

对于查询库的数据纠错更正需求，一般是发生在客户提出异议申请之后，由系统检查、确认是否是源系统的错误，最后进行数据纠正。

3. 主数据的数据纠错更正需求

主数据包含身份整合信息，针对不同的信息采用不同的整合方式，一般都直接在主数据中修改信息。

4. 数据仓库的数据纠错更正需求

原则上，数据仓库不进行数据纠错更正，如果确实需要修改，应该记录数据修改前后的值，尽可能保证数据的可追溯性和审计的要求，同时保存在线纠错请求的发起人、发起时

间、原因等信息。

4.5.2 非功能性需求

关于数据架构的非功能性需求，主要包括以下几个方面，如图 4-20 所示。

图 4-20 数据架构的非功能性需求

1. 对并发和响应时间的要求

我们需要考虑系统在线纠错更正请求的数量是多少，这种并发量对系统造成的压力是否大，而客户提交请求系统响应的时间应该维持在几秒以内。

2. 数据可追溯性要求

当客户提交在线纠错更正请求后，将更新基础数据库、数据仓库、主数据和查询库中对应的数据。同时记录数据变化的情况，从而确保数据的可追溯性。

3. 权限控制与安全性要求

在线纠错更正属于风险较大的操作，可能会对数据的正确性、一致性和完整性产生影响。因此，需要对在线纠错更正的权限进行严格限制。

4.5.3 在线纠错更正的指导原则

对于系统的在线纠错更正，需要保证数据的一致性和完整性。在线纠错更正的请求应该尽可能发生在基础数据库中，因为当更新完基础数据后，再通过特殊的数据加工迁移到主数据、数据仓库和查询库中。

对于已经加工完成的数据进行在线纠错更正，如果无法通过修改基础数据中的数据来实现在线纠错，只能考虑在加工区中修改数据。对于所有的在线纠错更正相关操作，必须保留痕迹，从而保证数据的可追溯性。

4.5.4 数据查询

当系统客服人员接到客户的异议申请时，首先通过查询库查询相关数据，从而确定客户反映的问题是否存在，然后通过查询结果定位是属于数据源的问题还是数据加工导致的问题。为了避免数据泄密，需要对数据权限进行严格控制。

当客服部门收到异议处理请求时，需要通过查询相关数据确定是数据加工问题还是数据源的问题。当客服部门或者相应机构提交数据纠错更正请求时，如果提交的数据通过审核，那么系统将会更新对应的基础数据存储的数据，同时进行数据加工和迁移任务的操作。

4.6 某金融行业数据架构优化

某金融行业数据架构的优化主要包含以下几个方面，如图 4-21 所示。

1）优化数据采集策略。

优化数据采集策略，细化数据分类，根据数据分类制定不同的采集周期和采集模式策略。统一规划数据采集策略，灵活配置数据采集接口和调度策略等内容。

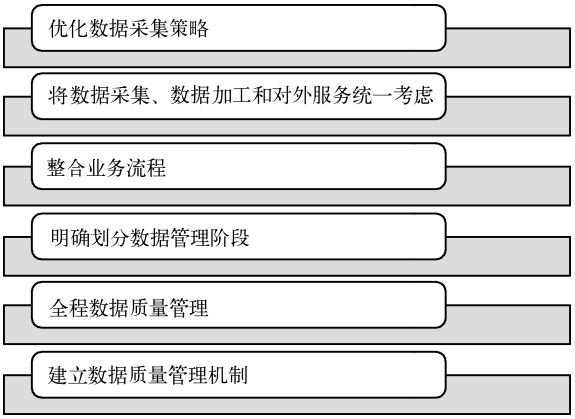


图 4-21　某金融行业数据架构的优化

例如，将客户的收入和个人资产独立采集和存储，用于不同的产品加工和应用。根据不同数据源业务发生的频率和周期，采用不同的数据采集策略。可以引入一些市场化的操作，让一些合作机构辅助数据采集的工作，扩大采集的范围，减少本系统采集的压力，同时增加数据采集的灵活性。

2）将数据采集、数据加工和对外服务统一考虑。

通过监控和调度管理实现任务之间的协调工作。统一监控各生产加工环节任务，根据阈值指标报警异常情况，建立针对事故、风险的应急处理机制，以优化资源的使用。

对于系统来说，主要考虑数据采集、数据加工和对外服务三大核心业务，它们是整个价值链优化的基础，如图 4-22 所示。

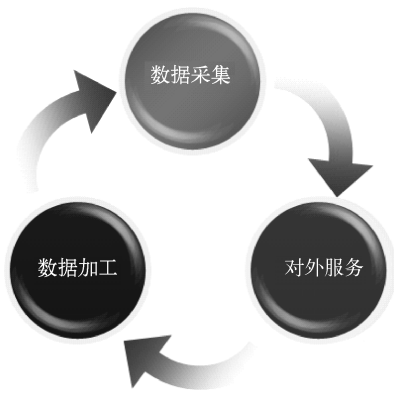


图 4-22　三大核心业务

● 数据采集

数据采集主要是建立稳定、高效的数据传输链路。建立数据采集的应急调整机制和监控调度机制。

● 产品加工

产品加工加强对数据产品加工能力的预测，尽量减少因为数据加工的问题而造成的对外服务的影响。

● 对外服务

对外服务将市场需求预测作为对外服务策略的重要依据，提高服务的准确性，按照服务水平信息，优化采集、加工、服务环节。

3）整合业务流程，加强信息系统支撑，尽量减少手工干预工作，提高自动化程度和系统的总体处理效率。

4）明确划分数据管理阶段，同时加强数据质量、查询匹配、数据整合等关键环节能力，打造核心竞争力。数据处理工作包括数据获取和整合、数据存储、对外信息服务三个阶段，质量管理、查询匹配、数据整合等组件作为核心竞争能力。

5）从数据采集、产品加工到对外服务的全程数据质量管理，优化关键质量管理策略，并提供数据质量、数据整合、测试等工具和组件作为公共基础组件。

一般做法是将数据质量工作前移，在入库前保证数据质量。可以采用抽样统计与逐条数

据校验相结合的校验方式，通过数据抽样和统计的方法，规避系统性数据错误，统计历史记录，作为制定数据质量提升策略的依据。同时，将手工质量管控工作与信息系统相结合，通过相关管理机构进行质量检查和质量绩效管理，提高数据质量。

6）建立数据质量管理机制，确保数据质量达到“适用”的要求，并且是“可管理的”，确保数据带来更大的社会和商业价值。

查看数据质量管理方面手段是否单一，建立数据质量跟踪和反馈机制，明确相关环节部门的权限和职责。定义数据质量，并明确各阶段数据质量管理要点，量化管理并制定相应激励措施。

4.7 某金融行业数据架构案例描述

下面分析某金融行业数据架构相关案例，如图 4-23 所示。

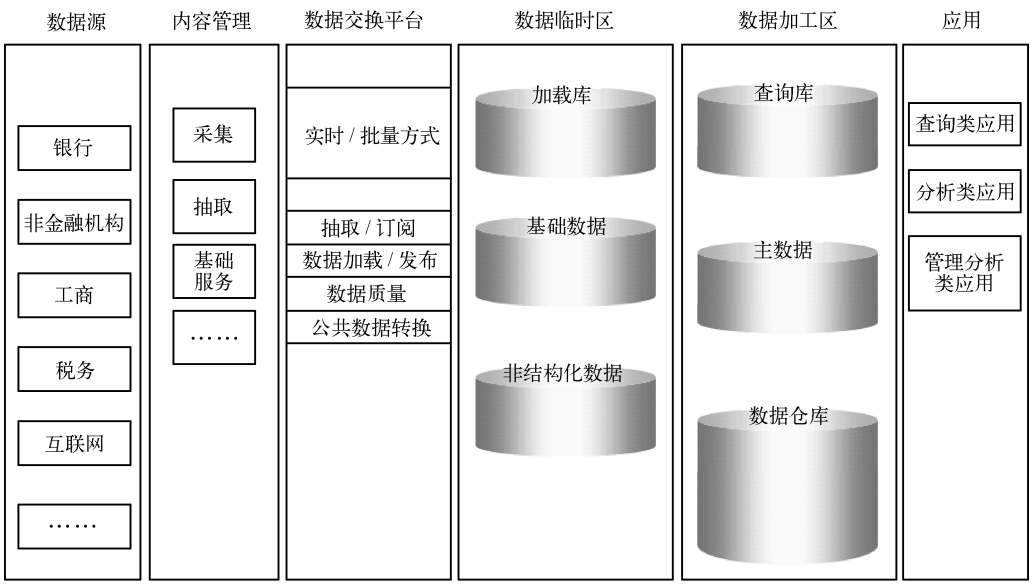


图 4-23　某金融行业数据架构相关案例

对于该数据架构，我们详细了解一下加载库、基础数据、主数据、数据仓库、数据交换平台、产品加工流程、数据架构实施规划和系统切换规划等内容。

4.7.1 加载库

加载库可以作为系统的数据质量控制中心，是合格数据进入到系统的唯一途径。加载库可以分成数据缓冲区和数据加载区。

缓冲区的目的是为了数据交换而设定的临时存储区，加载区是存储贴数据源的数据，一般只存储上一期的数据，为后续的逻辑校验做准备。

缓冲区数据和加载区数据关联进行逻辑校验，如图 4-24 所示。

缓冲区的数据首先经过格式校验，校验通过后再和加载区数据关联进行逻辑校验，校验都通过的数据存储到基础数据库中，如图 4-25 所示。

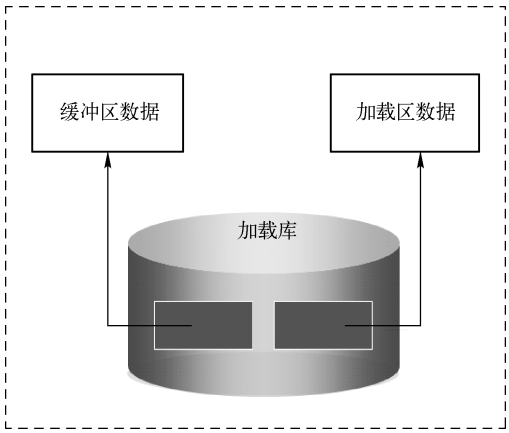


图 4-24　关联进行逻辑校验

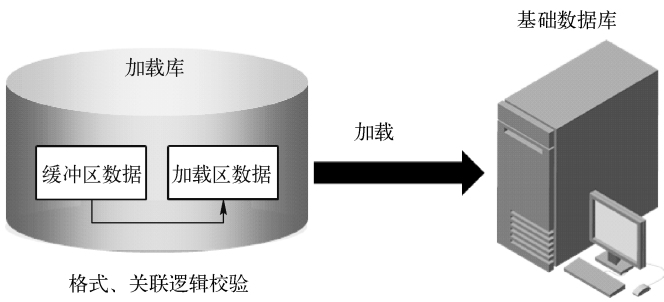


图 4-25　格式、关联逻辑校验

一般来说，数据加载区只存储最近一期数据，如果是新增数据，则直接插入到加载区中，如果是更新数据，则直接替换掉加载区上期的数据。最后，定时地将批量数据加载进基础数据库中。

4.7.2　基础数据

基础数据是系统唯一可信的数据源，它主要存储校验通过的数据，同时也可以存储非结构化数据结构化的内容。存储的期限可以根据业务需求去制定。基础数据库的数据可以到查询库、主数据和数据仓库中，如图 4-26 所示。

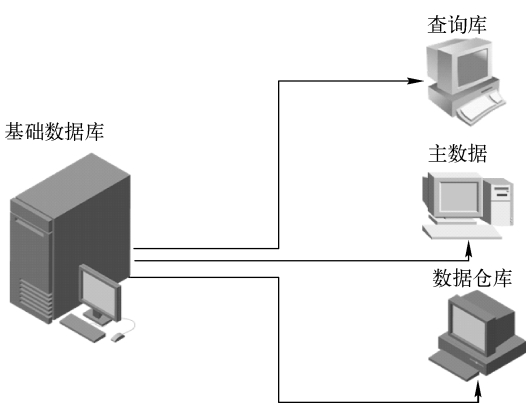


图 4-26　基础数据

4.7.3 主数据

主数据在整个数据架构的作用就是对身份信息的识别和归并，基于业务规则的识别、合并和覆盖原则，实现身份信息的唯一识别，同时增强信息的可信度。

身份信息可以使用唯一号码进行标识。然后将加工数据统一后，再对其他数据库供数。如图 4-27 所示，主数据将加工后的身份信息批量同步到查询类应用、数据仓库中。

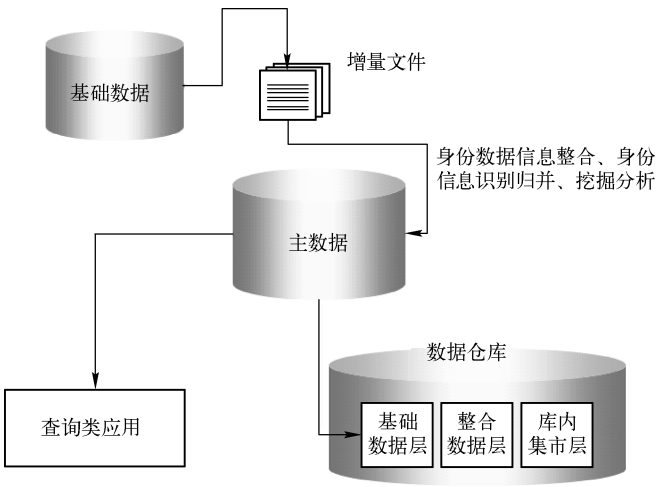


图 4-27　主数据

主数据相关技术包括主体数据的识别、主体数据的整合、主体数据的归并和主体数据关系的挖掘，如图 4-28 所示。

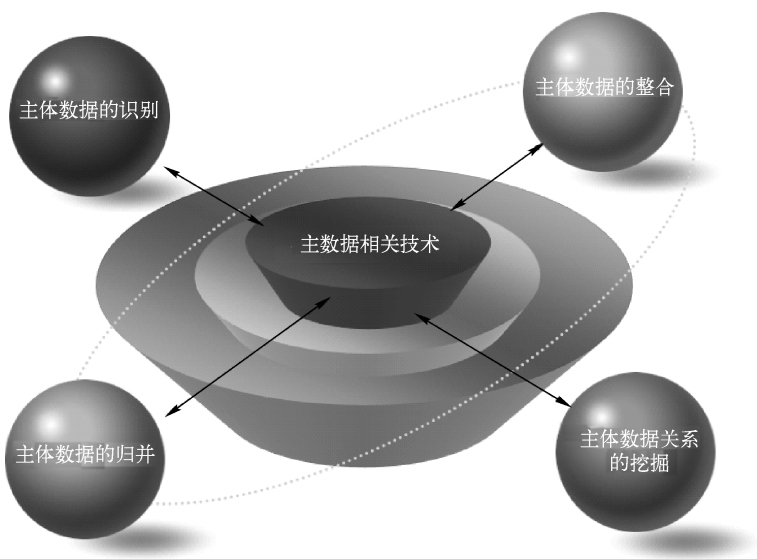


图 4-28　主数据相关技术

下面对主数据相关技术进行详细描述：

（1）主数据的识别

可以灵活地定义主体识别规则。例如，通过个人姓名、证件类型和证件号码识别个人身

份。如果识别规则复杂，则匹配效率低。然而，如果识别规则过于简单，则会导致匹配精度不高。

（2）主数据的整合

主数据的整合是对信息唯一码的分配，在主体识别的基础上，对新增主体信息分配唯一码，主体唯一码与原码比较后，分配唯一码，并且建立唯一码与原码的关系。

（3）主数据的归并

主数据可以灵活定义归并规则，但首先应该定位主体信息疑似名单，进行疑似客户的确认。可能需要业务人员的确认。

（4）主数据关系的挖掘

例如，主数据关系的挖掘可以包括个人与个人关系的挖掘、企业与企业关系的挖掘、个人与企业关系的挖掘等内容。

4.7.4 数据仓库

数据仓库整合系统的全局信息，包括基础数据层、汇总数据层和库内集市层。数据仓库中的数据包含历史数据，它记录了系统从过去某一时间点到目前各个阶段的信息。

一般来说，数据仓库不进行删除操作，通过这些历史信息，可以对企业的发展历程和未来趋势做出定量分析和预测。

数据仓库的数据来源是基础数据库、查询库和主数据中的数据，如图 4-29 所示。

一般来说，数据仓库的数据存储粒度较细，存储时间周期较长，基础层、汇总层和集市层之间的数据交换可以通过数据交换层完成。集市中的数据主要是统计性的，对明细数据保存较少。

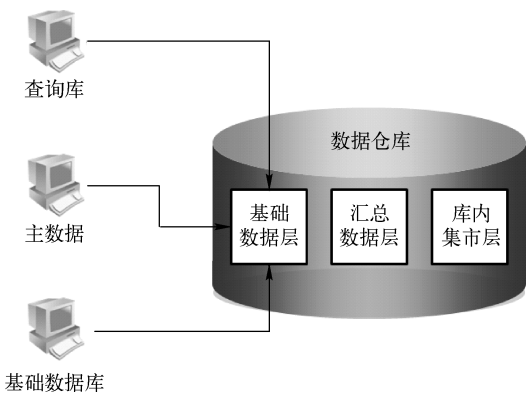


图 4-29　数据仓库的数据来源

4.7.5 数据交换平台

数据交换平台包括外部交换和内部交换两个部分，如图 4-30 所示。

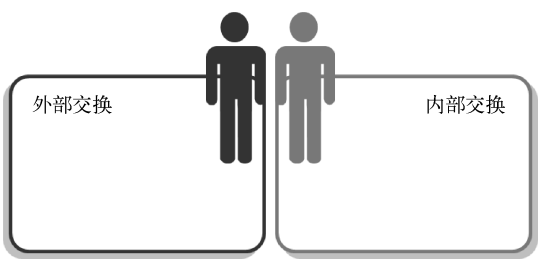


图 4-30　数据交换平台

外部交换：通过交换平台将外部文件数据加载进内部系统。

内部交换：是指系统内各个数据库之间的数据交换。

例如：

1）校验通过后的数据通过数据交换层到基础数据库中。

2）基础数据通过数据交换层到查询库、主数据、数据仓库中。

3）查询库的产品数据通过数据交换层到数据仓库中。

4）主数据库加工后的身份数据通过数据交换层到数据仓库中。

5）非结构化数据的元数据信息通过数据交换层到基础数据库中。

6）数据仓库加工后的结果数据通过数据交换层到分析类应用中。

数据交换平台的功能包括数据抽取、质量检查、数据转换和数据加载，如图 4－31 所示。

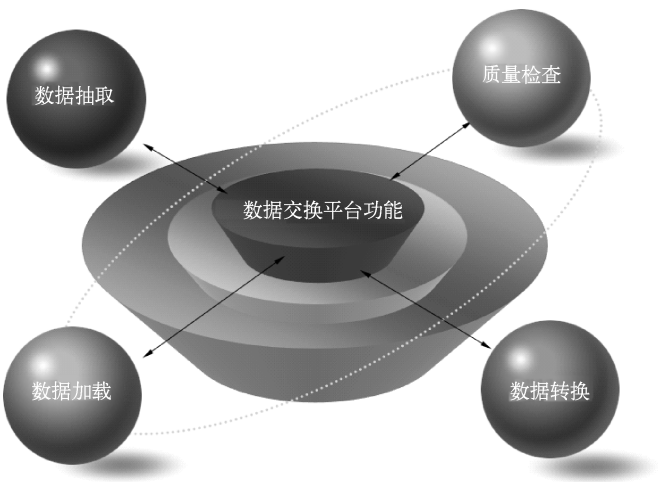


图 4－31　数据交换平台的功能

● 数据抽取功能

数据抽取功能是从数据源层获取原始数据，可以准实时或者实时地获取源系统的增量或者全量数据。抽取的范围是结构化或者非结构化数据。

● 质量检查功能

质量检查是数据交换层的重要工作，经过数据质量的检查，生成满足质量要求的数据文件。

● 数据转换功能

数据转换是对通过质量检查的数据进行转换，然后加载到数据库中，可以按照业务或者技术规则进行转换。

● 数据加载功能

创建可导入的文件，通过工具将数据批量导入到数据库中。

4.7.6　产品加工流程

为了提高产品加工的效率，可以支持加工的并行处理。在目标数据架构中，产品的加工流程包括对查询类产品的加工、管理类产品的加工和挖掘分析类产品的加工。

产品的加工流程如图 4-32 所示，详细介绍如下。

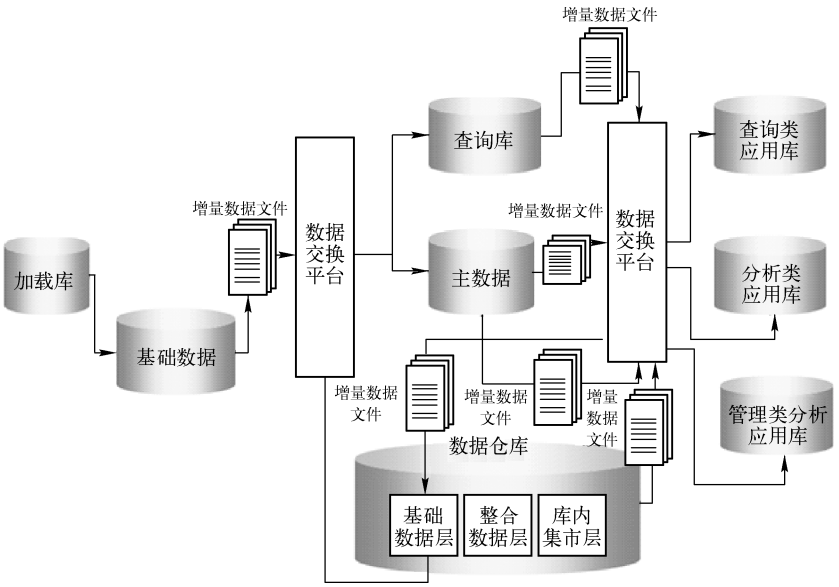


图 4-32　产品的加工流程

1）在加载库中进行数据质量校验，包括格式及逻辑校验。

2）基础数据导出增量数据文件，然后进行增量数据的迁移。

3）对主数据信息进行加工。例如，对基本身份信息的整合、唯一码的分配、疑似主体信息的识别等内容。

4）对数据仓库的数据进行加工，生成各种分析类产品。

5）对查询库的数据进行加工，生成查询类应用产品。

6）最后，对产品数据的加工结果进行迁移。

4.7.7　数据架构实施规划

系统建设策略

关于系统建设策略，前文已经介绍过，主要包括统一开发和推广、快速建设方式。

我们总结一下：项目最好的建设方式是抓住项目的核心应用，对重要核心的需求形成快速突破。然后在统一规划的基础上建设基础平台，统一开发和推广的建设方式和快速建设方式相结合，大大缩减了项目建设周期。因此，可以将整个项目划分成三个阶段，包括：系统建设、应用推广、业务提升，如图 4-33 所示。

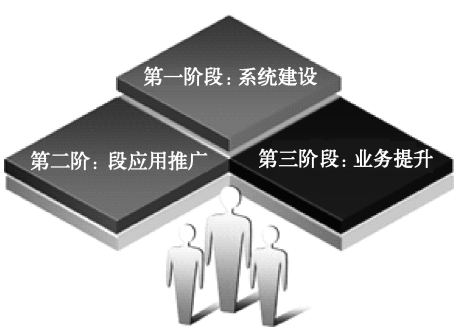


图 4-33　项目划分的阶段

（1）第一阶段：系统建设

遵循需求和总体架构设计的要求，完成核心应用的开发，同时搭建软硬件基础平台。例如，

在项目开始阶段，首先建设基础设施、容灾系统、数据采集系统、数据处理系统、产品加工系统、对外服务系统。然后，在此基础上可以建立客户服务系统、数据质量管理系统、管理分析系统等。

（2）第二阶段：应用推广

选择试点，试运行系统，然后按照推广计划，逐步推广第一阶段建设的核心系统。例如，在此阶段可以建立基础设施建设二期、数据采集系统二期、产品加工系统二期、容灾系统建设二期、对外服务系统二期、数据仓库等，同时应该统一管理和实施。

（3）第三阶段：业务提升

主要配合业务运营，优化系统。在总体架构的基础上，完成新增需求和应用的建设。同时可以拓展系统的数据采集、服务对象和产品的范围。

在项目建设时，也需要考虑可能存在的风险，如组织风险、业务变化风险、技术风险和管理风险等。

4.7.8 系统切换规划案例

系统切换规划的原则，如图 4-34 所示。

（1）稳定过渡的原则

系统推广和切换需要保证稳定过渡。

（2）系统影响最低原则

尽可能减小对原有系统的影响。

（3）风险最小原则

在切换过程中，不能对正常业务造成任何影响。

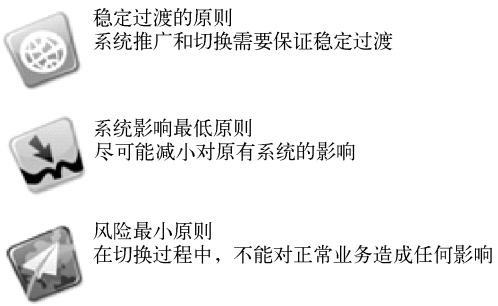


图 4-34 系统切换规划的原则

下面对系统切换方案进行详细描述。

方案一概述

新系统可以不支持一代数据采集接口和查询服务接口。此方案使得未切换源系统的数据采集只能在旧系统中完成。对于数据采集，只能在旧系统和新系统中同时进行，为了保证数据的一致性，需要对新旧系统的数据进行双向同步，直到新系统推广完成，旧系统始终需要并行，当切换完成之后，旧系统才可下线。

对于源系统数据采集端，需要逐步完成新旧切换，这种方式对于双向增量同步实现难度很大，第一次新旧系统切换是将数据采集服务随着新系统投产而启用，查询服务依旧使用旧系统；第二次新旧系统切换是将查询服务从旧系统切换到新系统。

总结

新系统不支持旧系统的查询接口，在新系统的查询服务启动之前，查询服务都在旧系统进行。在新系统全部完成切换之后，查询服务从旧系统切换到新系统。

方案一的工作主要是对新系统的接口开发和测试工作，旧系统不需要进行额外的接口开发和测试。为了不停止对外的查询服务，旧系统必须保持全量的数据，会一直并行到新系统全部完成切换为止。最难的工作和技术就是对新旧系统的数据库双向同步。

方案一的工作流程如图 4-35 所示。

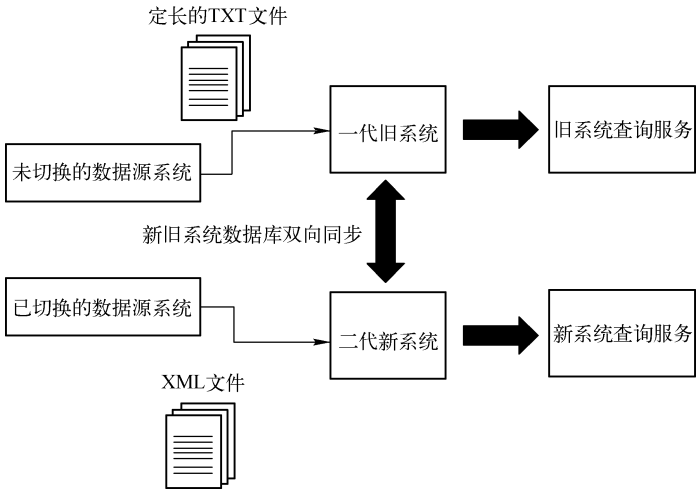


图 4-35　方案一的工作流程

方案二概述

方案二是在方案一的基础上修改的，重点在于对查询服务接口的修改。新系统不支持旧系统的数据采集接口，但是支持对旧系统的查询服务接口。新系统的数据采集和对外服务一次性进行切换，未切换的数据在旧系统中进行采集，已切换的数据在新系统中进行采集。查询服务支持旧系统，同时也支持新系统。当新系统推广完成之后，旧系统可以下线。需要保证新系统是全量数据。方案二的工作流程如图 4-36 所示。

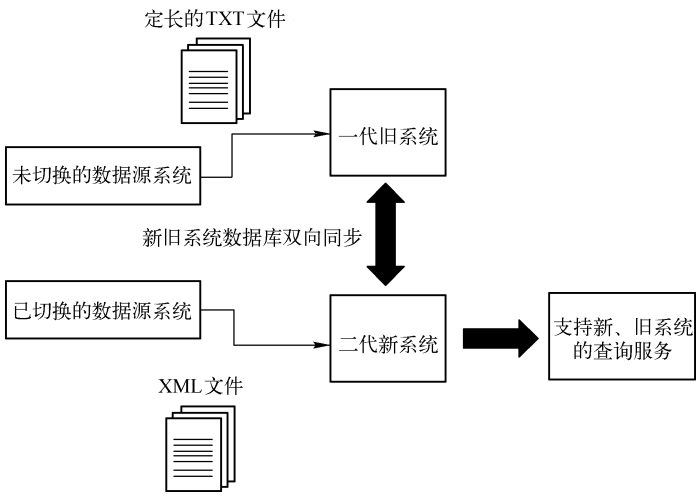


图 4-36　方案二的工作流程

方案三概述

方案三与方案一类似，新系统和旧系统互相支持对方的数据采集接口，但是新系统不支持旧系统的查询服务接口。

当首次切换时，查询服务使用旧系统，随着切换的慢慢推广，一直到完成，新系统的查询服务正式使用。新旧系统会一直并行，直到上线完成之后。最大的难点是对数据一致性的

校验。方案三的工作流程如图 4–37 所示。

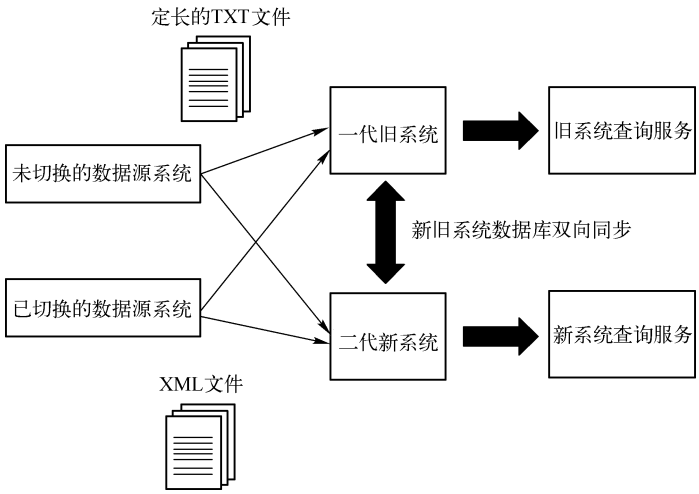


图 4–37　方案三的工作流程

方案四概述

方案四与方案二类似，新系统和旧系统互相支持对方的数据采集，新系统支持对旧系统的查询服务接口。采集数据增量双向加载，并行至推广完成。难点是对数据一致性的校验。方案四的工作流程如图 4–38 所示。

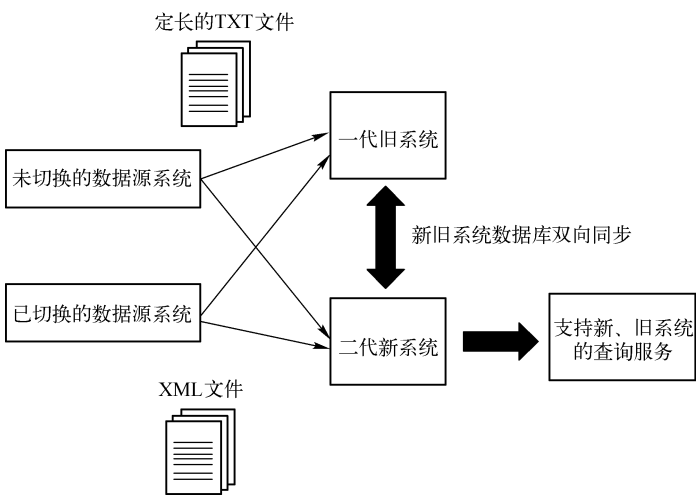


图 4–38　方案四的工作流程

方案五概述

新系统支持旧系统的数据采集接口，但是不支持旧系统的查询服务接口。对于投产切换，数据采集与查询服务可以分成两次切换，切换完成后，全部的数据在新系统中进行采集。查询服务在推广完成之后再切换到新系统中。新旧系统会一直并行，直到推广结束。此方案要求新系统保持全量数据。方案五的工作流程如图 4–39 所示。

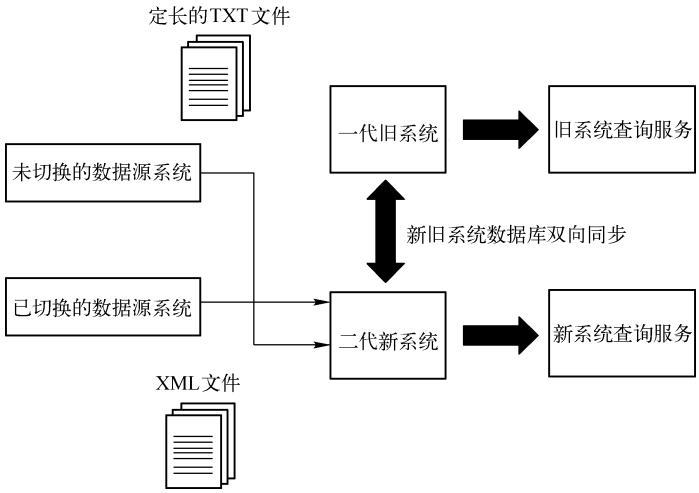


图 4-39　方案五的工作流程

方案六概述

新系统支持旧系统的数据采集接口、查询服务接口。新系统对外服务一次性切换完成。在切换稳定后，旧系统可以选择下线。此方案要求旧系统保持全量数据。方案六的工作流程如图 4-40 所示。

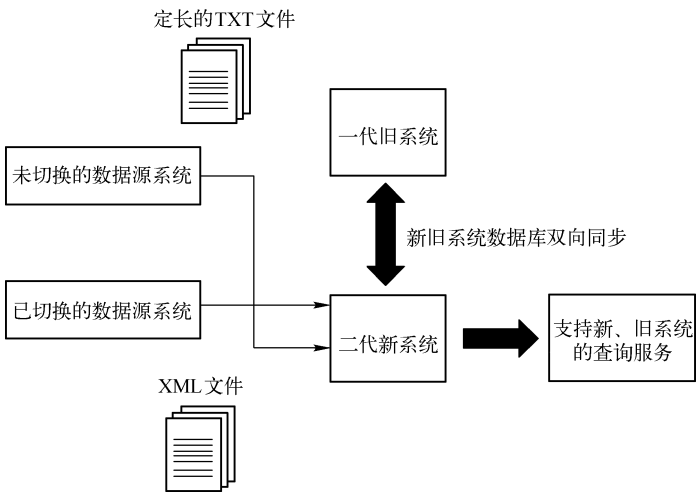


图 4-40　方案六的工作流程

方案七概述

新系统支持旧系统数据采集接口、查询服务接口。新系统一次性切换全部的数据采集和查询服务接口，然后逐步推广。新系统支持旧系统的数据采集与服务接口，服务一次性切换，无须新老系统并行。方案七的工作流程如图 4-41 所示。

但是方案七需要验证新系统采集旧数据的能力，包括：验证新系统对旧系统数据采集接口的支持能力和验证新系统对旧系统查询服务接口的支持能力，如图 4-42 所示。

方案八概述

新系统不支持旧系统数据采集接口、查询服务接口。新系统数据采集与查询服务一次投

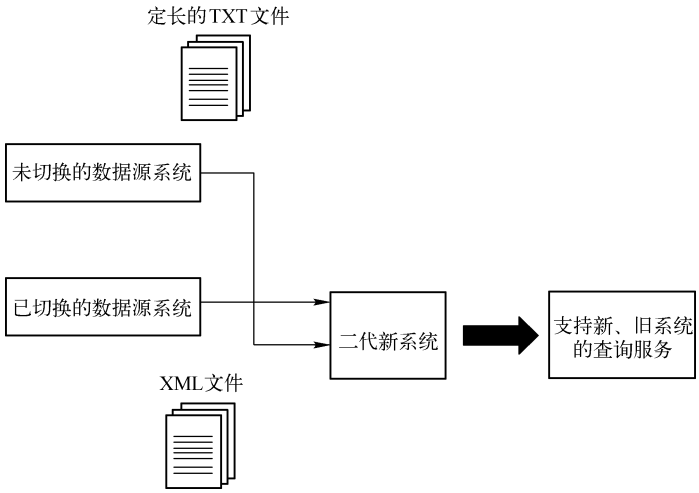


图 4-41　方案七的工作流程

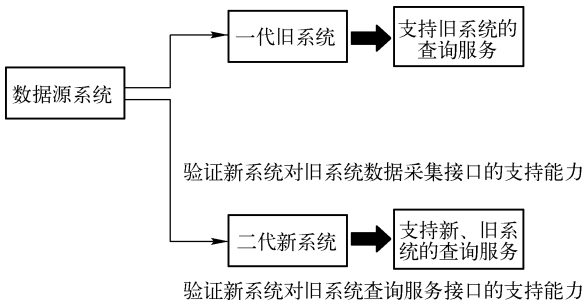


图 4-42　方案七的验证工作

产切换，新系统与旧系统会一直并行，直到推广结束，旧系统才可以择机下线。方案八的工作流程如图 4-43 所示。

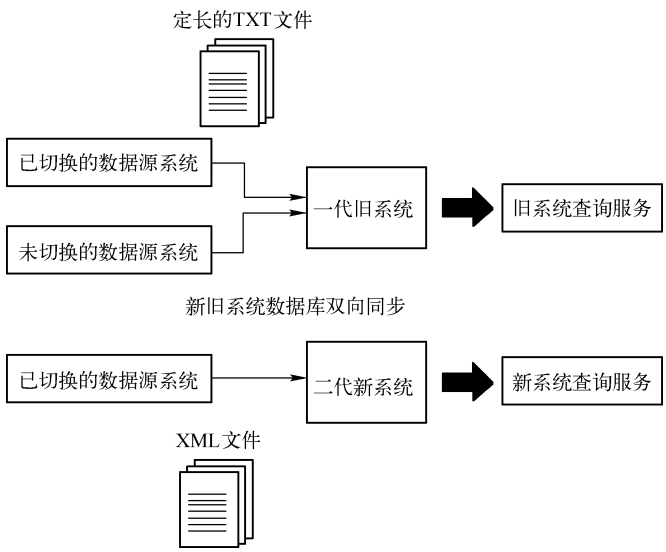


图 4-43　方案八的工作流程

新旧系统数据迁移的问题：

（1）分析新系统需要补录哪些数据

在新系统中，分析数据采集接口比旧系统采集接口增加了哪些内容，有哪些历史数据可以补录到新系统中。在新旧系统切换的时候，需要将历史数据一次性地提交到新系统的数据库中。也可以在系统切换前，提前将历史数据补录到新系统中。

（2）制定海量数据的迁移方案

第一种方式：使用数据迁移程序进行迁移，如图 4-44 所示。首先将源数据导出成原始数据文件；经加工后成为中间数据文件；然后将文件直接导入生产数据库中。

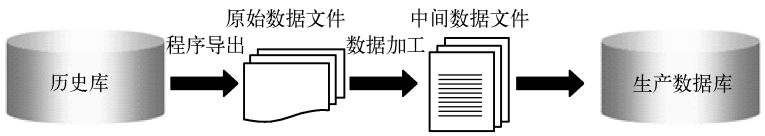


图 4-44　海量数据迁移的第一种方式

对于海量数据的迁移时间需要进行测试和验证。

第二种方式：使用数据迁移程序和中间库，如图 4-45 所示。可以采用中间库，如果数据迁移时间超出投产时间窗口，考虑分批导入的方式。

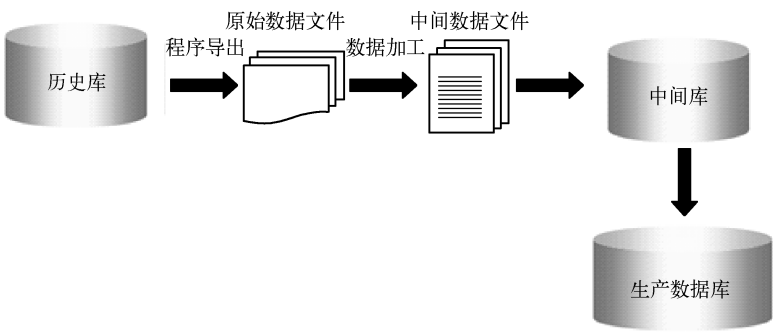


图 4-45　海量数据迁移的第二种方式

小结

- 一般来说，项目阶段分成以下几个部分：项目启动阶段，现状评估、高阶需求分析阶段、架构设计和规划阶段以及实施规划和运维阶段。
- 在系统总体规划过程中，主要包含三个过程：现状分析和需求分析阶段、总体规划设计阶段和总体架构实施规划阶段。
- 系统建设策略主要包含以下几种方式：统一开发、统一推广，快速建设方式。
- 项目阶段的建设计划主要包含以下几个方面：项目启动、需求分析、系统设计、开发和测试以及项目验收等。
- 预算主要包含两个方面的内容：一是对硬件、软件平台、应用软件和各种服务的投资和维护的费用估算，二是对人工服务费用的估算。

- 针对某金融行业信息化建设，可以分成以下几个任务：

（1）信息采集任务

信息采集任务主要是充实采集内容，优化采集方式，根据业务需求，动态地增加采集信息。例如，在个人欠款信息中增加欠款发生的日期。同时需要扩大对公共信息的采集，包括各种的税务信息、司法信息和电信信息等。

（2）数据处理和加工任务

数据处理和加工任务是建立数据处理和快速加工响应机制，能够将各种新业务快速纳入到系统中，提高数据的自动化处理和快速加载能力。

例如，可以将客户的信用评分能力、身份验证、关联查询、风险预警和各种的数据统计功能快速接入到系统中。

（3）应用

应用是建立多样化的产品交付方式，如离线交付、专网交付等，尽量做到 7 ×24 对外服务。

- 数据分布主要包括数据业务分布和数据系统分布。数据分布可以分析数据业务和业务各个环节的创建、修改和删除关系，同时可以分析数据在应用系统中的数据结构和应用系统各个模块之间的关系。
- 在规划数据分布时，需要考虑合适的技术方案来满足以下需求：

1）明确不同位置之间的数据定位和数据流向。

2）保证对海量数据的快速加载和不同数据库之间数据的快速增量迁移。

3）保证海量数据的快速产品加工。

4）应该适应数据采集的多样化、产品加工的多样化和对外服务配置化等特点。

5）可以适应数据的纠错更新机制。

- 对于数据架构的流转来说，主要是降低数据冗余度、提高数据一致性，进而达到灵活、高效的目的。
- 某金融行业系统的纠错更正需求主要包括基础数据的数据纠错更正需求、查询库的数据纠错更正需求、主数据的数据纠错更正需求和数据仓库的数据纠错更正需求。
- 基础数据可以作为唯一可信数据源，在基础数据做的任何修改也都会通过增量的方式同步到数据加工区中进行加工，然后在应用层得到体现，因此，尽量在基础数据中进行纠错更正，这样有利于数据的一致性。但是为了更好地控制数据，应该严格管理数据纠错更正的权限，所有的动作都应该被记录，以备后续查询使用。
- 对于查询库的数据纠错更正需求，一般是发生在客户提出异议申请之后，经过系统确认是否是数据源存在错误，由源系统在自己系统上经过检查，确认是数据错误之后，登录到本系统进行数据纠正。
- 主数据主要包含身份整合信息，针对不同的信息采用不同的整合方式，一般都直接在主数据中修改信息。
- 对于系统的在线纠错更正，需要保证数据的一致性和完整性。在线纠错更正的请求应该尽可能发生在基础数据库中，因为当更新完基础数据后，再通过特殊的数据加工迁移到主数据、数据仓库和查询库中。对于已经加工完成的数据进行在线纠错更正，如果无法通过修改基础数据中的数据来实现在线纠错，只能考虑在加工区中修改数据。

对于所有的在线纠错更正相关操作，必须保留痕迹，从而保证数据的可追溯性。

- 当客户提交在线纠错更正请求后，将更新基础数据库、数据仓库、主数据和查询库中对应的数据。同时记录数据变化的情况，从而确保数据的可追溯性。
- 为了提高产品加工的效率，可以支持加工的并行处理。
- 数据仓库整合系统的全局信息，包括基础层、汇总层和集市层。数据仓库中的数据包含历史数据，它记录了系统从过去某一时间点到目前各个阶段的信息，一般来说，数据仓库不做删除操作，通过这些历史信息，可以对企业的发展历程和未来趋势做出定量分析和预测。
- 数据交换平台包括外部交换和内部交换两个部分。外部交换是指通过交换平台将外部文件数据加载进内部系统。内部交换是指系统内各个数据库之间的数据交换。
- 某金融行业数据架构的优化主要包含以下几个方面：

1）优化数据采集策略。

2）将数据采集、数据加工和对外服务统一考虑。

3）整合业务流程，加强信息系统支撑，尽量减少手工干预工作，提高自动化程度和系统的总体处理效率。

4）明确划分数据管理阶段，同时加强数据质量、查询匹配、数据整合等关键环节能力，打造核心竞争力。

5）从数据采集、产品加工到对外服务的全程数据质量管理，优化关键质量管理策略，并提供数据质量、数据整合、测试等工具和组件作为公共基础组件。

6）建立数据质量管理机制，确保数据质量达到“适用”的要求，并且是“可管理的”，确保数据带来更大的社会和商业价值。

- 随着大数据时代的到来，数据应用可以产生更大的机遇和挑战。只有更好地利用数据，才能在未来的竞争中获得更大的优势。一般来说，数据的应用主要包括：报表功能、统计分析和数据挖掘三种方式。

（1）报表功能

报表功能是数据应用的基础，是较为传统和常见的数据应用。报表是决策分析的基础。报表功能的完善、灵活程度能够影响工作的效率。

（2）统计分析功能

统计分析功能是常见的数据应用方式。随着统计分析工具的推广，统计分析在很多行业中得到了越来越广泛的应用。例如，通过假设检验或者方差分析帮助分析经济运行的规律。

（3）数据挖掘功能

数据挖掘是数据统计分析的进一步发展，是对数据的深度应用。

数据挖掘虽然起源于 20 世纪 70 年代，但在最近 10 年内得到了广泛应用和发展，特别是被金融行业、互联网行业广泛使用。

第 5 章　大数据架构与实践

本章目标

通过前几章的学习，我们已经理解了数据架构的工作方法和指导原则，同时也了解了金融行业数据架构的相关案例，还学习了数据架构的流转、加工的处理时序、数据纠错方案介绍、数据架构的优化和数据架构实施规划等内容。

但是，随着数据采集的范围不断扩大，一些例如文档、视频等半结构化和非结构化的数据逐渐成为主要的数据源，可以这样说，80% 的数据可能都来自于非结构化数据，如图像、音频、微博、网页、电子邮件等。商业银行一直饱受着这些大量的非结构化数据没有更好地创造业务价值的困扰，我们可以把大数据视为挑战。

同时对于商业银行来说，大数据更是机遇，客户在不断与银行的交易过程中，创造出多种形式的数据，这也为银行实时或者准实时地分析数据提供了便利，同时可以对客户进行针对性的营销。因此，本章我们将重点介绍大数据。

学习本章后，读者将掌握：

- 大数据的建设背景
- 大数据面临的挑战和重要性
- 大数据的定义和特点
- 大数据下的数据架构
- 大数据分析平台基础框架
- 大数据技术如何落地
- 相关生产厂商大数据技术介绍
- 大数据与云计算
- 大数据和传统商业智能分析
- 大数据在金融行业的应用
- 大数据在其他行业的应用

5.1　大数据概述

5.1.1　大数据的建设背景

“大数据的真实价值就像漂浮在海洋中的冰川，第一眼人们往往只看到冰山一角，而绝大部分都隐藏在表面之下，数据总是从最不可能的地方被提取出来。”这段关于大数据的精彩论述来自维克托·迈尔 – 舍恩伯格所著的《大数据时代》一书。

经过许多年的发展，目前的信息积累已经到了一个新的阶段，它比以往有着更多的信息，数据的增长速度也在不断加快。

据 IBM 公司预测，到 2020 年，全世界产生的数据规模将达到目前数据量的 44 倍，在这

些数据中，只有 1% ~5% 的数据是结构化数据，这意味着非结构化数据和半结构化数据将占据绝大部分。因此，人们创造出了大数据的概念。

在理解大数据之前，首先应该理解什么是数据信息。数据信息好像是地球上的空气，无处不在、四处漂移，如图 5-1 所示。同时这些信息又是看不见、摸不着、无孔不入的。这些信息可以包括文字、图像、声音和影像等。信息实质上是人类思想外化的一种方式。

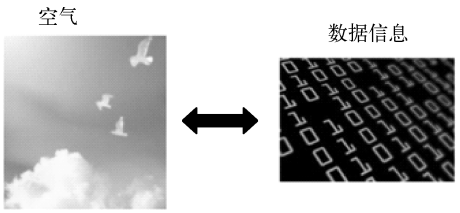


图 5-1　数据信息好像是地球上的空气

那么，什么是大数据呢？

大数据是指巨量的信息，规模巨大，已经无法用常规的软件工具在短时间内进行存储和管理。大数据的主要功能就是预测，可以将算法应用到海量的数据中，预测事件发生的可能性。但是我们不要拘泥于大数据的概念。

目前来说，大数据多数仍然停留在概念上，真正大数据落地的成功案例寥寥无几。我们应该去探寻大数据的真正内涵和价值。如何分析和使用大数据才是本章的重点。

例如，商业银行拥有大量的客户信息和交易信息，特别是客户在互联网上的每一次点击和评论，都是大数据的数据来源。通过对这些数据的分析，洞悉客户的潜在和真实需求。实质上我们每天都在创造着海量的数据，数据在“包围”我们，我们正在进入“大数据”时代。

大数据包括什么数据呢？

例如，交通和天气预报的数据、人们在社交网络上的信息、购物信息，以及各种视频、音频、短信等，均可视为大数据。

一般将 2012 年视为大数据时代的元年。很多行业经过多年的数据积累，已经具备了利用大数据的挖掘分析创造价值的能力。对于金融行业来说，它们每天都处理千万量级的交易数据，在银行卡中也保存了大量的收入和支出信息。进入大数据时代后，如何更好地利用大数据创造财富是不可回避的话题。很多银行可以根据对客户的深入了解，为客户提供多样化和个性化的服务。同时还可以针对相关热点、各种犯罪行为进行预测。特别是在国外已经形成了多渠道的客户分析、天气预测预警分析和交通堵塞预警分析等应用。

在互联网上，我们每天都会留下大量的浏览网页的痕迹。互联网技术很像人的神经系统，可以通过感官获取信息。大数据可以视为人的大脑中枢，各种信息集成到大脑中枢，然后对数据进行整合、集成和挖掘。举例来说，在社交网站上，记录了我们和朋友之间的交往信息。因此，我们应该做好对大数据的管理和利用工作。对于不同的行业来说，大数据都意味着巨大的商业机会，它可以帮助我们提高客户的忠诚度，增强客户的体验感。所以说，对于这些数据的收集和分析已经成为提升企业品牌形象的手段之一。

实质上，大数据在金融、互联网的应用非常广泛，这些企业或商业银行在日常运营过程中产生了大量的数据，尤其在人口众多的国家，大数据的应用更为广泛，通过这种挖掘和利

用大数据的能力，可以大大提高服务的水平。其实大数据为市场提供了各种机会，创造出了巨大的商业价值。同时大数据可以帮助各个企业找到适合自己的发展模式和客户群体，强化自身的特色。

图 5-2　大数据的应用非常广泛

传统的数据分析思维是要求数据准确无误，数据关系清晰。但是大数据的分析思维是接受数据的复杂性，单个数据的重要性不高，主要关注事物之间的关联关系。当我们完成对关联关系分析之后，就可以研究更深层次的因果关系，找出背后的原因。例如，将啤酒和尿布摆放在一起，蛋挞和飓风用品摆放在一起；通过了解人们生活上的喜好，分析患某种疾病的概率；利用人们的社交数据，分析个人的偿还意愿和偿还能力。

对于商业银行来说，为了保证在金融市场的竞争地位，将数据转化为可以洞察的信息和知识，推动业务的发展，提升管理的效率。通过大数据分析平台，接入客户的社交网络，终端媒介产生的各种非结构化数据，构建客户的全方位视图，获取客户的反馈信息和真正需求，才能对银行产品进行合理的规划和设置。

大数据分析可以帮助银行内部加强管理，增强透明度，优化各种业务流程和工作效率。提高银行系统交易的性能，减小运营和管理的压力。

大数据分析还可以帮助银行了解客户的风险信息，建立完善的风险管控体系。另外，可以及时地获取客户的反馈信息，对客户需求进行深入分析，对银行产品进行合理设置。同时构建客户的全方位视图。

例如，根据客户的偏好、年龄、收入、地域、历史购买水平、兴趣广度，构建客户的全方位视图，了解客户最真实的信息。在此基础上，对客户进行细分和风险评估，从而进行有针对性的营销，如图 5-3 所示。

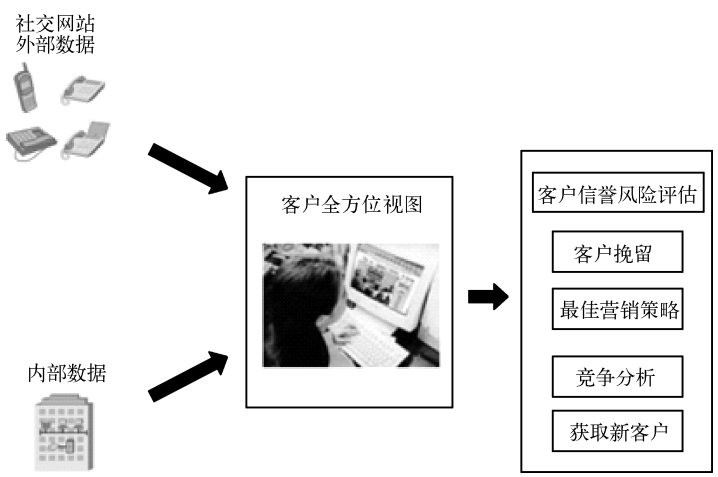


图 5-3　针对性的营销

同时可以制定云计算、物联网等规划，特别是云计算、物联网、社交媒体为大数据提供了丰富的数据来源。随着移动终端技术的应用，特别是数据挖掘技术的发展，已经彻底改变了客户的消费模式。如果从数据的角度来看，我们其实已经进入到了大数据的智能化时代。

我们总结一下国内大数据的建设背景：

国内大数据应用的基本现状较为复杂，目的是为了追求大数据技术而进行各种大数据项目的建设，这样可能会导致很多企业“掉进”以技术为导向的误区。

大数据的项目必须有明确的业务需求，用商业思维来推动大数据的建设，只有这样，大数据的价值才能充分体现出来。

5.1.2 大数据面临的挑战和机遇

1. 在大数据时代，我们面临的挑战

（1）企业或者商业银行将数据的重要性提升一个层次

首先要求企业或者商业银行将数据的重要性提升一个层次，对于数据的应用已经不仅仅是业务经营，而是已经扩展到客户服务和营销领域中，特别是可以通过大数据的应用，预测未来业务发展的方向，这对于数据驱动业务提出了挑战。

（2）大数据管理上的成本大大提高

基于大数据的分析可以让企业高层的经营决策更具有客观性，但是也导致了大数据管理上的成本大大提高。

（3）产品创新不足

在大数据时代，数据不仅仅是企业日常经营活动中的记录，而是一种资产，目前来说，依赖数据标准体系，以及数据架构、数据仓库等手段进行产品的管理和应用。但是在产品创新上仍然不足。

（4）数据整合和数据质量管理的难度很大

对于大数据来说，数据整合和数据质量管理的难度是非常大的。为了保证数据的一致性，应该运用合适的技术和管理手段去保障大数据的应用。

（5）一些企业和银行在数据利用上有一定的局限性

在大数据时代，国内的一些企业和银行在数据利用上有一定的局限性，特别是商业银行，很少有对网点的监测数据进行利用的，导致数据的应用局限在特定的用途和场景中。在国外很多机构中，可以将各种非结构化数据，如影像和视频文件，转化成对用户的行为分析。

（6）应用与理论研究的成本很高

从技术上来说，大数据的应用离不开 Hadoop、云计算。这也增加了应用与理论研究的成本。

（7）业务需求和技术之间的协调

大数据意味着更大的机遇，拥有巨大的应用价值，企业的 IT 技术部门希望业务部门提出大数据具体的分析需求，业务部门希望 IT 技术部门针对大数据提出分析建议。只有协调好业务需求和技术之间的关系，才能发挥大数据真正的作用。

（8）人才方面储备不足

大数据面临着人才方面储备不足的问题。大数据需要企业具备既有 IT 技术，又对业务十分熟悉的复合型人才。

举例来说，某银行的发卡量迅速增长，随着业务迅猛发展，数据也呈线性增长。面对着传统的商业智能分析，旧的系统架构无法支撑大数据的快速增长和灵活分析，无法实现秒级

营销和精准营销。同时大数据分析面临着人才缺失、数据共享难度大和落地困难等问题。虽然大数据面临着各种挑战，但是大数据分析也带来了巨大的经济利益。据全球权威的咨询公司 Gartner 统计，2012 年和 2013 年大数据分别带动了 280 亿美元和 340 亿美元左右的 IT 支出，按照此速度的增长，2016 年全球在大数据上的总花费可能会达到 2320 亿美元。

2. 大数据为各行各业带来了巨大的经济利益

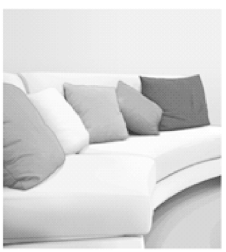
2011 年，大数据为欧美部分产业带来的收益如表 5-1 所示。

表 5-1　大数据带来的经济利益

美国医疗保健业	美国制造业的产品开发和组装环节	美国零售业的净利润	欧洲的公共管理部门	全球个人定位服务供应商
带来 3000 亿美元的价值	节省一半的成本支出	增长了 60%	节省了 2500 亿欧元的支出	增加了 1000 亿美元的收入

对于中国企业来说，大数据技术的研发和投入相对较少，目前很多企业没有利用好大数据。大数据的发展对于我们的启示（见图 5-4）是：

必须把握好大数据技术，推进企业的转型创新



企业制定新的大数据人才战略，以价值体系激励员工

培养洞察分析的能力，以个性化服务赢得客户

图 5-4　大数据的发展对于我们的启示

1）必须把握好大数据技术，推进企业的转型创新。

2）需要企业制定新的大数据人才战略，以价值体系激励员工。

3）培养洞察分析的能力，以个性化服务去赢得客户。

5.1.3　大数据的定义和特点

虽然目前大数据没有明确的定义，但是我们每天都在产生海量的数据，数据将我们“包围”起来，我们正在进入到“大数据时代”。根据 Gartner 的定义，大数据的特征具体涵盖了称为 4 V 的内容：数据量大（Volume）、实时性强（Velocity）、商业价值（Value）、数据多样化（Variety），如图 5-5 所示。

对大数据关注也是因为它蕴藏巨大的商业价值。在有些资料和文档中，将大数据的特征定义为 3 V 特性，包括数据量大、数据多样化以及数据产生频率、更新频率高。在这里我们主要讨论大数据的 4 V 特性。

数据量大：例如，互联网、物联网每天都在产生大量的数据，数据量持续以前所未有的速度增加。数据量大是大数据相关的重要特征之一。

实时性强：主要是指数据产生的速度快，数据变化的频度可以到毫秒级。举例来说，我们每天都通过传感器或者监控视频产生新的数据，数据以比从前更快的速度产生、获取和分

图 5-5　大数据的 4V 特性

析。特别是订单、微博、监控视频、传感器、支付等每时每刻都在不停地产生数据。

数据多样化：多样化是指数据类型的复杂性和数据种类的繁多，用来描述不同类型的数据和数据源。随着传感器和一些智能设备的发展，数据呈现了爆炸性的增长态势，包括如电子表格、声音、图片、视频、文本、微博、传感器数据、点击流、日志文件、手机呼叫、地图 GPS 等内容。

商业价值：通过对大数据的挖掘和分析，可以发掘出巨大的商业价值。

我们总结来说，大数据的定义就是通过快速采集、挖掘和分析，从大数据量、多样化的数据中获取价值。形象地说，大数据就是沙里淘金的过程。

对于传统的数据仓库技术和大数据处理，它们之间最大的区别就是数据仓库更多地是对过去事物的分析，而大数据主要分析我们即将面对的问题，也就是预测和分析未来的情况，具有更高的价值。

对于大数据来说，有结构化数据、半结构化数据和非结构化数据三种类型。

1）结构化数据：主要存在于关系型数据库，在过去几十年里一直是主流的应用。

2）半结构化数据：包括类似于电子邮件、文字处理文件以及网上新闻等内容。

3）非结构化数据：包括社交网络、物联网、移动计算和各种传感器产生的各种信息，可以有音频、视频和图片等内容。目前超过 80% 的数据属于非结构化数据。

大数据对于系统的需求主要包含了高性能、高存储、可扩展和低延迟等几个特性。高性能是指可以高并发地对海量数据进行读写，同时依靠并行处理，快速响应查询、分析。高存储是指对海量数据的存储。可扩展是支持可扩展性。低延迟是指能够快速响应。

下面详细介绍大数据的几个特点：

（1）数据量大

大数据应该有多大呢？

举例来说，1999 年，美国沃尔玛公司的数据仓库容量是 100 TB，2012 年，Facebook 每天的数据量超过 500 TB。目前，互联网上一天的内容就可以刻满 1.68 亿张左右的 DVD，发出的社区帖子在 200 万个以上。

截止 2012 年，数据量已经从 TB 级跃升到 PB 级、EB 级甚至 ZB 级。2008 年全球产生的数据量为 0.49ZB，2009 年产生的数据量达 0.8ZB，2010 年产生的数据量是 1.2ZB，2011 年的数据量已经达到 1.82ZB。目前全世界数据的年增长量达到 50% 左右。又如，2000 年美国

新墨西哥州数字巡天望远镜启用几周后，搜集的数据量就已经超过了天文学历史上的数据总和。这一切都意味着每两年全世界的数据总量就会增加一倍。

据 IBM 公司 2012 年研究报告，在整个人类文明产生的全部数据中，有大约 90% 的数据是过去两年内产生的。到 2020 年，全世界产生的数据量可能会达到今天的 44 倍左右。

（2）实时性强

大数据作为感知世界的“仪表盘”，它的增长速度很快，数据变化与处理的频度可以到毫秒级，例如各种订单、支付、监控等，每天不停地产生着数据，同时对海量数据进行及时分析。对于某些应用来说，要求在几秒钟之内得出答案，否则就错过了最佳时机。这种实时性强的特点也是区别于传统数据仓库和商业智能技术的关键特征之一。

实时性强的原因是数据创建的快速性。目前数据是以传统系统不可能达到的速度在获取、产生和分析。例如，各种的股票实时分析、实时动态的传感数据、各种的交通路况信息、每一秒中淘宝平均成交 178 笔订单等。这种数据产生的速度，已经完全超乎了人们的想象。

（3）商业价值

价值密度低是大数据的一个典型特征。犹如淘金的过程，虽然大多数都是沙子，但是这些沙子中仍然存在着宝贵的黄金，我们需要做的就是将大多数的沙粒去除和清洗掉，将黄金提取出来，如图 5-6 所示。同样对于大数据来说，多数的数据是低价值的，例如影响天气因素的数据很多，但是每一条单独的信息都是价值很低的，只有将这些信息汇总和综合到一起，才能具备对天气预测的能力。

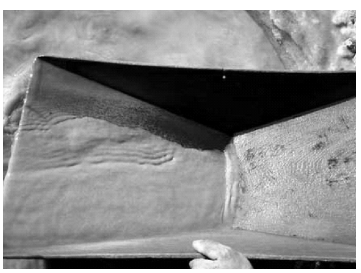


图 5-6　价值密度低

（4）数据多样化

对于大数据来说，数据种类繁多，80% 以上的数据来自于半结构化数据和非结构化数据，如文档、视频、电子邮件等。

随着传感器、智能设备技术的发展，数据的类型呈现多样化的态势，包括文本、微博、音频、视频、传感器数据、日志文件、手机呼叫、地震勘探、气象云图、卫星遥感、物联网、环保监测、舆情监控、地图 GPS 和各种的点击流等。将这些不同类型的数据进行交叉分析，是大数据的核心技术之一。特别是语义分析和各种地理位置信息技术都会在大数据时代得到广泛应用。

5.1.4　大数据下的数据架构

分析前一章的数据架构规划图，其中在数据临时区中有非结构化数据一项，如图 5-7 所示。

如何处理非结构化数据呢？如图 5-8 所示。

首先可以使用“网络爬虫”手段收集非结构化的数据，在 Hadoop 平台中建立非结构化信息的标签、摘要、索引、日志、内容等，然后提取结构化的元数据信息，如类别、摘要等内容，最后与基础数据中的结构化数据进行整合。

对于流数据来说，它强调的是实时处理与分析，而不是数据存储，所以只在内存中进行处理，不落在具体的磁盘中。随着时间的流动，它只对一段时间内的数据进行处理。例如，

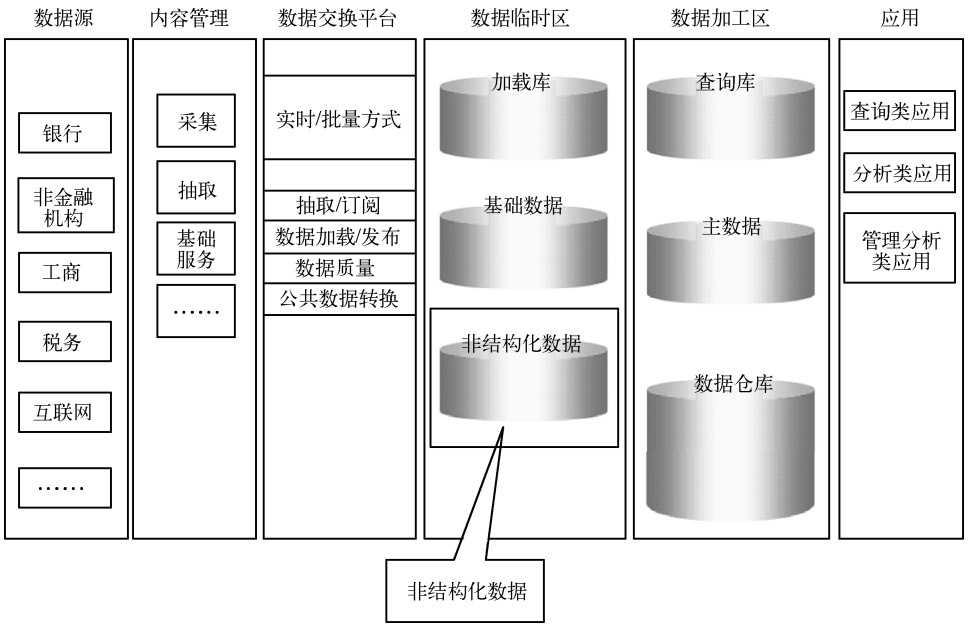


图 5-7　数据架构规划图

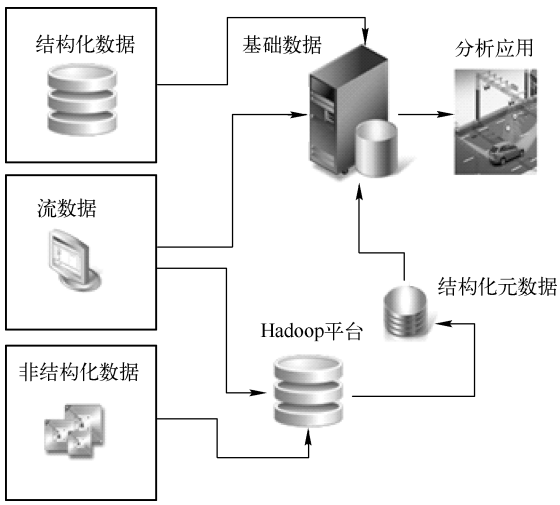


图 5-8　非结构化数据的处理流程

它把银行交易系统的日志信息实时地放到流平台当中，进行反欺诈的实时监测，流计算一般可以在几秒钟之内对海量数据中的异常行为进行预测和分析。

总之，对于基础数据来说，它存储的都是有用的信息，类似于存储的都是“黄金”。Hadoop 平台存储的是从网络中收集来的沙子，我们的目的就是将沙子里的黄金筛选出来。非结构化数据通过网络爬虫等手段把数据放入到 Hadoop 平台中，再转化成结构化数据进行分析。

大数据的一个重要应用就是舆情分析，利用网上收集的信息，如正面、负面的信息，分析人们的情感和进行预警分析。舆情分析包括企业的声誉分析、品牌分析、服务质量分析、

竞争产品分析、市场动态跟踪等内容。

随着业务的扩展，用户应该对大数据进行数据架构规划，如图 5-9 所示。

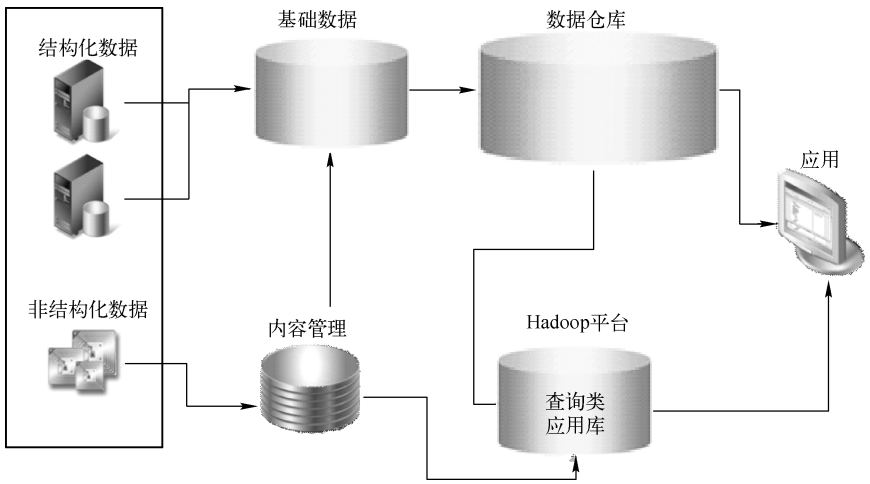


图 5-9　大数据的数据架构规划

大数据的数据架构规划可以采用 Hadoop 技术，通过与结构化数据的关联，进一步拓展对非结构化数据的处理，其中数据源包括结构化数据、半结构化数据、非结构化数据，特别是非结构化数据和半结构化数据通过网络爬虫的方式收集信息，经过内容管理平台的处理，将非结构化数据、半结构化数据结构化处理，其中可以将内容管理平台处理得出的非结构化数据的元数据信息存放到基础数据存储中。

对于 Hadoop 平台来说，它是基于 HDFS 或 Hbase 存放非结构化/半结构化数据。对于应用来说，它是基于结构化数据、半结构化数据、非结构化数据进行综合分析。

对于我们熟知的流数据，具有哪些特性呢？如图 5-10 所示。

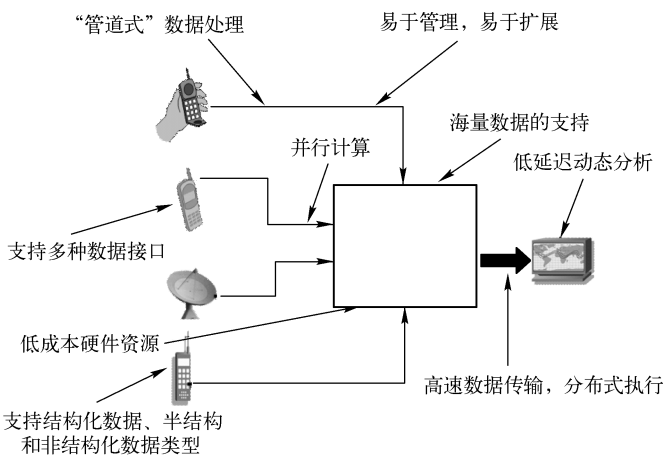


图 5-10　流数据特性

流数据具有“管道式”的数据处理方式，易于管理、易于扩展，支持并行计算和多种数据接口，以及各种低成本硬件资源。同时支持结构化数据、半结构化数据和非结构化数据

类型，也支持高速数据传输和低延迟动态分析等。

流分析的主要过程如图 5-11 所示。

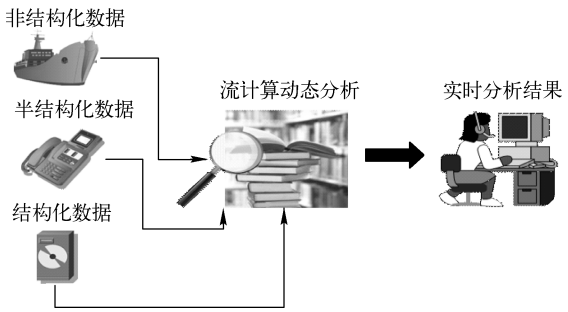


图 5-11　流分析的主要过程

流数据有哪些作用?

流数据可以保障数据处理的实时性，提高数据分析和决策的实时性，同时实现数据挖掘、分析和展现的有效融合，降低延迟性。

大数据的处理流程

大数据的处理流程主要包括大数据的采集、对数据的统计分析和对数据的挖掘等三个阶段。

（1）大数据的采集

通过数据库接收来自客户端的数据，同时进行查询和处理。例如，Oracle、MySQL、HBase 和 MongoDB 等，这些产品有各自的特点。

（2）对数据的统计分析

对于繁杂、粗糙的、庞大的数据来说，一旦经过提炼和加工，便可能带来巨大的经济效益。可以利用分布式技术对海量数据进行查询和汇总。特点是查询的数据量大，查询的请求多。包含的产品包括 Hadoop、Oracle Exadata，可以做离线分析和实时分析。

（3）对数据的挖掘

对查询的数据进行挖掘分析，满足高级的数据分析，但涉及的算法复杂，数据量巨大。

银行每天都在处理千万量级的交易，它记录了我们每一笔的收入和支出情况，包括资金的汇入和汇出情况。在未来，数据将以 40% 的速度快速增长，大数据为银行带来的价值是不可估量的。

商业银行可以分析客户使用网银的习惯，将最常用的功能展示在登录界面上，省去了用户在菜单中跳转所花费的时间。同样，我们也可以基于对数据的采集和识别，评估信用卡申请人提交的信息和证明材料，包括其他信用卡发行商提供的申请人交易信息和还款信息。一些营销专家和数据分析专家可以借助数据挖掘工具，对用户的信息进行提炼和分析，然后基于对海量数据的挖掘，进行风险控制和用户营销。

5.1.5　大数据分析平台基础框架

大数据分析平台主要包括大数据基础平台、平台组织团队、数据治理和应用系统等。

（1）大数据基础平台

在统一调度下，整合各类数据，以支撑应用。

（2）平台组织团队

平台组织团队主要包括大数据需求分析、平台建设和运维等组织和团队。

（3）数据管控

建立数据标准管理、数据质量管理、元数据管理和数据生命周期管理机制，为基础平台提供保障。

（4）应用系统

建设各类数据应用系统，发挥大数据的价值。

5.1.6 大数据技术如何落地

很多企业都知道大数据应用的重要性，但是不清楚如何更好地利用大数据，很多企业在大数据应用时最大的难题就是如何保证大数据的落地。下面介绍大数据如何落地，如图 5-12 所示。

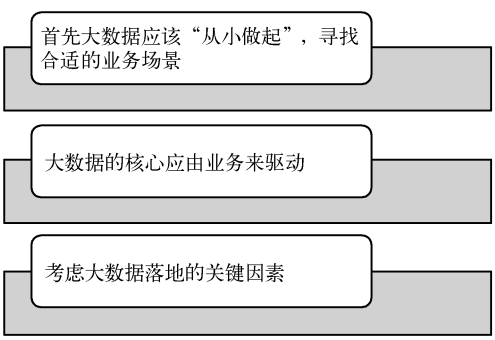


图 5-12 如何保证大数据的落地

（1）首先大数据应该“从小做起”，寻找合适的业务场景

企业应该避免缺乏具体且可测量的相关应用，对企业面临的问题和各种业务需求进行深入分析，理解企业最迫切的需求是什么，从哪里入手最容易产生效果。

（2）大数据的核心应由业务来驱动

对于企业来说，大数据的核心应由业务来驱动。特别是跨行业的业务场景，如数据探索、风险管理、反欺诈等。具体的行业主要包括医疗、零售、商业银行等，它们都有自己独特的业务需求，如基于地理位置的精准客户营销。

（3）考虑大数据落地的关键因素

大数据落地的关键因素包括：如何实时获取非结构化数据，如何组织和集成大数据，如何使用工具和技术分析大数据，如何为企业提供实时的、共享的、全面的业务决策分析。

5.2 大数据相关技术概述

大数据相关的技术主要包括：云计算、物联网、分析工具、社交工具、移动计算等，如图 5-13 所示。

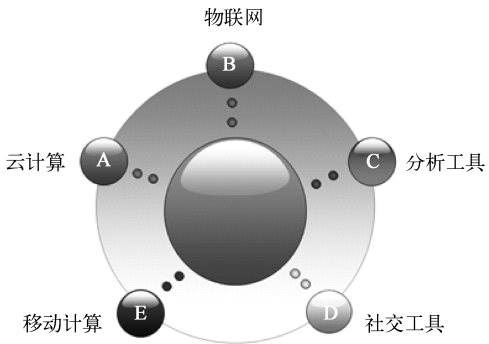


图 5-13　大数据相关的技术

其中，云计算技术是为大数据时代进行的技术准备，它可以突破边界存储技术。而物联网技术主要是证明世界是联系的，而我们现在火热的智慧城市就是利用物联网技术实现的，将来还会出现智慧乡村、智慧社区和智慧家庭，如图 5-14 所示。

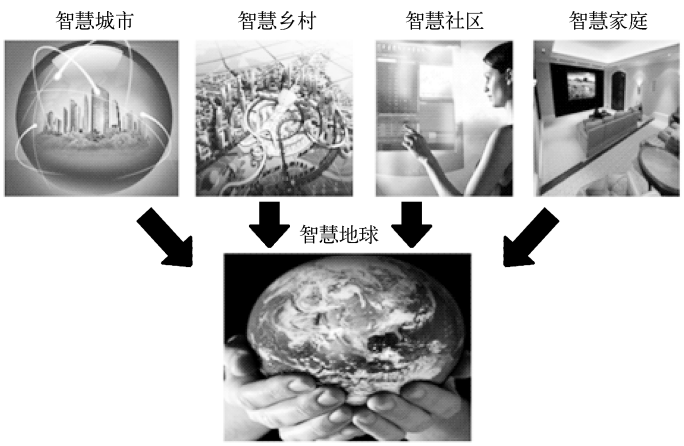


图 5-14　大数据相关的应用

使用大数据技术的目的是为了让我们的生活变得更智能化、更美好。IT 技术的终极目的就是为了实现智慧地球。其中移动计算技术是为了传递信息，使得人们获得大幅度的信息自由。而社交工具可以为客户提供方便快捷的服务，帮助企业开展全方位的营销。

5.2.1　相关生产厂商大数据技术简介

大数据技术相关厂商包括 IBM 公司、微软公司、EMC 公司和甲骨文公司等，如图 5-15 所示。

1. IBM 公司相关技术

IBM 公司提供的大数据服务主要包括：数据分析、文本分析、监测和各类商业服务。其中在一些大数据产品中，比较新的产品是 IBM InfoSphere BigInsights，它是基于开源的 Hadoop 技术，目的是从大量的数据中提取相关的信息。它为金融等行业制定了大数据的解决方案。IBM 公司一直致力于对大数据、信息流和结构化数据的研究。

在短短几年时间内，IBM 公司投入大量的资金进行并购和研究。例如，2009 年收购了

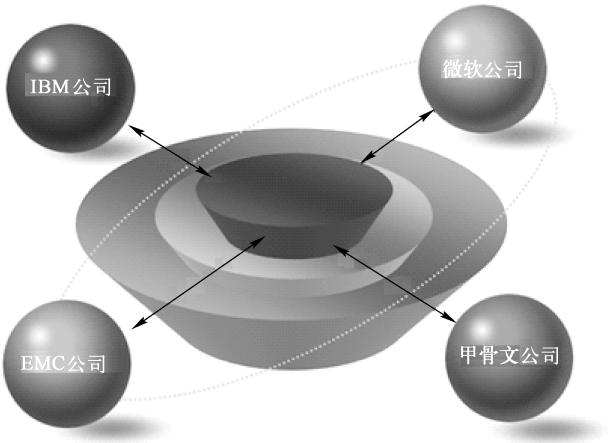


图 5-15　大数据技术相关厂商

数据分析和统计软件提供商 SPSS，2010 年收购了数据库分析供应商 Netezza 公司等。

下面分析一下该公司产品具有哪些特点，如图 5-16 所示。

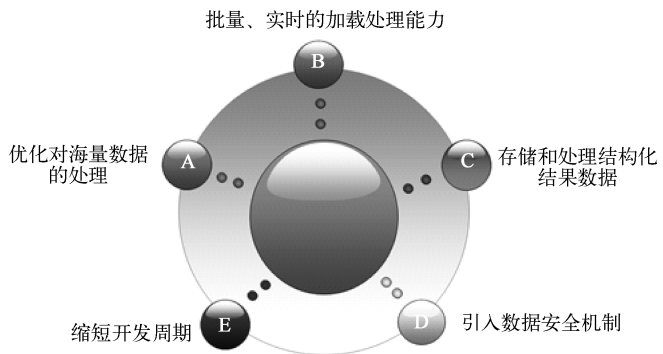


图 5-16　IBM 公司相关产品具有的特点

（1）优化对海量数据的处理

基于 Hadoop 技术，实现对海量数据的分析，包括对大数据的存储和分析。

（2）批量、实时的加载处理能力

基于 Hadoop 技术，实现对结构化数据、非结构化数据批量和实时地加载处理。

（3）存储和处理结构化结果数据

可以存储和处理结构化结果数据。其中内置的文本语义分析和预测组件可以实现对非结构化数据的结构化处理。

（4）引入数据安全机制

引入专业的数据安全机制，对数据进行有效的审计和保护，使得数据处理更加快速、可靠、安全和稳定。

（5）缩短开发周期

该产品可以让开发人员能够关注业务逻辑，而不是技术细节的实现，大大降低了开发的复杂性，缩短了开发周期，屏蔽了 MapReduce 的实现细节。

2. 微软公司相关技术

微软公司提供的 Windows HPC Server 2008 是一种基于 Windows Server 技术的高性能计算解决方案。同时微软公司也开发了并行处理技术，向 Windows HPC Server 的用户提供处理大数据的工具。特别是与惠普公司合作开发了一系列能够提升决策速度的设备。

3. EMC 公司相关技术

对于 EMC 公司，大数据解决方案涉及多达几十个产品。这些大数据解决方案可以有效使用来自不同数据源的数据，包括网页、监控系统和传感器的信息。

例如，EMC Greenplum 的设备，通过大规模并行处理（MPP）架构去解决大数据相关的问题。

4. 甲骨文公司相关技术

甲骨文公司为大数据提供了多种软硬件方案，同时在大数据的市场上提供了多种核心产品。例如，Oracle 大数据机与 Oracle Exadata 数据库云服务器、Oracle Exalogic 中间件云服务器一起组成了广泛和集成的产品系列。

甲骨文公司面向大数据的解决方案主要包括：数据的捕获、组织、分析和决策，如图 5-17 所示。

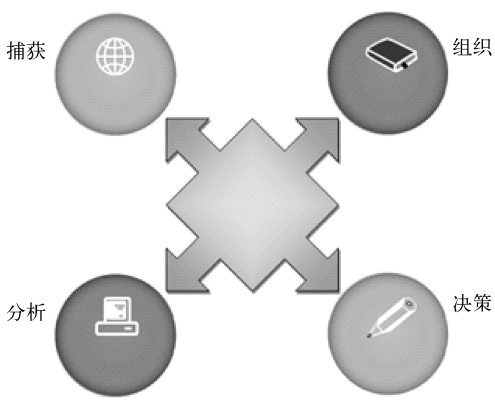


图 5-17　甲骨文公司面向大数据的解决方案

5.2.2　大数据与云计算

对于云计算来说，相当于提供一个快捷的海量数据处理的平台，它为大数据提供了访问、管理的渠道和场所。云计算本质上就是利用数据处理技术实现企业的各种业务模式。例如，企业的经营数据、银行的交易信息，互联网中的交互信息，以及物流行业中的商品及物流信息，都可以利用云计算技术进行存储、计算和访问。大数据和云计算等信息技术为非结构化数据管理提供了支撑，对于企业来说，决策者将脱离经验和直觉，更加倾向基于大数据分析做出决策。

举例来说，如果把商业智能转移到云计算平台上，可以在很大程度上提高商业智能的运行效率和数据分析能力。特别是金融行业，已经明确提出了“云 + 大数据”的战略，如图 5-18 所示。我们可以把云计算当做基础设施建设，而大数据作为资产，数据挖掘是实现价值的手段之一，预测分析是要达到的目的。

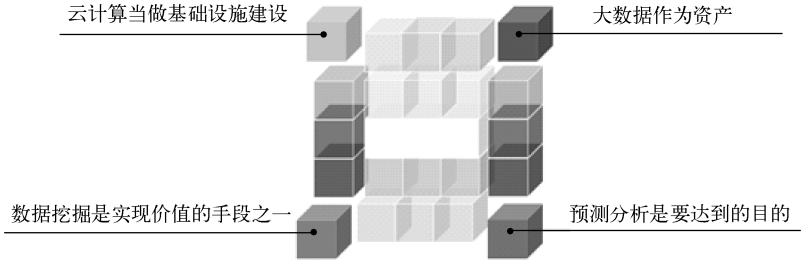


图 5-18 “云 + 大数据”的战略

从技术创新的角度来说，大数据的处理技术应该增强安全性、高可用性，包括对大数据的解密、加密、动态口令认证等内容。在业务创新上，应该提供更深层次的挖掘，有效地提升业务能力，为大数据提供广泛的管理平台。

大数据时代下的超大数据量，包括占到一半以上的半结构化和非结构化数据，已经远远超出了传统数据库的管理能力，大数据技术可以帮助人们存储和管理大量的数据。可以从低价值、高复杂度的数据中提取有用的价值，特别是相关的产品和技术不断涌现。从本质上来说，大数据也是数据，依然离不开对数据的存储、检索和管理，如挖掘分析等。我们可以利用大数据技术和云计算，改善和提高各个行业的经营模式。

关于大数据和云计算的处理技术，主要包括以下内容：

（1）大数据和云计算共同改变商业运营模式

大数据和云计算共同改变着企业的商业运营模式，在目前社会中，充斥着各种海量数据，如博客、微博、邮件、视频、音频、文档等非结构化数据，利用大数据和云计算技术将任务分布在资源池上，满足对大数据的计算和存储需求。

大数据和云计算的结合满足低成本硬件、软件的要求，同时能够处理各种类型的海量数据，正在悄悄改变着商业运营模式。

（2）关于大数据和云计算的存储和管理

云计算对关系型数据库产生了巨大的影响。它可以提高对海量数据的并行处理能力和实时分析能力，同时提供在线分析处理和在线事务处理的能力，也可以满足大数据环境下的业务需求。通过大数据技术和云计算的结合，除了降低建设大型数据仓库和软硬件设备的成本，也大大减轻了运营、运维和推广的压力。通过云计算和大数据技术进行海量数据的统计、分析、预测处理，可以促进传统商业智能系统的发展，快速适应商业模式的变化。

例如，云计算可以满足对海量数据的处理，能够处理 PB 级的数据量。同时可以简单部署，快速响应，减少磁盘 IO 时间，降低建设、运营成本，特别是大幅度地降低了硬件成本、软件成本和人力成本。

5.2.3 大数据和传统商业智能分析

大数据分析和传统商业智能分析在内容、分析方法和各种时效性要求上都有很大不同，传统数据仓库平台已经很难支持所有的分析应用，需要开发各种标准接口，支持 MPP 架构、内存计算和 Hadoop 技术等。只有构建混合型的大数据云平台，才能够支持传统的商业智能和大数据分析。

传统商业智能分析主要是面向内部的结构化数据，依赖数据仓库，以报表查询和挖掘分

析为主。大数据分析包含结构化、半结构化和非结构化的数据，一般数据量都在 TB 级以上，主要以挖掘分析、实时预测为主。

特别是主要的商业智能供应商都宣称对大数据技术的支持，或者在一些解决方案中使用了大数据技术，大数据可以作为传统数据库、数据仓库的扩展。它们是相互促进的关系，而不存在互相取代的问题。因此，为了满足未来商业智能的发展，应该将大数据技术和商业智能技术结合起来。

5.3 大数据的应用情况

近几十年，随着计算机技术的发展，信息已经积累到了一定程度，它比历史上任何一段时期充斥着的信息都多，而且数据的增长已经达到了前所未有的速度。对于中国企业来说，应该利用大数据，将传统模式转变成以数据服务为核心的商业模式。

大数据在现代社会应用非常广泛。例如，在电子商务中，每天可以访问 1 亿次，每年可以由 10 亿人访问，并且进行网络交易。对于保险业来说，可以进行大量的图片上传工作和索赔分析工作，每天可以有 100 万次。例如，可以为客户提供在线透明分析，对于 2000 万辆汽车来说，每天大约有 10 亿条同步记录。对于医疗卫生业来说，每天可以有 2000 万次的监视。

大数据的应用还包括很多方面，如数据的可视化技术，可以更清晰和准确地展示多维数据，反映趋势变化等。同时可以提供更快、更便宜的预测分析。

大数据应用的行业如图 5-19 所示，主要包括金融服务业、数据媒体、交通运输、司法执法和零售等行业。

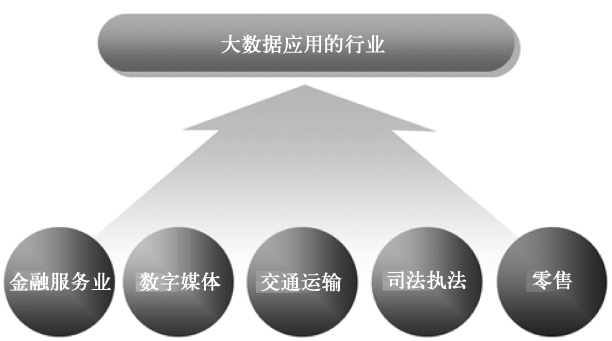


图 5-19　大数据应用的行业

对于金融服务业来说，大数据的应用主要包括：金融欺诈检测、反洗钱等。可以跨多个实时信息流关联复杂的风险分析，并且得到实时响应，每天可以增加 10TB 的数据，甚至更多。还可以全方位分析客户视图。

下面分析一下随着互联网金融时代的到来，对商业银行造成了哪些冲击。主要表现在以下几个方面：

1）传统的存贷款业务受到很大的压力，因为互联网金融有着强大的技术创新能力，导致金融脱媒的现象越来越严重。

2）长期以来，传统的商业银行一直依赖利息，创新的动力不足，同时机构冗余，变革缓慢。

3）互联网金融已经成为我国金融服务的有效补充，在一定程度上可以解决了中小企业融资困难的问题。

4）互联网金融可以通过社交网络和电子商务平台挖掘与金融相关的各种信息，满足用户的需求，同时对客户的服务更具有针对性。

但是对于商业银行来说，同样具有自己的优势，例如：

1）商业银行在金融领域中长期处于领先的地位，已经建立起自己的品牌，获得了客户的信任。

2）商业银行具有专门的监管机构，例如银监会体系，它同时具有成熟的风险管控体系。

3）商业银行正在努力提升网银和电子银行的客户满意度和交易活跃度，同时提供了与P2P不同的差异化服务，利用长期建立起来的品牌和信用去吸引投资者和融资者。

可以这样说，基于大数据的应用，对未来金融行业的发展将会起到关键性的作用。

同时，对于其他行业来说，大数据技术也会促进其不断发展，见表5-2。

表5-2　大数据同样可以促进其他行业的发展

名　　称	属　　性
数字媒体	实时广告定位、精准广告投放、属性分析
零售	全渠道营销、实时促销
司法执法	多点监测、网络安全检测
交通运输	物流优化、缓解交通拥堵

目前来说，很多IT企业都在积极推出大数据相关的产品和方案。

1）IT企业根据客户的实际需求来进行商品推荐，根据客户购买商品的历史记录，推荐其偏好的相关产品，或者根据用户的浏览历史，推荐符合用户喜好的商品等。

2）如何挽留客户，更好地为客户提供服务，数据起到了重要的作用。可以对客户进行分类，针对不同的客户群体，制定不同的营销策略。例如，向新注册用户发送一些优惠券；向老客户发送一些折扣信息等。

5.3.1　大数据在金融行业的应用

“大数据”的特征为：数据量大，数据种类繁多，数据的增长速度加快，数据来源的多样性。在大数据时代，关于大数据的挖掘工作迅速增加，它的数据来源更加广泛，可以通过数据交换、整合发现市场的趋势，让企业或者商业银行发现商机，创造新的价值。同时可以使用仿真和复杂的计算，在计算速度极快的条件下完成工作任务。当然，在大数据时代下，我们面临的主要问题是数据的真实性，因此，需要大量的数据模型去分析，以保证数据的准确性。

在几十年前，商业银行使用传统的核算记录各类数据，而在目前，商业银行是以计算机、各种电子化设备采集数据，因此形成了目前的海量数据。

对于以前的银行数据，因为过于分散，源头单一，无法表现客户的交易行为，以及客户的喜好和消费习惯等特征。因此，银行很难了解客户对于产品和服务的满意程度，无法从根本上弥补信息的不对称性。同时，商业银行拥有大量的客户数据，可以通过数据分析获得很

多信息，但是因为信息的不全面性，可能在管理和营销上得到错误的结论。

例如，某位信用卡用户月均刷卡 10 次，月均刷卡 300 元，每年平均拨打 5 次客服电话，但是从未投诉。那么按照这些信息，该客户是一名满意度较高、流失率很低的客户。但是真实情况是：该客户多次打客服电话都没有接通，客户多次在微博和博客上进行抱怨还款不方便，客户服务不好，可以看出该客户的流失风险很高，如图 5-20 所示。

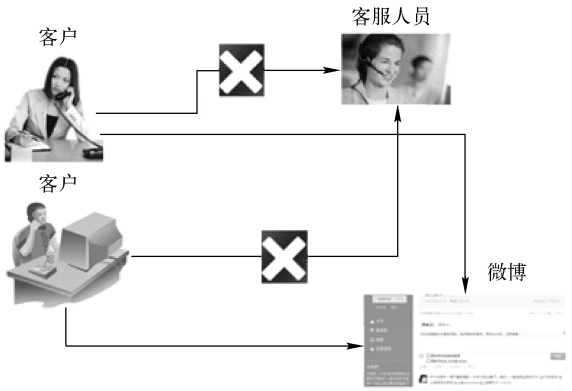


图 5-20　信用卡客户示例

在大数据时代，商业银行面临的压力不仅来自于同行业之间的竞争，同时来自于外部挑战的压力也越来越大，特别是在互联网和电子商务等企业中，它们的产品创新能力和大数据应用能力明显超过商业银行，同时这些企业也在涉足金融领域，改变着人们的金融消费模式，银行将在以后的发展过程中，承受着巨大的压力。

举例来说，阿里小额贷款公司可以根据人们的信誉度发放贷款，不需要提供担保。其具体做法是根据其电商平台、淘宝网和支付宝等信息数据，依赖大数据分析技术，判定哪些个人和企业可以发放贷款，贷款额度是多少等。同时也可以使用大数据算法找出竞争对手产品价格的变化，从而改善自己的价格以保持竞争力。

所以说，大数据已经改变了我们的生活模式，提供了产品创新的新思路。网民和消费者的区别正在模糊，数据成为核心的资产。在大数据时代，如何能够利用大数据技术，深刻理解消费者的需求，做出预测和判断，是企业和银行需要考虑的问题。

大数据在金融行业的应用除了行业分析、风险评估外，还可以了解各系统的交易情况、分析客户行为特征。

1）通过各种网络渠道，及时获取各种与商业银行相关的事件，针对网上的各种信息及时反应。

2）通过社会渠道，获得了解客户对商业银行的评价反应，及时调整和优化，维护商业银行的形象。通过获取网上信息，及时了解行业动态，为存、贷款工作提供数据支持。

3）通过客户网站及其他客户披露的数据，及时获取客户的信息。

4）通过新闻媒体、社会化网络，及时获取与客户相关的事件，获取营销机会，规避风险等。

5）通过各种社交网络（微博、微信、博客、社区等），获知客户感兴趣的热点话题，了解客户行为，通过关注客户的网络行为，获取销售信息。

6）对于商业银行来说，可以利用大数据技术分析宏观的经济变化，寻找信用优质的小微企业等内容。

1. 在大数据时代，大数据的应用给金融行业带来了哪些挑战？

1）金融同行业的竞争开始加剧，同时金融脱媒产生了很多新型业态，它们共同参与到金融市场的竞争中。很多金融机构都在向综合经营方向发展，商业银行也纷纷发行各自的金融产品和理财产品。

很多第三方支付公司通过对各类产品的创新，替代了大量的银行支付业务，逐步吞食银行支付结算的市场份额。

2）很多商业银行都把电子银行业务当做重要的交易渠道，它具有低成本、高效率的特点，大大减轻了银行柜面的压力。随着大数据时代的来临，要求对商业银行的电子渠道进行创新，保证商业银行以电子渠道为基础，逐渐扩大交易渠道，制定个性化和综合性的银行产品。

3）在大数据时代，商业银行传统的业务价值观被削弱，要求银行可以提供个性化的金融服务和解决方案，提高客户对产品和服务的认同度。目前来说，可以通过收集客户的社交网络信息，分析客户的购买力和偏好，提高商业银行的利润率。

4）在大数据时代，很多互联网企业从网络购物和供应链服务转向属于传统银行业务的支付、清算等领域，对商业银行的传统地位造成挑战。商业银行可以通过全场景的金融解决方案，为客户提供资金流，整合银行的资源，提高利润率。

5）在大数据时代，商业银行的盈利模式有很多，例如可以通过银行的业务赚取中小企业的利息收入和大型企业的中间业务收入。商业银行可以依赖数据服务能力，为客户提供电子商务解决方案和财富管理服务。

6）在大数据时代，商业银行可以充分利用业务数据和社交网络数据。通过集中、整合、挖掘和共享发挥数据的价值，提高风险管控能力。提高商业银行的整体管理水平。

我们总结来说，金融行业普遍存在以下问题：数据丰富，但是知识贫乏；创新动力不足；IT 观念落后；人才匮乏，如图 5–21 所示。

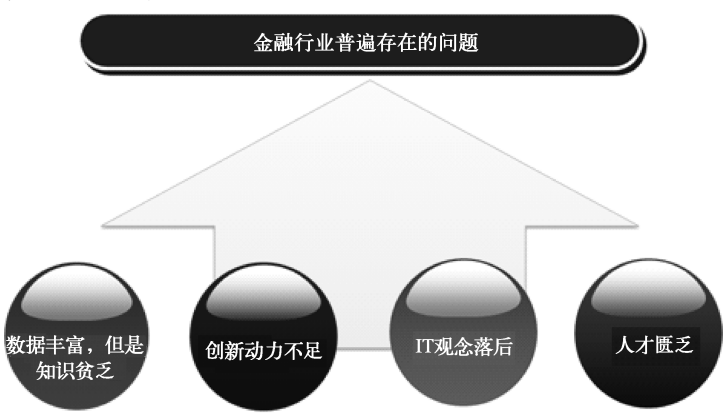


图 5–21　金融行业普遍存在的问题

2. 金融行业应该重视大数据应用的哪些问题呢？

首先应该加强对数据的整合工作，改进数据的处理架构，保障数据的安全体系，完善数据的运维体系，最后加强对专业化技术团队的建设，如图 5–22 所示。

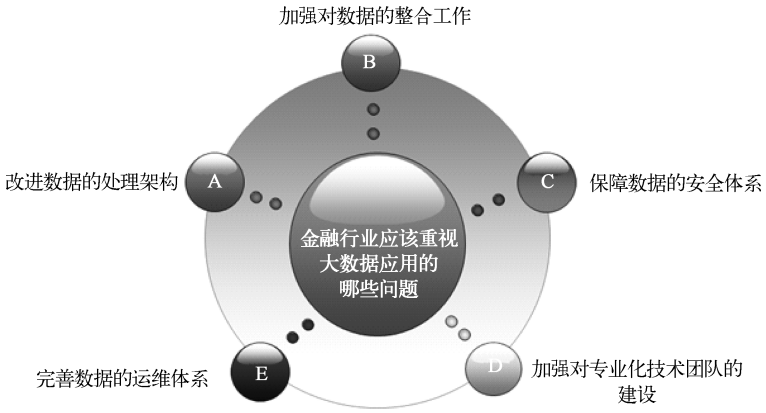


图 5-22　金融行业应该重视大数据应用的哪些问题

大数据的应用还可以作为银行创新的催化剂，引导银行对业务模式的变革，推动商业银行在经营理念、组织架构、业务流程上进行全面调整，不断增强核心竞争力，提升运营效率。大数据为商业银行提供了重要的战略发展契机。“大数据”对于银行的作用主要表现在以下几个方面，如图 5-23 所示。

图 5-23　“大数据”对于银行的作用

（1）对客户的消费趋势进行预测

商业银行可以利用大数据技术对客户的消费趋势进行预测，同时增强对客户的细分和市场趋势的分析力度。例如，我们可以基于人口统计特征，通过查询客服、银行柜员的记录，以及各种网站的点击流和客户的支付历史等信息，对客户行为进行洞察。

（2）对风险和欺诈进行洞察

利用大数据技术，商业银行可以对风险和欺诈进行洞察。例如，可以利用财务风险分析、贷款风险评估、实时欺诈检测等手段。通过各种社交媒体、市场新闻，获取对银行客户和潜在客户的洞察，以提高对各种风险的预测水平。

（3）评估商业银行的服务质量和客户满意度

利用大数据技术，可以评估商业银行的服务质量和客户满意度。例如，通过与客户的会

谈、录音等各种交互记录，识别客户的问题，以此提高服务的质量和客户的满意度。

（4）开展精准营销

商业银行可以利用大数据技术开展精准营销以提高利润，降低成本。同时扩展了营销的手段，从网点坐售、电话营销扩展到短信、微博和微信等平台，如图 5-24 所示。

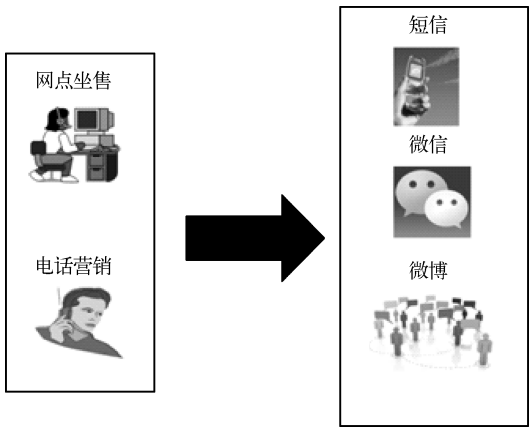


图 5-24　营销的手段

（5）提高商业银行的管理水平

利用大数据技术，提高商业银行的管理水平。实现“以数据说话”，为银行的市场营销、资产负债管理、客户关系管理等方面提供决策支持。

（6）拓宽商业银行的业务领域，加速产品的创新

利用大数据技术，可以拓宽商业银行的业务领域，加速产品的创新。例如，社交媒体为商业银行创造了新的客户接触渠道，从银行网点、ATM 等固定设备扩展到移动终端设备，甚至扩展到微博、微信等社交网络。渠道的创新也引起对银行支付模式的创新，从传统支付、电子支付和第三方支付过渡到移动支付上来，如图 5-25 所示。

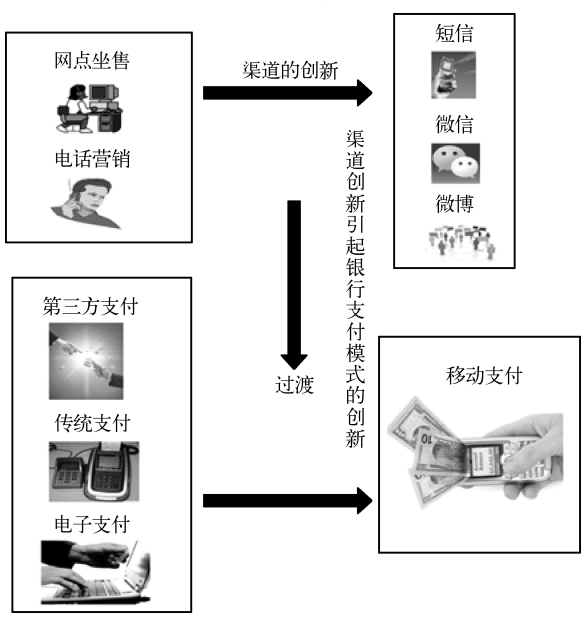


图 5-25　拓宽商业银行的业务领域

当客户与银行发生交易的时候，会产生大量的数据，这些数据为银行进行有针对性的营销创造了机会。因为数据隐含着大量的信息，所以我们最主要的工作就是将这些信息挖掘出来，并且加以利用。

在大部分的应用中，随着数据量的指数级增长，特别是一些非结构化数据的快速增长，这些海量的数据会导致数据分析的时间延长，传统的商业智能发展会出现“瓶颈”，而在大数据时代，这些问题会成为缺乏为客户创造价值的动因。

在很长的一段时间内，银行的多数应用都是建立在客户与银行的交易过程中，例如银行开户、存款和取款等业务。要深入理解客户的需求，更好地为客户服务，仅仅依赖这些交易数据是远远不够的。随之社会的发展和科技的进步，银行可以通过多种途径收集客户的信息，例如在一些移动终端上收集客户的位置信息，然后进行有针对性的营销。在大数据时代，这些非结构化的数据量远远超过传统的结构化数据量。

举例来说，某银行客户进入一个购物广场，在某超市里面进行了一笔 120 元的消费，客户信息是：30 岁，女性，有一个孩子。这时该女士会收到一条短信，提示她刚进行了一笔 120 元的消费，可以在某儿童商店享受 5 折优惠一次，于是该女士很有可能会给孩子买一套衣服或者一双鞋子。该流程就实现了大数据的秒级营销，如图 5-26 所示。

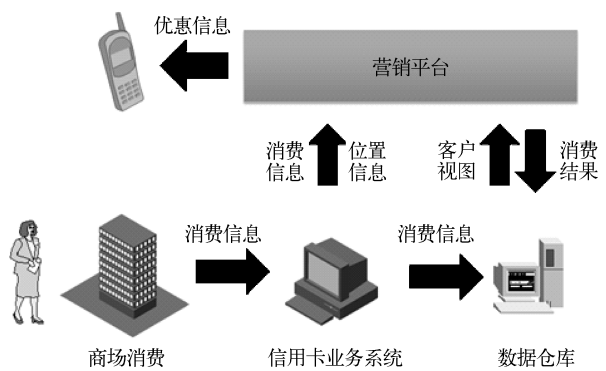


图 5-26　大数据的秒级营销

随着互联网行业的发展，客户可以通过互联网或者其他电子渠道去发表自己的一些看法，甚至是购买商品，这些动作都会为商业银行收集客户的信息创造了条件，降低了信息的不对称性。也就是说，在以前，客户对银行的情况可以有多种渠道去深入了解，但是银行却很难深入了解客户的需求、真实想法和自身的资金实力。

目前来说，很多商业银行可以收集客户在互联网上的一些言论、微博发表的信息和购买商品的信息，然后去分析客户最喜欢的服务和产品，包括客户自身的信用信息和资金实力等内容，从而正确理解客户，统计和分析出一些商机，有针对性地进行精准营销，并且更好地提供服务。这也为商业银行实现从“以业务为中心”向“以客户为中心”的转变提供了条件。

对数据的分析逐渐成为银行实现核心业务价值的重要手段之一，特别是在利率市场化阶段，会出现存款的稳定性降低和存贷款利差普遍收窄的情况。金融脱媒，导致大量客户流失和客户的忠诚度降低。银行如何为客户提供个性化的服务已经成了迫在眉睫的课题之一。因此，银行需要进一步提升数据分析的能力，提高对业务的洞察力。

目前一些商业银行的数据量已经达到了几十 TB 以上，特别是非结构化数据的快速增

长，这种指数级的增长，对数据分析的能力提出了挑战。特别是“金融脱媒”现象越发明显，银行作为“支付中介”的垄断地位已经动摇，同时客户对银行服务的要求越来越高。银行业这个长期以来一直变化缓慢的行业现在应该放下“架子”，及时且更加全面深入了解客户的基本信息和属性，对客户进行精准营销，提升业务运行效率，逐步提升客户体验。

举例来说，商业银行可以基于大数据的分析和查询，特别是收集客户的地理环境、年龄和交易喜好信息，有针对性地为客户提供理财产品建议和提醒，同时通过对大数据的分析和挖掘，评估客户的信用风险和资金偿还能力，降低银行的各种风险，如图 5-27 所示。

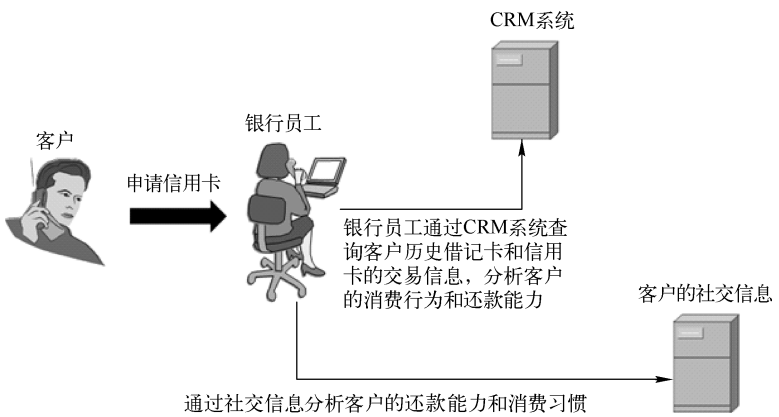


图 5-27　降低银行的各种风险

总结：大数据分析可以实现从“以业务为中心”向“以客户为中心”的转变，降低了信息的不对称性。

3. 大数据在金融行业的主要应用

应用方式如图 5-28 所示。

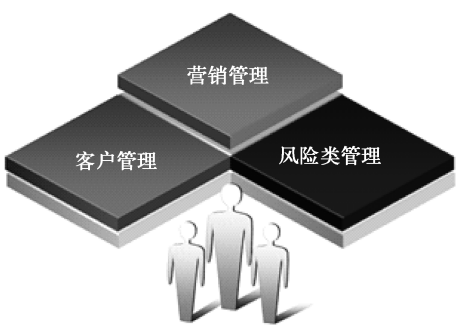


图 5-28　大数据的主要应用

（1）客户管理

可以构建客户的全方位分析，见表 5-3。

表 5-3　构建客户的全方位分析

客户维度名称	基 本 属 性
客户基本信息	客户名称、证件类型、证件号码
客户资产信息	与客户资产相关的信息

（续）

客户维度名称	基 本 属 性
客户风险信息	与客户相关的信用评级信息等内容
客户财务信息	客户产生的利润等内容
客户事件信息	例如提前还款、逾期等信息
客户联系信息	客户主要联系信息，包括家庭地址、电话等
客户产品信息	包括存款类、贷款类等信息
客户关系信息	客户经理与客户之间的关系
客户信用评级	客户信用卡申请资料、客户的信用风险等级

其中在客户信用评级中，银行可以通过收集客户信用卡申请资料，分析客户的信用风险等级，帮助银行业务人员做出决策。特别是国外的银行机构，需要给客户多高的利率，是根据业务人员的分析决策决定的，客户的信用评级是一个重要参考。

在客户风险信息中，银行可以收集客户的基本信息、地理环境、年龄、交易信息和各种信用信息，对这些海量数据进行分析和挖掘，评估客户的信用风险和资金偿还能力，降低银行的各种风险。

商业银行以大数据为应用，借鉴行业先进模型，建立标准体系，保证数据的唯一性、完整性和共享性，同时商业银行也应该制定加强对客户数据的安全保护策略。

（2）营销管理

传统营销一般采用一对多的方式，这种针对群体性的营销，成本较高，同时准确性很差。应该引入大数据的概念，实现有针对性的智能营销，如图 5-29 所示。

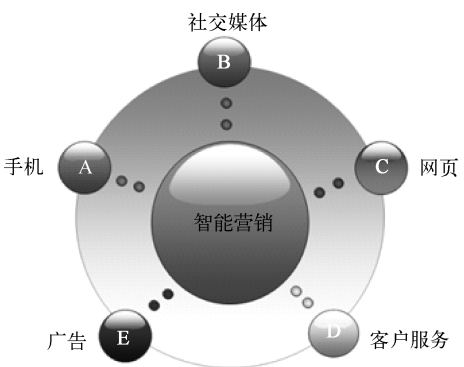


图 5-29　有针对性的智能营销

对于智能营销管理中的舆情分析来说，主要包括银行声誉分析、银行品牌分析、银行服务质量分析、竞争产品分析、产品评价分析等。主要是跟踪社交媒体的评论，了解影响客户的关键性问题，产生潜在的客户流失预警和满足客户服务的需要。也可以长期跟踪新闻热点，包括对正负面报道的分析，以提供个性化的市场分析结果。

对于客户与市场洞察方面，主要包括银行对市场的趋势分析。从社交媒体、市场新闻信息中提取信息，方便对市场的洞察。

对于运营洞察与优化，主要包括系统的数据保存与管理、系统日志维护和系统故障分析。对于数据保存与管理来说，是通过大数据平台对各种历史报表和分析数据进行保存和管理工作。对于系统日志维护来说，是为了实现更多的历史数据保存和更好的分析能力。对于系统故障分析来说，主要目的是为了对系统的故障进行预测与分析，从而更好地提升系统的运营效率。

（3）风险类管理

通过大数据技术，可以实现准确、高效的风险控制，基于历史数据和实时数据，实现欺诈监测。对于风险与欺诈洞察，主要包括财务风险分析、市场与组合风险分析、贷款风险评估分析、反洗钱与欺诈调查 、实时欺诈检测和市场监督等内容，如图 5-30 所示。

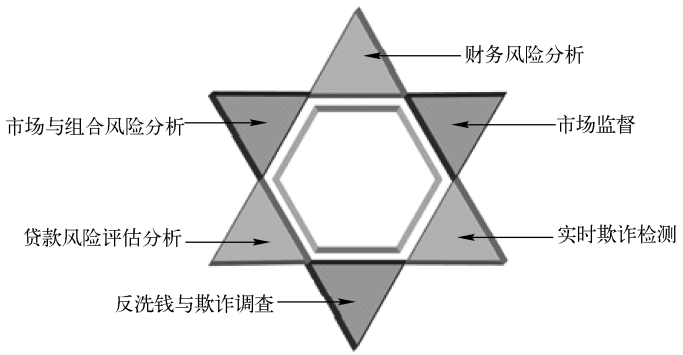


图 5-30　风险类管理

1）财务风险分析是通过评估信用风险和市场风险所产生的详细数据进行分析，目的是为了符合监管的需要。

2）市场与组合风险分析是通过大量的历史市场数据和交易数据，实现更多的实时预测风险分析。

3）贷款风险评估分析是从媒体或者社会公共信息中提取企业客户和潜在客户的信息，以提高风险预测能力和预警能力。

4）反洗钱与欺诈调查是提取犯罪记录信息、法律数据等内容进行欺诈调查的分析。

5）实时欺诈检测是通过大量的欺诈数据进行分析。

6）市场监督是通过实时交易监控实现对市场的监督作用。

大数据在金融行业未来的应用方向

大数据在金融行业未来的应用可能会很多，如图 5-31 所示。

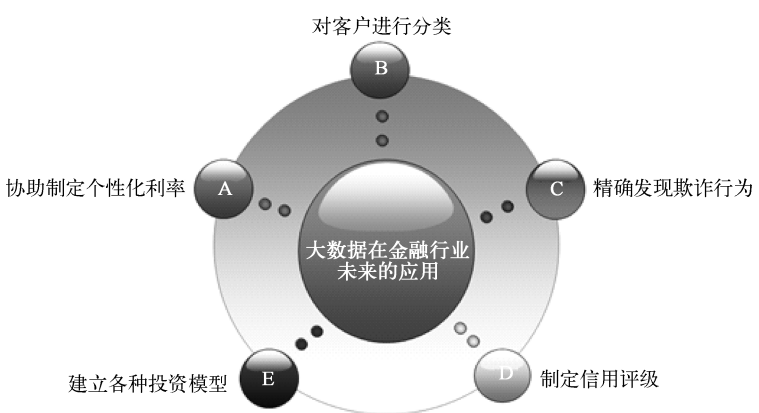


图 5-31　大数据在金融行业未来的应用方向

在大数据时代，商业银行需要做好哪些工作呢？

1）在日常运营过程中，商业银行应该加强对数据的管控和数据处理。其中，数据管控应该参考标准，保障数据采集的准确性和数据应用的可视化。尽量降低银行的声誉风险。

2）商业银行应该提高对大数据应用的支持力度，同时实现资源利用的最优化。

3）商业银行应该重视对大数据技术人才的培养和储备。

“大数据时代”将会带动整个社会交易模式的变化，未来更多的客户服务将在互联网中进行，特别是对于商业银行来说，更应该注重挖掘相关的社交媒体信息，拓展获取客户信息的渠道，使之成为银行经营的有用工具，能够为客户提供更好的服务。

5.3.2 大数据在其他行业的应用

大数据在企业的应用主要表现在以下几个方面：

1）客户全方位视图，以增强企业对客户的了解。

2）进行可预测的运维分析。

3）通过大数据技术找出新的业务模式。

4）实时风险评估，降低风险管理成本。

总结来说，大数据在企业的应用主要体现在客户全方位视图、运维分析、找出新的业务模式和降低风险管理成本等方面如图 5-32 所示。

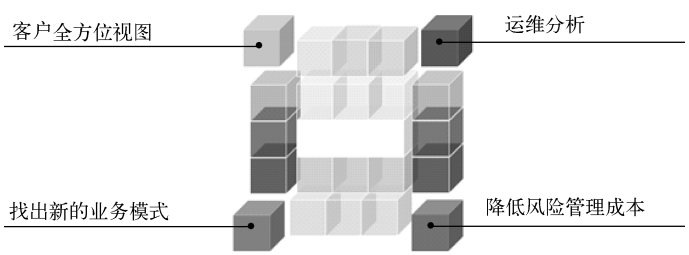


图 5-32　大数据在企业的应用

大数据除了在金融行业的应用外，在其他行业里面有哪些应用呢？

（1）电力行业

电力行业可以利用大数据技术平台分析和预测电力维修、产能和故障原因等。

（2）医疗行业

在医疗行业中，医院可以通过对大数据的应用，对远程病人进行监控，尽量做到预防保健，从而有效地降低病人的住院率。大数据在医疗行业的应用主要是分析全部的数据，而不单纯是样本数据，分析数据的目的是以预防和预测疾病为主。

例如，对传染病的传播趋势进行预测，为相关卫生机构提供快捷和近似的流行病预测。大数据技术可以支持区域卫生医疗，临床决策支持，建立全民健康档案，药物研发，健康结果分析等。同时还可以利用大数据技术对病人进行实时监控，提前发现病人的危险情况。也可以实现电子病历、诊疗移动化、智慧医院等。

如图 5-33 所示，健康中心利用健康管理门户网站对每个家庭实行健康监控，同时提供各种远程服务，将重要的信息传送给医院。医院根据这些信息将诊断结果再传送给健康中心，健康中心依据这些诊断结果对每个家庭提出健康意见。以上过程正好形成了一个闭环。

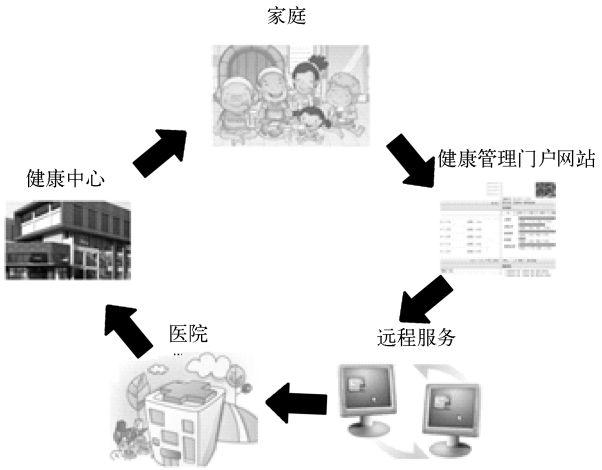


图 5-33　医疗行业大数据应用

（3）电商行业

电商行业主要关注 4 个方面的内容：东西卖给谁？去哪里找客户？卖给客户什么东西？怎么卖？如图 5-34 所示。

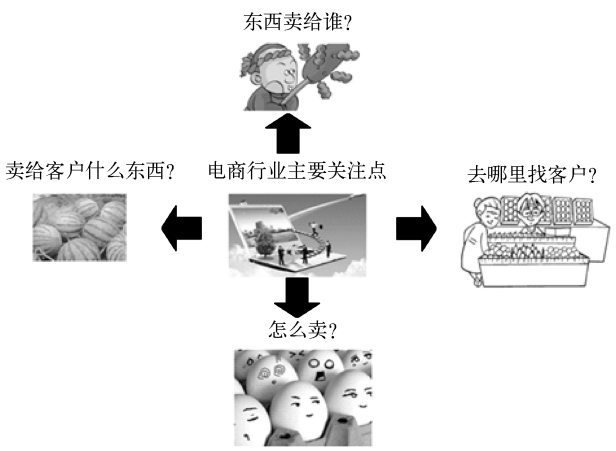


图 5-34　电商行业主要关注 4 个方面的内容

电商行业使用大数据技术的目的是让数据分析替代直觉。通过对数据的分析得到信息和知识的反馈。

举例来说，对于有很多休闲时间的老人来说，他们非常喜欢安全、舒适的按摩器材。对于按摩器材厂商来说，就解决了“东西卖给谁”的问题。

通过大数据平台，对用户的行为进行预测，这就解决了“怎么卖”的问题。而对于某种商品有特殊需求的客户，他们往往更看重商品的质量和品质，其次才会考虑价格的因素，这就解决了“卖给客户什么东西”的问题。

很多女性喜欢母婴类物品购物网站，很多电商可以为这些客户推送广告，这就解决了“去哪里找客户”的问题。

（4）交通行业

大数据平台主要分析交通状态信息、地理信息、警力分布信息、交通信息控制、车辆检测记录、查询统计、实时交通信息采集、交通流实时信息、交通流量统计等。其中交通指挥和调度包括各种的交通信息服务、短信提示、车载导航信息、热线、交通基础信息服务、动态交通信息服务等。

如图 5-35 所示，交通行业就是利用大数据的技术，通过收集交通基础设施数据、实时交通检测数据、GPS 汽车定位数据，进行数据的整合、分类，并加载到数据仓库中。然后，在此基础上，进行数据分析，并将分析结果再传送给交通指挥及调度系统。

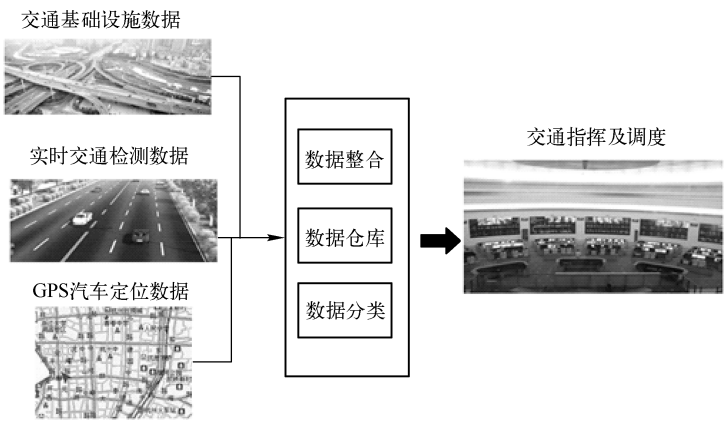


图 5-35　交通行业

（5）智慧城市

智慧城市主要包括智能城市交流、应急指挥系统、区域医疗系统、教育信息化。涉及的大数据技术包括云计算技术、物联网技术和信息安全技术。通过整合城市的信息资源，建设城市的劳动社会保险、电子商务、电子政务，使得城市更加智能化。

建设智慧城市的难点是：信息孤岛严重、缺乏有效的管理。这样会导致重复建设严重，缺乏安全、完整和科学的城市建设体系。

建设智慧城市的原则是：创新、高效服务、宜居、便利、健康、绿色、安全、智能和信息共享。具体的内容包括市民管理服务、社会保险、交通、医疗、公共管理、企业管理、行政审批、纳税、企业年检、就业和城市物流等方面。

（6）其他领域

其他行业，例如国防安全，可以利用大数据技术进行情报分析、舆情分析等。对于证券业，可以支持对异常行为的监测功能，同时支持商业决策。对于电信行业，还可以进行网络监控分析、客户流失率分析等。

大数据在其他行业的应用很多，还包括智慧乡村、智慧小区、数据化城市管理、情感分析、社交 CRM/网络分析、社交媒体分析、价格优化分析、客户行为分析、影响力分析等。

小结

- 据 IBM 公司预测，到 2020 年，全世界产生的数据规模将达到目前数据量的 44 倍，在

这些数据中，只有1% ~5%的数据是结构化数据，这意味着非结构化数据和半结构化数据将占据绝大部分。

- 大数据是指巨量的信息，规模巨大，已经无法用常规的软件工具在短时间内进行存储和管理。大数据的主要功能就是预测，可以将算法应用到海量的数据中，预测事件发生的可能性。但是我们不要拘泥于大数据的概念。
- 大数据在金融、互联网的应用非常广泛，这些企业在日常运营过程中产生了大量的数据，尤其在人口众多的国家，大数据的应用更为广泛，通过这种挖掘和利用大数据的能力，可以大大提高服务的水平。
- 国内大数据应用的基本现状都较为复杂，目的是为了追求大数据技术而进行各种大数据项目的建设，这样可能会导致很多企业“掉进”以技术为导向的误区。大数据的项目必须有明确的业务需求，用商业思维来推动大数据的建设，只有这样，大数据的价值才能充分体现出来。
- 在大数据时代，我们面临哪些挑战：

1）企业或者银行将数据的重要性提升一个层次。
2）大数据管理上的成本大大提高。
3）产品创新不足。
4）数据整合和数据质量管理的难度很大。
5）一些企业和银行在数据利用上有一定的局限性。
6）应用与理论研究的成本很高。
7）业务需求和技术之间的协调。
8）人才方面储备不足。

- 对于中国企业来说，大数据技术的研发和投入相对较少，目前很多企业没有利用好大数据。大数据的发展对于我们的启示是：必须把握好大数据技术，推进企业的转型创新。同时需要企业制定新的大数据人才战略，以价值体系激励员工。培养洞察分析的能力，以个性化服务去赢得客户。
- 对于商业银行来说，为了保证在金融市场的竞争地位，将数据转化为可以洞察的信息和知识，推动业务的发展，提升管理的效率。同时随着移动终端技术的发展和应用，已经改变了客户的消费模式。如果从数据的角度来看，我们其实已经进入到了大数据时代。
- 虽然目前大数据没有明确的定义，但是我们每天都在产生海量的数据，数据将我们“包围”起来，我们正在进入到“大数据时代”。根据Gartner的定义，大数据的特征具体涵盖了称为4 V的内容：数据量大（Volume）、数据多样化（Variety）、实时性强（Velocity）、商业价值（Value）。
- 我们总结来说，大数据的定义就是通过快速采集、挖掘和分析，从大数据量多样化的数据中获取价值。形象地说，大数据就是沙里淘金的过程。
- 对于大数据来说，有结构化数据、半结构化数据和非结构化数据三种类型。
- 大数据分析平台主要包含：大数据基础平台、平台组织团队、数据管控和应用系统等。
- 大数据对于系统的需求涵盖了“三高一低”：高性能、高存储、高扩展和低延迟。

- 对于云计算来说，相当于提供一个快捷的海量数据的平台，它为数据提供了访问、管理的渠道和场所，它本质上就是利用数据处理技术实现各种业务模式。
- 在大数据时代，有一些代表性的例子，例如银行可以根据对客户的更深入了解，提供有个性化的服务。还可以进行相关的热点分析、犯罪行为分析、多渠道的客户分析，天气预测告警分析、交通拥堵预测分析等。
- 近几十年，随着计算机技术的发展，信息已经积累到了一定程度，它比历史上任何一段时期充斥着的信息都多，而且数据的增长已经达到了前所未有的速度。对于中国企业来说，应该利用大数据，将传统模式转变成以数据服务为核心的商业模式。

第 6 章　数据治理体系

本章目标

通过前几章的学习，我们已经理解了数据架构的基本知识和相关案例，同时了解了大数据的架构实践。为了提升数据架构各个层次的管控及其协作能力，我们同样需要理解数据治理方面的知识。

在本章中，我们将重点学习数据治理方面的知识，包括数据治理的概念、数据治理建设的关键要素和成功手段、数据治理建设的意义和必要性、数据标准的定义、数据标准项目总体规划和设计、数据质量管理总体规划、数据质量管理的解决办法、元数据管理的设计方法和数据生命周期的设计方法等内容。

学习本章后，读者将掌握：

- 当前企业和商业银行的总体现状和面临的问题
- 关于相关问题的改进措施
- 数据治理的概念
- 数据治理体系框架
- 数据治理建设的关键要素和成功手段
- 数据治理建设的意义和必要性
- 数据标准的定义
- 数据标准的分类和应用价值
- 数据标准体系框架
- 如何推进数据标准建设的实施
- 数据标准项目总体规划和设计
- 数据标准规划方法
- 数据标准实施优先级
- 数据质量管理的概况
- 数据质量管理总体规划
- 数据质量管理的解决办法
- 数据质量管理的执行
- 元数据管理概况
- 元数据管理的设计方法和流程
- 数据生命周期概况
- 数据生命周期的设计方法和流程

6.1 数据治理体系概述

6.1.1 当前企业和商业银行的总体现状和面临的问题

数据是企业的原始材料，也是金融、电信、互联网等行业最大的价值来源之一，如何利用这些数据，以及如何更好地对数据进行挖掘，已经成为提高企业竞争力最重要的手段之一。

1. 当前企业和商业银行的总体现状

目前来说，很多企业和商业银行都处于数据治理的初级阶段，很多系统的数据仍然面临着各种问题，例如数据不一致、不完整，数据质量较差，甚至不同的系统之间采用的数据标准规则都不一致，这样都会导致数据共享成本的上升和数据清洗工作量大大增加。如果缺乏对这些数据的有效管理，不仅会造成数据的价值和潜力不可能被挖掘出来，同时也会严重影响企业的利益和决策。对于这些问题，表面上是数据的问题，但是更深层次的原因是对数据管理的缺失或者相关制度不健全，以及人员的职责划分不清晰。

举例来说，对于数据管理缺失的问题，为了保障系统能够采集到完整、真实和有效的数据，在进行系统建设的时候，必须通过数据标准给予规划和约束。对于令人头疼的数据质量问题，它的改进也是一个长期的过程，除了使用技术手段保障数据的质量外，还可以通过对数据的管理来保证数据质量问题的快速解决。很多商业银行建立数据质量管理体系和数据治理机制，通过对数据质量问题的预防、识别、分析和监控等活动，满足数据质量管理的要求。

2. 企业和商业银行面临的问题

对于多数企业的系统建设，总会暴露出一些弱点和缺陷，例如系统多、数据标准不一致、很多数据难以共享等问题，这对核心业务系统的运行效率有很大的影响。所以对于大多数企业来说，应该着眼于长期的数据治理，挖掘数据的潜力，为企业增加业务价值。

对此，我们应该考虑如何对这些问题进行解决。

6.1.2 关于相关问题的解决办法

关于上述问题，我们有以下几种解决办法，如图 6-1 所示。

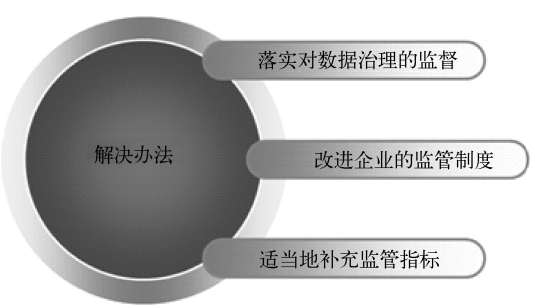


图 6-1　关于上述问题的解决办法

（1）落实对数据治理的监督

从战略角度来说，对数据治理的监管有利于实现企业的科学管理和可持续发展，例如可以将数据质量管理纳入到企业的规章制度中，建立数据质量管理的相关政策、流程、人员角色和职责，确立数据质量管理的目标，保障相关的管理部门和人员对数据质量管理进行有效评估和检查，同时落实数据质量责任制。

（2）改进企业的监管制度

将监管内容细化到业务流程的每一步，建立有效的激励和惩罚制度，并且按照各个环节的职责要求，保障相关人员能够履行职责。举例如下：

通过建立统一的数据字典，确保客户、产品和机构等基础信息的名称、定义、来源的一致性。各个系统之间可以建立统一的数据标准，规范数据名称和定义，然后在此基础上，逐步健全数据仓库，实现数据的标准化和规范化。同时保障监管标准的本地化，贴近监管的实际情况，做好监管数据治理的顶层设计，从而引导企业的高层领导从战略高度认识数据治理对于企业的管理转型和可持续发展的作用，然后将数据治理纳入到公司的规章制度中。对于高管层来说，应该确立数据治理的目标，建立机制和流程，明确职责和人员，通过各种审核、控制的方式保障相关部门对数据治理的评估和检查，有效落实问责制。

（3）适当地补充监管指标

增强对核心指标的验证作用。

因此，我们引出了数据治理的概念。

6.1.3 数据治理的概念

数据治理是一套包含策略、原则、组织结构、管理制度、流程以及各种相关技术工具的管理框架。它是数据管理与应用行使权力控制的活动集合，在数据管理与应用层面上进行规划、监督和控制。数据治理是为数据管理、应用与服务提供保障的一种机制。

换句话说，数据治理实质上就是治理数据的政策和管理的方法，具体应该落实到相应的岗位和人员职责上，通过业务流程和数据流程的规范，把数据当成核心财富。如果将数据看做矿山的话，数据治理就是具体的开采方法和手段，如图 6-2 所示。

图 6-2　数据治理类似矿山开采的方法和手段

一般来说，数据治理可以分成两个部分：

1）数据的保障机制，包括政策的制定，考虑使用何种机制、流程和工具去保障数据的

规范性。

2）需要考虑数据的质量标准和数据质量的任责体系。数据治理是企业的责任，需要统一的解决方案和治理模型来保护及共享不同层面的数据。

数据治理可以看做是一门新的学科，能够把企业的独立系统结合起来，重新定义数据的价值和保护机制。从技术上来讲，数据治理是从 OLTP 系统到后台业务数据库，再回到前端的一个闭环的过程。一般来说，数据治理可以解决以下几个方面的问题。

1）制定完善的数据管理机制。

2）对数据进行规范化、标准化和制度化。

3）降低数据维护的难度和成本。

对于商业银行来说，数据治理主要包括建立数据治理机制、数据管理制度及流程，以及对数据标准的制定等内容。数据治理的最终目的是为了提升数据的质量，通过有效的数据整合、清洗、应用和对外服务使商业银行能够具备真正的管理能力和竞争能力。

6.1.4 数据治理体系框架

对于数据治理体系的框架结构，可以包括规划、机制、治理专题和对象、实现 4 个部分，如图 6-3 所示。

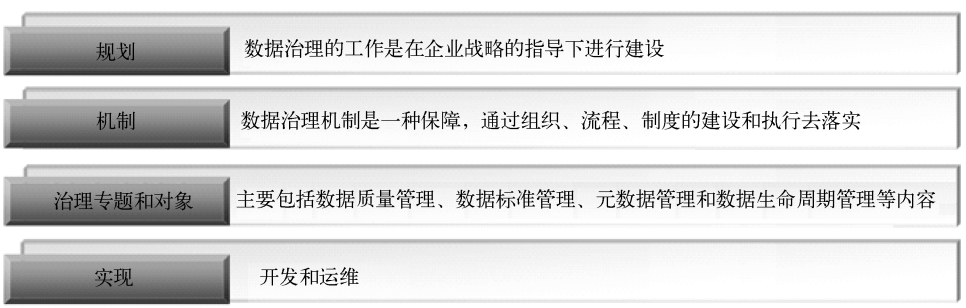


图 6-3 数据治理体系的框架结构

数据治理规划：数据治理的工作是在企业战略的指导下进行建设。

数据治理机制：数据治理机制是一种保障，通过组织、流程、制度的建设和执行去落实，其中数据治理的机制是核心内容，数据治理的执行实质上就是数据治理机制的落实和实现。

数据治理专题和对象是数据治理的主要工作内容，主要包括：数据质量管理、数据标准管理、元数据管理和数据生命周期管理等内容。

数据治理的实现：数据治理的实现包括开发和运维等内容。

6.1.5 数据治理建设的关键要素和成功手段

1. 数据治理建设的关键要素

（1）以数据标准为基础

数据标准为治理体系提供了基本的业务层面保障，统一了业务含义。并且通过对数据使用者和管理者的角色定义，建立了基本的数据管理任责体系。

（2）以提高数据质量为核心

数据治理实质上就是为了提升企业的数据质量，提高企业的运营效率和管理分析的能力，从而最大化地实现企业的业务价值。保证数据质量是数据治理工作最重要的出发点之一。

（3）明确数据治理的职责

一般来说，数据治理是企业高层的职责，可以由高层中的某人负责全企业的数据治理工作，将数据治理的职责赋予管理层的某个委员会，由该委员会确定数据治理的目标和原则，审核数据治理的相关制度、流程，对数据治理的重大问题进行决策。同时对核心数据进行分类，为每类数据分别指定相应的责任部门和责任人。

数据治理建设的成功手段

数据治理建设的成功手段主要由以下几种，如图 6-4 所示。

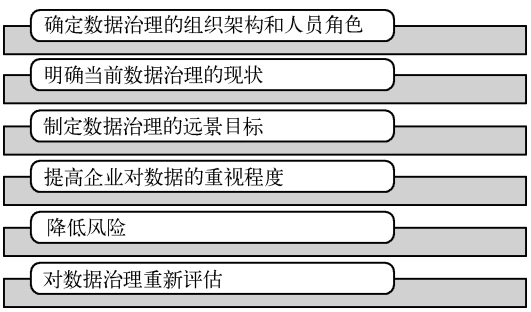


图 6-4　数据治理建设的成功手段

（1）确定数据治理的组织架构和人员角色

首先需要明确数据治理的含义是什么，以及规定数据治理的组织架构和在架构中的各个角色应该承担的职责是什么。在规定的数据治理管理框架下，分别制定每个数据治理对象的管理办法。例如，制定数据标准和数据质量的管理办法。同时，还包括它们的整体管理方针、组织架构划分、职能角色分工以及具体的工作方法、实施细则等内容。

（2）明确当前数据治理的现状

明确数据治理相关的人员组织架构，调研当前的数据治理现状。

（3）制定数据治理的远景目标

当明确数据治理的现状之后，可以由数据治理相关的委员会牵头制定数据治理的远景目标，明确数据治理在几年后达到何种地步，然后制定符合实际的项目计划和里程碑。

（4）提高企业对数据的重视程度

数据不是一种普通的商品，又像水一样重要，但这些宝贵的财富往往会被企业所忽视，因此，提高企业对数据的重视程度已经成为未来研究的必要课题。

（5）降低风险

了解数据在企业的使用情况，研究数据长期的趋势，分析过去事件发生的原因，预测未来数据可能有哪些损失，通过修改当前的政策和管理手段，改善和降低各种风险。

（6）对数据治理重新评估

因为企业每天都可能会发生变化，包括它们的组织机构、人员角色等，它们的数据、价

值和风险也可能会发生变化，所以当企业的组织机构、流程和机制发生变化的时候，应该对数据治理重新评估。

6.1.6 数据治理建设的意义和必要性

数据治理建设的意义，主要包括：对风险进行预警，理解数据，提高数据的管理能力，解决安全运营和风险管理等需求，保证数据的一致性、完整性和可用性等。

我们在了解数据治理的基本情况后，再去深入理解数据治理的几个对象。一般来说，数据治理包含数据标准、数据质量、元数据管理、数据生命周期管理等内容。

1. 数据治理建设的意义

（1）对风险进行预警

数据治理可以帮助企业或者商业银行对各种风险进行预警，从而发挥真正的价值。

（2）理解数据

数据治理可以帮助企业或者商业银行理解并解决它们需要什么数据、如何获取等一系列问题，只有这样才能真正实现对数据的决策分析和数据治理。

（3）提高数据的管理能力

目前国内商业银行的目标是从“以账户为中心”向“以客户为中心”进行转变，经过多年的数据积累和整合，数据治理可以大大提高商业银行的数据管理能力。

（4）解决安全运营和风险管理等需求

数据治理可以解决企业或者商业银行的安全运营、风险管理等多种需求。

（5）保证数据的一致性、完整性和可用性

数据治理体系可以保证数据的一致性、完整性和可用性。

数据治理是保障企业和商业银行安全稳定运营的基础，特别对于商业银行来说，如何避免数据的泄露、篡改，保证数据的一致性和完整性是实现业务连续性的关键。

总的来说，数据治理对商业银行等金融机构尤为重要：

1）数据作为商业银行或者企业的重要资产，相当于人体的血液一样，是非常重要的。

2）高质量的数据，有利于管理决策层进行准确的分析。

3）数据治理有利于保护核心业务数据。

在了解数据治理的基本概况之后，再去深入理解数据治理的几个对象。一般来说，数据治理包含数据标准管理、数据质量管理、元数据管理、数据生命周期管理等内容。

2. 数据治理的主要对象

从技术上来说，不准确的数据会导致系统产生更多的压力和成本，特别是很多数据仓库项目因为数据质量问题而导致失败，所以降低因为数据质量问题而造成的损失和希望得到IT投资回报是实施数据治理的动力。对于企业或者商业银行来说，在交易过程中会产生大量的数据，例如客户基本信息、各种业务信息和系统日志信息等内容。

数据治理工作对于确保银行安全、稳定运营，实现业务创新，具有重要的意义。数据治理是建立数据治理机制，明确责任人，建立数据管理制度和流程的过程。

数据治理的目的就是为了提升数据架构各个层次的管控及其协作能力。数据架构为数据治理提供基础能力支撑，同时把数据当成资产去管理，将价值挖掘出来。

数据治理可以有4个管控机制：政策、组织、流程、技术手段和工具。

对于企业来说，无论是数据、人员还是资产，都可以从这 4 个方面进行分析。首先制定管理政策、流程，建立管理组织，然后建立一个管理系统或者平台，接着把相应的政策、组织和流程固定化和稳定化，再通过企业的管理制度去保障数据治理的执行。

数据治理可以包含 4 个领域：数据标准管理、数据质量管理元数据管理、数据生命周期管理。这 4 个领域都是为了提升数据价值。

下面分别介绍数据标准管理、数据质量管理、元数据管理和数据生命周期管理等相关内容。

（1）数据标准管理

数据标准管理主要解决系统间数据不一致的问题。通过建立规范、政策体系、组织、管控流程和使用相应的技术工具来保证核心数据的一致性和准确性。数据标准是企业级的数据定义，全企业所有的系统都应该遵守和执行数据标准。

（2）数据质量管理

对于数据质量管理来说，可以使用技术工具或者管理平台把可能引发的各类质量问题进行修正，通过改善和提高组织的管理水平，执行相关的政策和流程，使得数据质量得到进一步提高。

（3）元数据管理

元数据管理主要是管理数据，告诉用户系统有什么数据，以及如何去管理数据。它同样通过规范、政策体系、组织、管控流程和使用相应的技术工具来满足对元数据的管理。通过元数据管理可以了解数据的变化过程，包括这些变化会给系统带来什么影响等。

（4）数据生命周期管理

数据生命周期管理解决的是系统效率问题和数据存储问题。首先可以划分 4 个阶段来描述数据的生命周期，包括：数据创建、数据使用、数据归档和数据销毁。然后使用技术工具或者管理平台解决 4 个阶段的问题。通过改善和提高组织的管理水平，执行相关的政策，加强对数据生命周期的管理。

如果企业缺少数据治理，则会产生不一致的业务定义和数据格式，间接导致数据的准确性差，数据交换和共享的成本高，难以解决各种复杂的问题。但是如果企业非常重视数据治理，就会形成统一的业务定义和数据格式。数据会在跨部门和跨系统间得到共享，对数据问题形成跨部门的协调解决机制。

下面从政策、组织、流程、技术工具或管理平台等 4 个方面对数据质量、数据生命周期、数据标准和元数据管理进行分析。

（1）政策

通过制定相应的政策，明确部门的责任，确定数据治理在各个领域的政策、规范，通过制定政策相应的去规范相关人员的行为。

（2）组织

通过建立组织架构和人员角色，确定数据治理相关的责任人，定义不同责任人的角色和职责。

（3）流程

通过制定数据治理各个领域的工作方法和步骤，确定相关人员的分工和合作关系。

（4）技术工具或管理平台

通过技术工具或管理平台保证数据质量的管理成效，支持数据标准和元数据的发布和查询，以及对数据生命周期进行管理。

1）用户可以基于数据治理的成熟度，制定数据治理体系建设的发展路径，优先发展薄弱环节，遵循各个方面均衡发展的原则，保证应用的健康发展。

2）通过数据任责管理机制，建立数据资产的管理体系。把数据看做是银行或者企业宝贵的资产，通过建立一整套的管理体系，对数据进行管理和访问，从而建立有效的、长期的数据治理体系文化。

3）在业务管理和经营过程中，使各个部门的人员都能够体会到数据的作用，从而推动数据标准管理、数据质量管理、元数据管理和数据生命周期管理的建设。最后带动业务的发展，保证数据管理和业务应用相互促进，共同发展。

下面分别叙述数据标准管理、数据质量管理、元数据管理和数据生命周期管理等相关内容。

6.2 数据标准

6.2.1 数据标准概况

一、数据标准的定义

在多数企业和商业银行中，几乎都面临着相同的问题：如何提高对客户的服务水平，如何提高商业银行或者企业的运营效率。其中比较有代表性的解决办法就是采用新技术，突出自身特点，从而吸引客户，同时建立有效的数据治理机制，利用已经积累的数据进行科学化的管理。

因为大多数企业和商业银行的业务系统都是独立建设的，在数据共享过程中，保证数据一致性是最大的困难，数据标准体系就成了解决这个问题的“救命稻草”。数据标准体系为企业或者商业银行的数据整合提供了有力的基础支持。具体来说，数据标准体系为企业建立了标准的数据定义和口径，为数据共享提供了可能性。

那么，什么是数据标准呢？

数据标准是一套完整的数据规范，是数据在使用和交换过程中，为了保持数据一致性和准确性而制定的规范，它主要包括数据分类、业务标准和技术标准的详细定义。数据标准是数据治理中基本的业务和技术层面的保障。

数据标准有利于企业各个部门之间的信息共享，它是数据治理重要的工作方向之一，通过数据标准体系的制定，有利于提升数据管理的水平，保证数据质量的提高，同时确保核心数据的一致性和准确性。

数据标准的工作内容主要包括以下两个方面。

（1）对数据标准分类的划分

如果按照数据的使用范围，来源以及业务逻辑划分，可以将数据标准划分成基础类的数据标准和公共类的数据标准。其中基础类的数据是通过各种业务处理产生的基础数据，例如客户信息、产品信息和各种账户信息等内容。公共类的数据是在基础类数据的基础上，按照

一定的业务规则汇总的数据。

（2）建立数据标准的基本框架

一般来说，基础类的数据标准是标准定义的重点，可以参考行业内先进的经验和数据模型。例如，可以将基础类的数据标准划分为：客户、产品、渠道、交易和活动，如图 6-5 所示。

1）客户。

通过梳理客户相关的业务流程，获取关于客户的核心数据项。包括数据项的组成、分类、业务描述和技术描述等内容。

2）产品。

通过对产品的标准定义和分类，提供统一的产品定义和产品代码等内容。

3）渠道。

通过对渠道的分类，确定渠道主要的信息子类以及该信息子类包含的数据项和定义等内容。

4）交易。

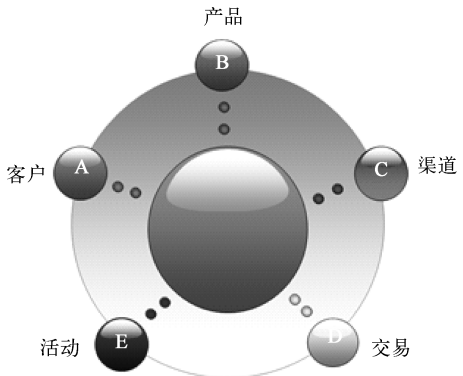


图 6-5　数据标准的基本框架

通过对交易的分类，确定交易核心的信息项及其属性。

5）活动。

根据活动的流程，定义活动主题的信息项、业务描述和技术描述等内容。例如，营销计划、营销方式、营销内容等信息项的组成。

二、数据标准的分类和应用价值

1. 数据标准的整体分类

从整体上来说，数据标准可以分为业务数据标准和技术数据标准。

（1）业务数据标准

业务数据标准是从业务层面上对数据的统一解释和要求，包括重要数据项的业务含义和该数据项在处理、加工过程中应该遵循的业务规则等内容。

从业务的角度来说，数据标准又可以分成基础类数据标准和分析类数据标准。其中基础类数据是企业或者商业银行在日常业务中产生的基础数据，同时按照数据所属的业务主题，进一步划分成不同的主题，例如客户、产品、协议和交易等。对于分析类数据来说，是为了满足企业内部管理的需要，在基础类数据的基础上，按照分析规则进一步加工而成的。

（2）技术数据标准

技术数据标准是从技术实现层面上对数据的统一规范和定义，包括字段长度、数据格式和数据默认值等内容。

从技术角度来说，数据标准可以分为结构化数据标准和非结构化数据标准。

2. 数据标准的价值

数据标准体系的建设对业务部门和技术部门都有较高的应用价值，如图 6-15 所示。

（1）数据标准对于业务部门的价值

对于业务部门来说，可以通过对数据标准的定义，梳理业务需求与流程，通过数据标准确定业务需求蓝本，通过数据标准规范业务分析。

（2）数据标准对于技术部门的价值

对于技术部门来说，在系统设计中可以直接使用数据标准，在开发中直接应用数据标准的映射信息，还可以根据系统建设需求提出对数据标准的修正要求。

数据标准的目的就是在系统内实现数据标准的统一，同时能够为外围系统提供标准化的服务。数据标准可以促进数据质量的提高和数据共享，从而提高整体的业务运营效率和 IT 实施能力。

三、数据标准体系框架

1. 体系框架

数据标准的体系框架可以包括：文化和战略，数据标准内容，数据标准制度和流程，数据标准的组织和角色，数据标准工具。

（1）文化和战略

文化和战略包括数据标准的政策、原则、沟通和协作、宣传等几个方面。政策、原则主要包含数据标准的战略。沟通和协作主要包含协调机制和沟通机制。宣传主要包含数据标准的推广和培训计划等内容。

（2）数据标准内容

数据标准内容包括基础数据标准、公共数据标准。其中基础数据标准是比较重要的，可以包含客户数据标准、产品数据标准、交易数据标准、营销数据标准等内容。

（3）数据标准制度和流程

数据标准制度和流程包括管理制度、管理流程。其中管理制度可以包含数据标准管理制度、数据标准化平台管理制度。管理流程可以包含数据标准的新建流程、变更流程、复审流程和考核流程。

（4）数据标准的组织和角色

数据标准的组织和角色主要包含管理组织和核心角色。其中管理组织包括信息技术委员会、数据治理工作组。核心角色包括数据标准决策者、数据标准管理者、数据标准业务专家、数据标准使用者。

（5）数据标准工具

数据标准工具主要包括标准管理工具和标准知识库。其中标准管理工具包括标准主题管理、业务标准管理功能、技术标准管理功能和标准代码管理。标准知识库包括外部监管和行业最佳实践，以及行业最佳标准化案例。

数据标准的体系框架如图 6-6 所示。

数据标准化的过程实质上就是数据标准设计、管理和应用的过程，目的是为了统一全企业核心的业务定义和技术定义，从而提升企业的业务规范性、业务之间的协作能力和数据的质量。同时，用户可以参考制定数据标准的依据并了解数据标准的功能。

2. 制定数据标准的依据

1）数据标准的制定可以参考外部的标准，例如一些国际、国内的公共标准。

2）数据标准的制定应该参考系统的数据字典和公共代码。

3）数据标准的制定应该参考业务制度和一些管理条例等。

4）数据标准的制定可以参考先进的行业经验。

图 6-6　数据标准的体系框架

3. 数据标准的功能

1）为外部提供标准化的数据服务。

2）指导跨系统的数据整合和模型设计。

3）有效推动跨部门数据的共享。

6.2.2　如何推进数据标准建设的实施

数据标准建设的实施主要包括以下几个方面的内容。

1）首先，将企业战略和规划作为数据标准化建设的指导依据之一。

2）然后，通过合理高效的组织机制能够有效消除业务和技术之间的隔阂，从而有效地推动数据标准的落地。同时由数据标准组负责制定各类数据标准。

数据标准管理者包括：数据标准组长和数据标准专家等。对于业务部门人员和技术部门来说，他们都是数据标准的使用者和执行者。数据标准管理者的组织层次主要为决策层、管理层和执行层。

数据标准决策层主要负责审批数据标准方案，协调重大数据标准事件，同时听取汇报和指导工作。

数据标准管理层主要制定、维护数据标准化的政策、流程和制度等内容。协调和推动数据标准问题的解决。

数据标准的执行层主要包括数据标准使用者。他们主要参与数据标准的制定，配合数据标准管理层组织和实现数据标准的落地。

为了保障数据标准的实施落地，在开发过程中应该设置相应的检查点以保证数据标准的执行管理。

数据标准的开发流程主要包括需求阶段、设计阶段、开发阶段、测试阶段和上线阶段，如图 6-7 所示。

- 在需求阶段可以设置检查点，由需求人员、治理工作组相关的人员检查对于需求的描述是否遵循了数据标准的规范。

- 在设计阶段可以设置检查点，由测试人员将数据标准纳入到测试计划中。
- 在上线阶段可以设置检查点，由数据治理相关负责人审核系统上线时是否遵循了数据标准规范。
- 建立数据治理文化体系，让数据标准化在企业各个部门之间得到广泛宣传。

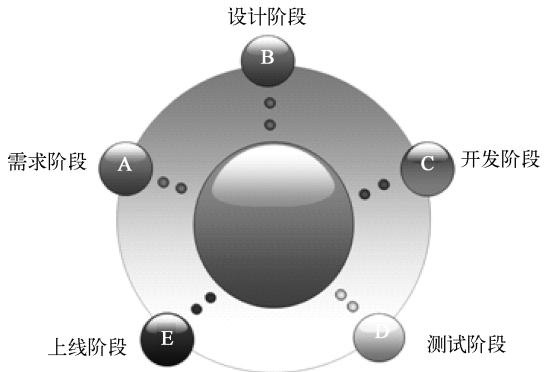


图 6-7　开发流程

3）最后开展数据标准化的专题工作，包括健全数据标准的管理体系，监控数据标准的执行情况，检查数据标准的落地实施。

如图 6-8 所示，我们应该建立由组织规划、制度、技术和专项考核等多种因素相结合的管控机制，从而有效保障数据标准管控机制的执行。

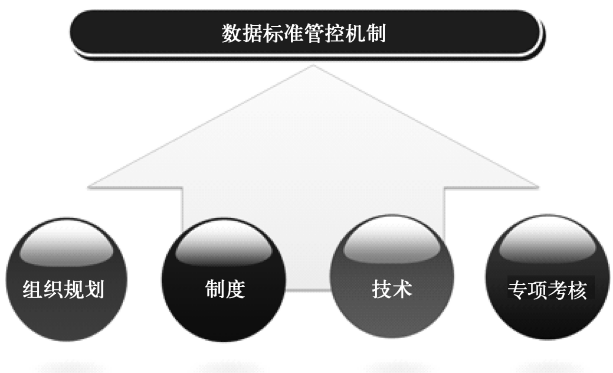


图 6-8　数据标准管控机制

（1）组织规划

完成数据标准管理制度与流程体系规划，建立数据标准管控组织。

（2）制度

为数据标准管控机制的执行提供制度保障。

（3）技术

从技术层面上对数据标准管理系统进行建设。

（4）专项考核

从考核层面上将数据标准的管控机制纳入到绩效考核体系中。

完成数据标准在重要系统的落地工作。通过标准的落地，实现数据定义的统一，促进数据的集中与共享，提升数据质量，支持业务的发展。数据标准在重要系统的落地工作主要包括客户数据标准的落地、公共代码数据标准的落地、产品数据标准的落地，如图 6-9 所示。

（1）客户数据标准的落地

由多个系统生成客户号，可能存在一个客户多个客户号的情况。当客户数据标准落地后，对于新增的客户，统一客户编号。对于存量的客户保留 5 位或者 6 位的编号。客户数据

标准的落地可以统一各系统的客户号，作为客户识别的依据，为客户的归并打好基础，有利于建立统一的客户视图，实现“以客户为中心”的目标。

（2）公共代码数据标准的落地

公共标准代码同步到多个系统中，可以降低代码维护的工作量和系统的复杂度，提高数据的一致性和准确性。

（3）产品数据标准的落地

产品数据标准可以规范产品的分类，有利于提高产品的数据质量。

图 6-9　数据标准在重要系统的落地工作

6.2.3　数据标准项目总体规划和设计

一、数据标准体系总体规划的指导原则

数据标准体系总体规划的指导原则，如图 6-10 所示。

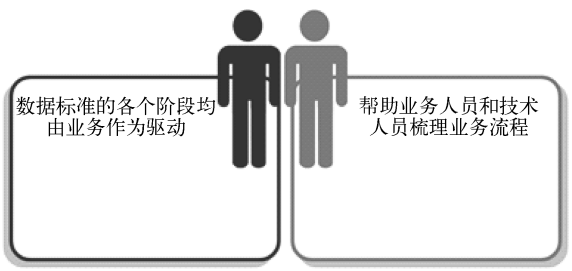


图 6-10　数据标准体系总体规划的指导原则

1）数据标准的各个阶段均由业务作为驱动。可以建立数据标准管理机制，包括每个阶段的主题以及未来落地的方向，同时对各个主题进行定义。

2）帮助业务人员和技术人员梳理业务流程。因为数据标准的主题横跨业务的方方面面，所以数据标准可以帮助业务人员和技术人员明确业务规则，梳理业务流程。

二、数据标准的规划方法设计

数据标准的规划方法可以参考国内外先进的实践经验，并且结合具有行业先进水平的逻辑模型以及专家的经验。

（1）数据标准规划方法

数据标准体系建设的规划方法可以遵循业界先进的方法论，通过调研、规划访谈、数据标准现状分析，了解业务部门对数据标准的期待和想法，将数据标准的需求转化成业务人员可以理解的文档，建立数据标准管理相关的治理架构和管理流程。数据标准规划的过程如图 6-11 所示。

（2）数据标准实施优先级

数据标准实施的优先级需要考虑实施的迫切程度、实施的难易程度和业务关注程度等 3 个方面。

图 6-11　数据标准规划的过程

1）实施的迫切程度。通过对各部门领导的访谈和其他调研工作，了解业务部门在发展过程中关于数据标准方面遇到的挑战和困难。对于那些挑战难度较大、困难较多的主题，在实施顺序上会优先进行考虑。

2）实施的难易程度。数据标准实施的难易程度主要是指从标准的现状，例如数据不一致的程度、整合的难度等方面自下向上地考虑数据标准的实施次序。

3）业务关注程度。业务关注程度是由业务部门针对数据标准的重要性组合而成的，回答了对各自领域的数据标准主题的关注程度。

三、数据标准定义方法设计

数据标准定义方法设计主要包括：数据标准分类、定义数据标准的流程，如图 6-12 所示。

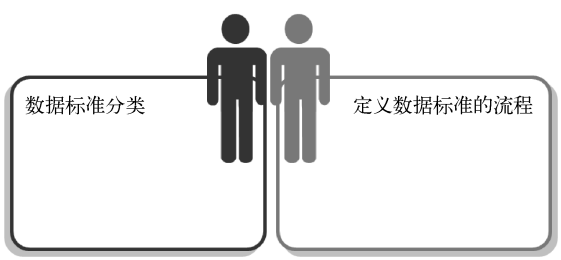


图 6-12　数据标准定义方法

1）数据标准分类。

基础数据标准是数据标准定义的重点。根据行业经验和金融建模的方法论，商业银行基础数据标准按照数据主题可以划分为客户、产品、客户资产、员工与机构、账户、营销活动、交易、渠道、财务和地理位置。这些数据主题既彼此独立，又互相关联。可以参考业界先进的逻辑金融模型，如图 6-13 所示。

基础数据的标准定义框架包括业务主题、基础信息类、信息子类及其业务属性和技术属性。所谓业务属性是根据现状，对客户、产品、渠道、内部机构、协议、地域、财务、事件和资产在内的几大主题进行定义，并对每个主题的重要信息类和子类进行业务规则说明。技术属性定义为数据在应用层面上的技术要求，包括数据长度和格式要求等。

在业务需求和数据整合方面，可以将业务和业务之间的关系抽象成数据之间的关联关

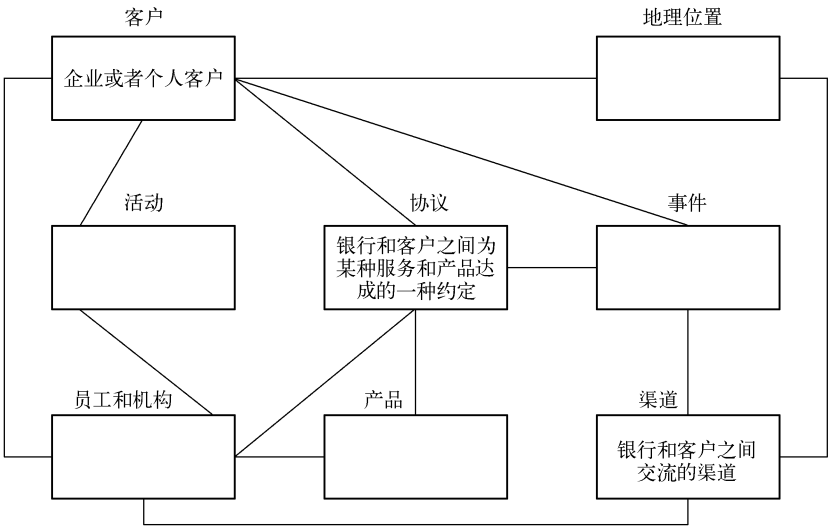


图 6-13　业界先进的逻辑金融模型

系。例如，商业银行的业务数据可以抽象成客户、产品、渠道、内部机构、协议、营销、地域、财务、事件和资产等几大主题。而商业银行的业务领域，例如存贷款、信用卡业务、国际业务、票据业务和投资理财业务等内容都可以包含在这几大主题之中。

通过这几大主题中对业务的描述，可以将银行所有的业务整合起来，例如可以为客户关系管理、风险管理、绩效分析、产品管理分析、渠道分析和利润贡献度分析提供重要的参考。对于基础信息类来说，它是对业务数据的高度概括，例如客户信息、产品信息和渠道信息等，我们把这些由于围绕业务领域而汇集在一起的数据称为信息类。

例如，基础数据标准将业务数据分成客户、产品、渠道、协议、营销等内容，而每个主题又可以分成多个信息类。例如，客户主题包括个人客户信息、对公客户信息、同业客户信息等，而每个信息类又包含一个或者多个信息子类。

对于信息子类来说，它是在信息类的基础上对数据的进一步细分，这种细分是为了描述信息类中的数据项内容。一个信息子类可以包含一个或者多个数据项内容。

数据标准体系中基础数据框架的内容和范围主要包括主题定义、主题间关系、信息类和信息子类等。业务标准和技术标准的例子分别如图 6-14 和图 6-15 所示。

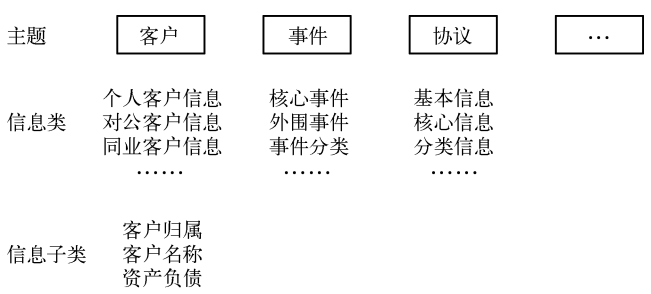


图 6-14　业务标准

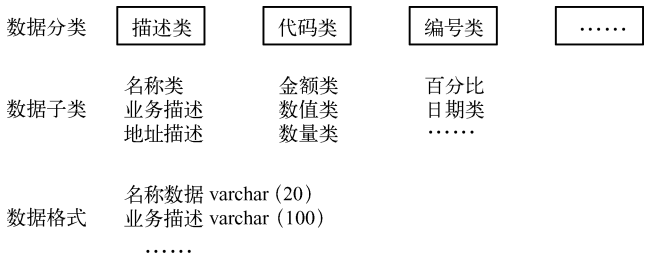


图 6-15　技术标准

2）定义数据标准的流程。

定义数据标准的流程主要包括现状分析、主题定义、标准的审核和标准执行建议，如图 6-16 所示。

图 6-16　定义数据标准的流程

① 现状分析。现状分析主要搜集和整理现有系统中与主题相关的业务和数据定义，对问题进行诊断和分析。

② 主题定义。确定主题的定义、分类以及信息项的范围等。建立数据项的清单，定义每个数据项的标准，包括业务属性和技术属性。

③ 标准的审核。标准的审核是由相关成员对数据标准进行评审，由高层领导最终确认。

④ 标准执行建议。提出数据标准应该遵循的原则以及具体的执行建议。

四、数据标准执行方法设计

数据标准执行方法设计主要包括：以业务需求作为数据标准执行的驱动力、按照计划逐步推进数据标准的建设、制定数据标准的执行策略、完善和管理数据标准的落地和执行。

（1）以业务需求作为数据标准执行的驱动力

数据标准的执行依赖业务部门的需求，只有执行数据标准，才能体现业务的价值。数据标准的执行是以依赖业务需求的迫切程度为前提的。数据标准具有长期性、基础性、迫切性等特点。

（2）按照计划逐步推进数据标准的建设

可以按照计划逐步地推动数据标准的建设。首先选择业务价值高的项目或者专题进行，可以进行一系列的可行性研究和业务价值分析，制定详细的标准落地方案。然后由相关人员进行组织和统一管理。最后，对于一些新建的系统建设项目，需要在开发和设计过程中设置检查点

来确保数据标准的执行，并且不断地完善和充实数据标准。

（3）制定数据标准的执行策略

从业务和技术等多个方面去验证数据标准执行方案的可行性，同时根据分析结果，给出合理的数据标准执行建议。

（4）完善和管理数据标准的落地和执行

在数据标准的定义和落地过程中，不断地完善数据标准的管理办法和规章制度，组织架构和流程。同时还需要加强对数据标准执行过程的评审和监督工作，并且逐步建立和细化数据标准的评审规范。

五、数据标准制定的工作步骤

数据标准制定的工作步骤主要包含以下几个部分：准备阶段；对数据标准的需求数据项进行采集；由业务部门确认关键数据项；制定数据标准，以形成数据标准的初稿；对数据标准进行研讨和确认。

1）准备阶段。

准备阶段主要包括对数据标准现状的调研、工作方法和工作模板的准备工作等内容。

2）对数据标准的需求数据项进行采集。

该步骤主要内容包括数据项的来源类型、数据项的来源、主题域、主题域大类、主题域细类、共享项名称、数据项中文名称、系统表中文名称、系统表英文名称、表内字段英文名称、说明、是否纳入共享项等内容，见表 6–1。

表 6–1　对数据标准的需求数据项进行采集

数据项的来源类型	数据项的来源	主题域	主题域大类	主题域细类	共享项名称	数据项中文名称	系统表中文名称	系统表英文名称	表内字段英文名称	说　明	是否纳入共享项
业务系统	CMP	对私客户	基本信息	个人客户名称、个人工作情况	客户编号	集中签约客户号	客户基本信息	t_cust_info	Cust_no	客户编号类别 + 证件号码 + 后缀	1
业务系统	ECIF	对私客户	基本信息	个人客户名称、个人工作情况	客户编号	ECIF 客户编号	客户基本信息	M_indi	ECIF_no	客户唯一编号	1

3）由业务部门确认关键数据项。

该步骤主要内容包括：系统和业务现状，数据标准制定依据，数据标准制定建议，是否制定标准，数据标准名称等。

4）制定数据标准，以形成数据标准的初稿。

5）对数据标准进行研讨和确认。

根据提出的不同意见进行标准修改，最终形成数据标准。

总之，在数据标准执行过程中，需要对现有的数据标准进行管理和维护，在落地过程中，逐步地完善数据标准管理流程和规范。

六、推动数据标准落地的方法

推动数据标准落地的方法主要包括以下几种。

1. 通过业务驱动推动数据标准的落地

主要通过业务部门的需求、标准执行的效果和业务现实迫切程度等几个方面去推动数据标准的执行和体现业务的价值，同时它们都是数据标准落地执行的前提和重点。

2. 通过制定计划和采用监督评审的方式推动标准的落地

主要包括按计划进行系统的改造和监督评审等工作，如图 6–17 所示。

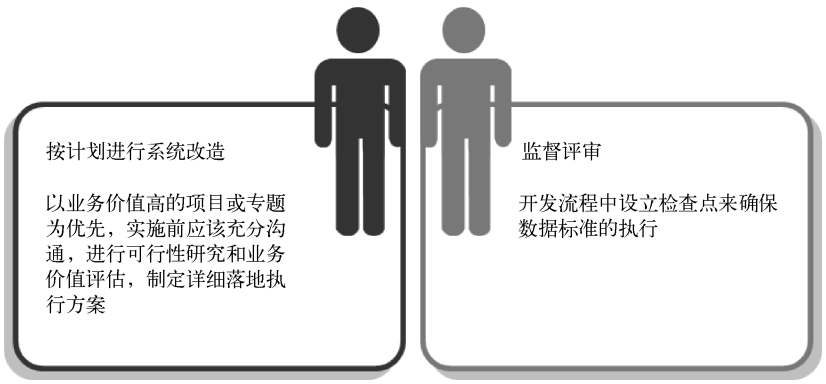


图 6–17　制定计划和采用监督评审的方式推动标准的落地

3. 通过阶段实施的方式推动标准的落地

主要包括可行性研究、价值评估、设定范围和差异执行等工作，如图 6–18 所示。

可行性研究	价值评估	设定范围	差异执行
• 业务影响	• 业务价值评估	• 主题范围	强制执行 参考执行
• 技术影响	• 业务部门支持	• 实施层次	全部执行 参考执行
• 系统关联	• 需要吻合度	• 预期目标	业务缺口 技术缺口
• 改造工作量	• 试点推进	• · · ·	• · · ·

图 6–18　通过阶段实施的方式推动标准的落地

4. 建立数据标准的闭环管理流程

数据标准只有在业务系统的日常运营过程中才能发挥其作用。数据标准可以提高数据的共享性和一致性。数据标准的闭环管理流程包括标准应用、标准发布、标准维护和标准监控，如图 6–19 所示。

5. 通过完善管理组织和流程去推动标准的落地

主要内容包括管理办法/规章制度、组织架构和流程，如图 6–20 所示。

举例来说，可以参考外部标准、监管要求。先进经验和逻辑模型来规划数据标准体系，如图 6–21 所示。

数据标准管理是一项具有系统性、复杂性和长期性特点的工作 。

随着标准的落地和执行，我们可以不断地完善数据标准，建立数据标准动态管理机制，

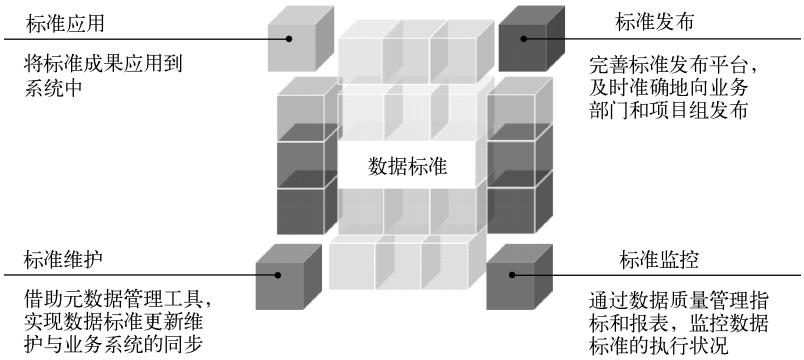


图 6-19　建立数据标准的闭环管理流程

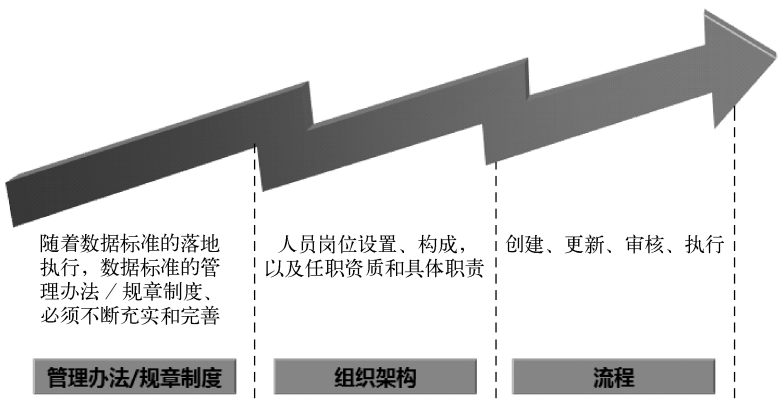


图 6-20　通过完善管理组织和流程去推动标准的落地

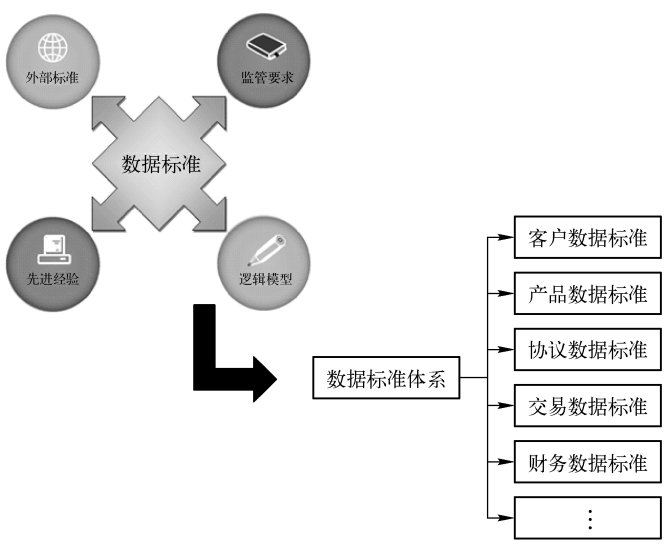


图 6-21　规划数据标准体系

通过数据管理系统进行标准的更新和发布，推动数据标准在业务领域的落地和执行。业务管理部门在制定业务制度和产品创新时应该遵循数据标准，IT 操作人员在系统内进行数据采集和维护过程中应该执行数据标准，加快数据标准在技术领域的落地。

对于数据标准工作来说，落地执行是重点，业务驱动是关键，配套落实是保障，如图 6–22 所示。

（1）落地执行是重点

数据标准只有在执行时才能体现标准的价值，包括对业务、技术和业务流程的借鉴，然后不断地修正和完善数据标准。

（2）业务驱动是关键

数据标准的建立和使用不能脱离业务需求，真正解决实际问题才是数据标准实施的动力。

（3）配套落实是保障

通过一系列的配套落实来保障数据标准纳入到整体的治理体系中，从而监控数据标准的执行状况。

图 6–22　数据标准工作

考虑数据标准执行的先后顺序。对于渠道、公共统计口径和产品目录及其定义等指标的数据标准，按照其重要性，分别划分成高、中、低三个部分，如图 6–23 所示。

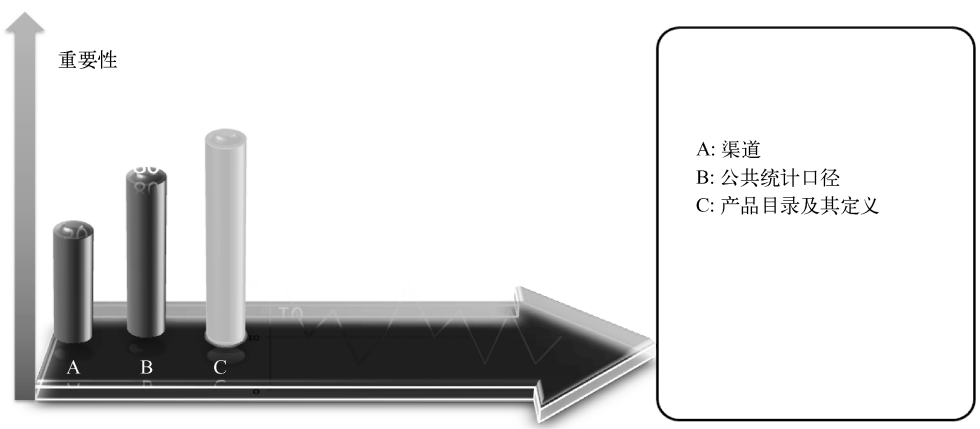


图 6–23　数据标准执行的先后顺序

标准体系的实施路线图的制定包括建立数据标准管控体系、数据标准定义和数据标准落地，如图 6–24 所示。

具体过程举例如下：

（1）数据标准管控体系 – 组织流程 – 建设初期

建立数据标准小组机制和管理流程。

（2）数据标准管控体系 – 组织流程 – 建设中期

建立专职机构和管控绩效指标体系。

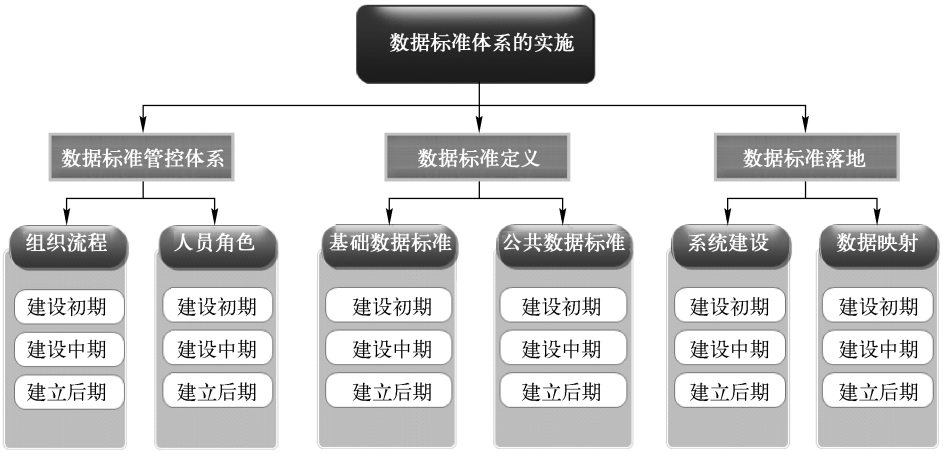


图 6-24　标准体系的实施路线图

（3）数据标准管控体系 – 组织流程 – 建设后期

定期复审数据标准体系，保证数据标准的合理性。

七、数据标准管控规范、管控原则、管理组织和管控流程

1. 数据标准管控规范

数据标准管控规范包括数据标准制定管理办法、数据标准审核管理办法、数据标准发布管理办法和数据标准管理规范等内容，如图 6-25 所示。

（1）数据标准制定管理办法

明确数据标准制定的部门；明确数据标准制定的工作环节和工作细节。

（2）数据标准审核管理办法

明确数据标准审核的部门；明确数据标准审核的工作环节及工作细则。

（3）数据标准发布管理办法

明确数据标准发布的部门；明确数据标准发布的工作环节及工作细则。

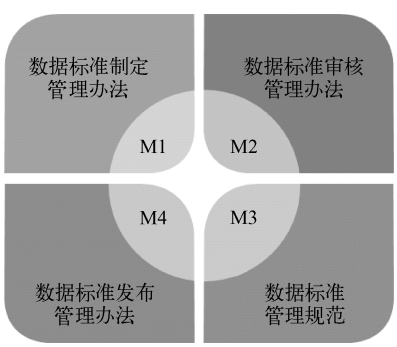


图 6-25　数据标准管控规范

（4）数据标准管理规范

明确数据标准管理工作方向与思路；明确数据标准管理部门以及各部门在工作中承担的角色与职责。

2. 数据标准管控原则

数据标准管控原则主要包含唯一性、稳定性、前瞻性、准确性、可执行性和低风险性。

- 唯一性

主要保证数据标准的命名、编码和业务解释的唯一性。

- 稳定性

主要维持数据标准的权威性和稳定性。

- 前瞻性

数据标准的调研、设计和执行要具备前瞻性。

● 准确性

对数据标准的业务定义、业务名称和口径都应该具备准确性。

● 可执行性

主要考虑业务实际情况和未来发展，保证数据标准具有可执行性。

● 低风险性

主要考虑各种业务风险和实施风险，保证数据标准能够顺利实施和落地，降低风险性。

3. 数据标准的制度规范、数据标准管理办法和数据标准制定的工作方法

（1）数据标准的制度规范

指明数据标准管理工作方向与工作思路，明确参与数据标准管理工作的部门以及各部门在工作中承担的角色和责任。

（2）数据标准管理办法

明确参与数据标准制定的工作部门以及数据标准制定的工作环节及工作细节。

（3）数据标准制定的工作方法

明确数据标准制定的工作方法和原则。

如图 6-26 所示。

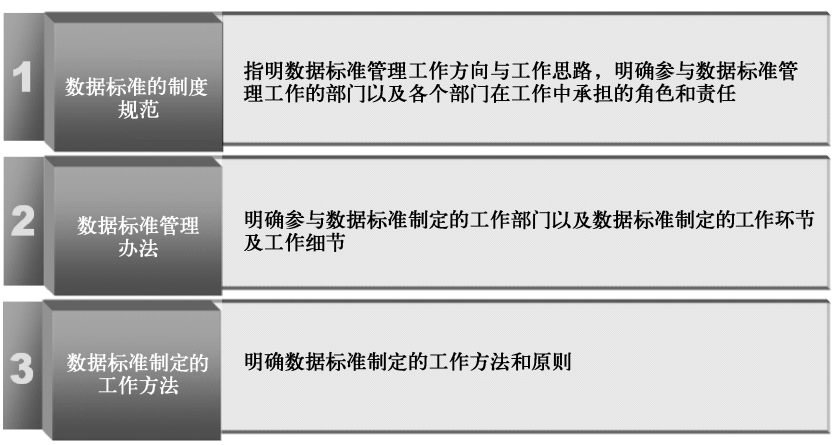


图 6-26　数据标准的制度规范、管理办法以及制定的工作方法

4. 数据标准的管理组织和管控流程

数据标准管理组织说明如图 6-27 所示，包括：建立数据管控办公室，设置数据标准主管和数据标准管理员；设立数据标准责任人，包含数据标准负责人、数据录入人员和数据使用人员等，设立系统责任人，分为数据标准负责人和系统负责人。

数据标准管控流程主要包括标准申请、标准规划、标准审核、标准实施和标准规划评估，如图 6-28 所示。

（1）标准申请

数据标准的申请流程是通过制定计划，提出修订数据标准的申请，同时提交给上层领导进行审核，最后明确责任人的过程。它的主要工作是由标准管理员制定相应的计划，再由数据标准的使用者或者系统负责人提交标准制定的申请，最后由标准的负责人审核相关申请，由数据标准管理员将标准分配给相应的责任人。

图 6-27　数据标准管理组织

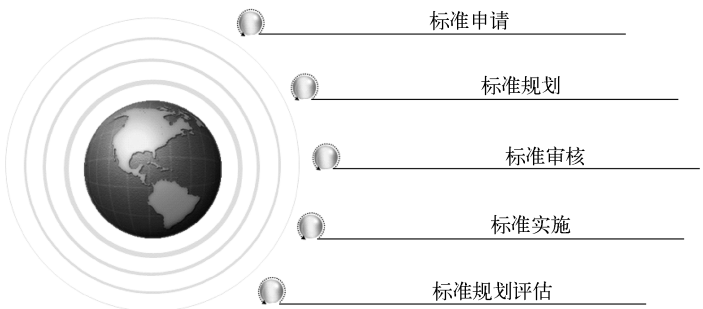


图 6-28　数据标准管控流程

相应的责任人可以包括数据标准管理员、数据标准使用者、系统负责人、业务数据标准负责人和技术数据标准负责人等。

流程主要工作内容包括：

1）制定数据标准的相关计划。

2）明确数据标准相关人员、角色和相应的职责。

3）记录数据在标准应用过程中存在的问题。

4）由相关人员提出数据标准新增、修改、删除的申请。

5）将数据标准的申请提交到决策层审核。

6）由数据标准管理员明确相应的责任人。

数据标准申请流程如图 6-29 所示。

1）制定计划。

由数据标准管理员制定计划。

2）提出申请。

由数据使用人员、系统负责人提出申请。

3）审核申请。

由业务、技术数据标准负责人审核申请。

4）明确责任人。

由数据标准管理员明确责任人。

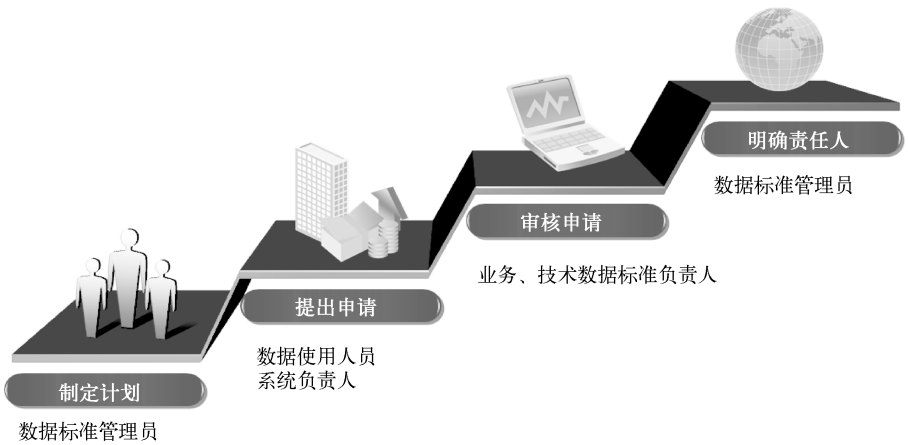


图 6-29　数据标准申请流程

（2）标准规划

数据标准的主要工作是通过对标准现状的调研和分析，制定业务数据标准和技术数据标准，最后形成数据标准初稿，如图 6-30 所示。

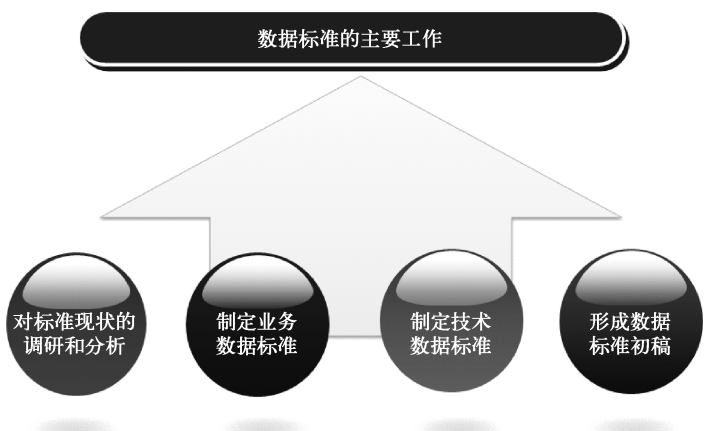


图 6-30　数据标准的主要工作

数据标准规划流程如图 6-31 所示。

- 现状分析

由业务数据标准负责人、技术数据标准负责人进行现状分析。

- 数据标准业务定义

由业务数据标准负责人进行数据标准业务定义。

- 数据标准技术定义

由技术数据标准负责人进行数据标准技术定义。

● 数据标准初稿制定

由业务数据标准负责人、技术数据标准负责人制定数据标准初稿。

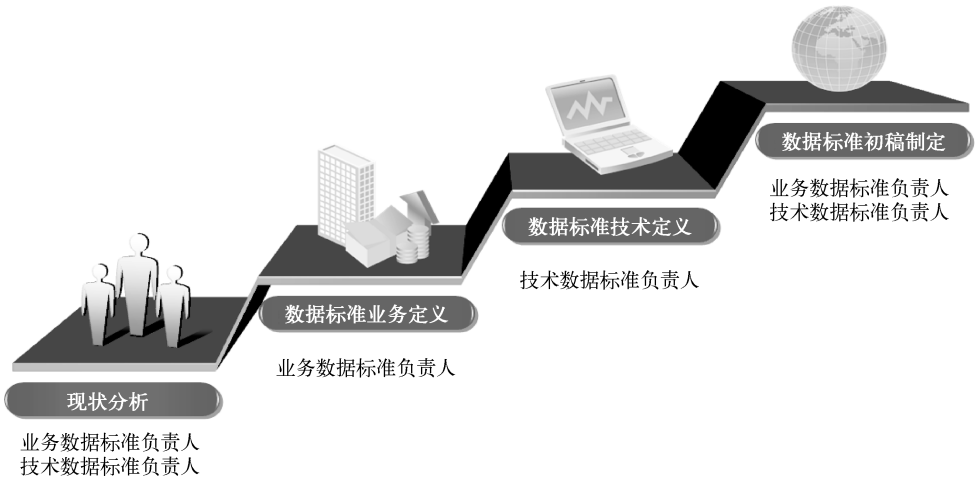


图 6-31　数据标准规划流程

（3）标准审核

标准审核的主要工作是对数据标准规划进行审核，审核通过后，再对相关部门进行批复和发布，如图 6-32 所示。

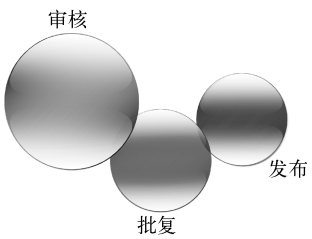


图 6-32　标准审核

数据标准审核与发布流程，如图 6-33 所示。

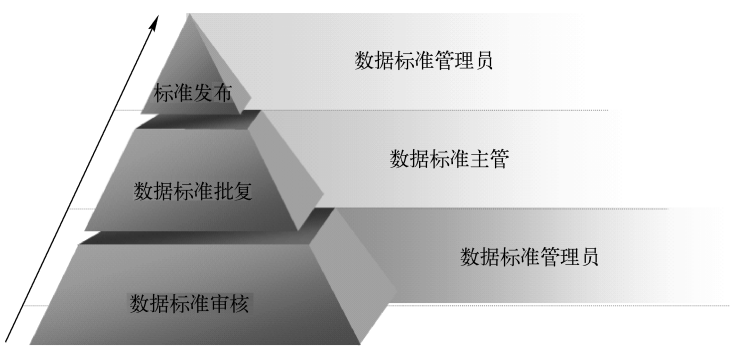


图 6-33　数据标准审核与发布流程

（4）标准实施

数据标准的实施的基本步骤包括制定数据标准实施方案、审核数据标准实施方案和数据标准的实施，如图 6-34 所示。相关的责任人可以是数据标准管理员、数据标准主管等，如图 6-35 所示。

1）制定数据标准实施方案

由数据标准管理员制定数据标准实施方案。

2）审核数据标准实施方案

由数据标准主管审核数据标准实施方案

3）数据标准的实施

由系统负责人进行数据标准的实施。

图 6-34　数据标准的实施的基本步骤

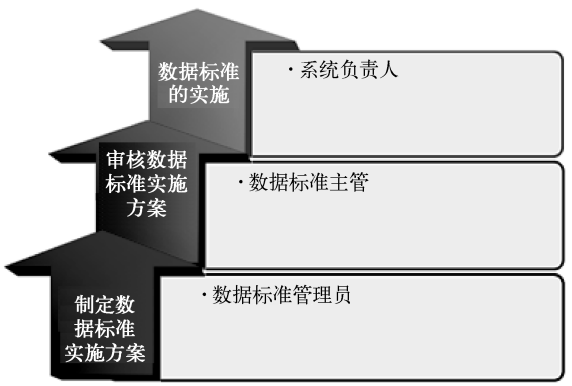


图 6-35　数据标准的实施流程的相关负责人

（5）标准规划评估

对数据标准规划进行定期评估，根据评估结果对数据标准规划进行修正，保证数据标准的正确性。标准规划评估流程主要包括评估规划、审核方案、标准评估和标准变更，如图 6-36 所示。

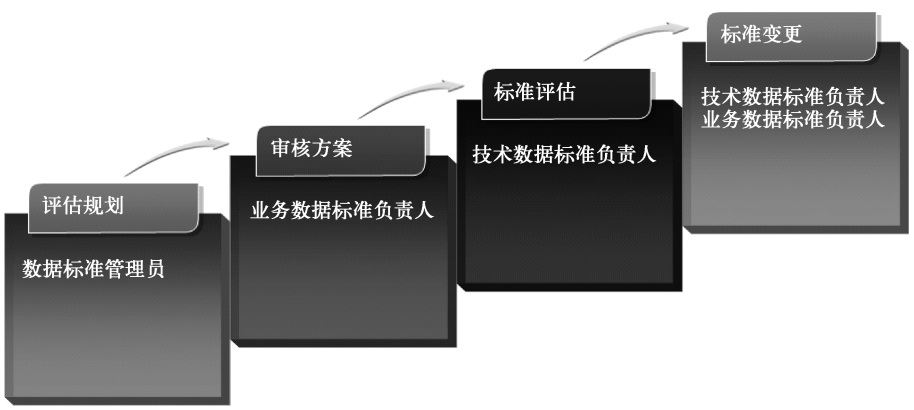


图 6-36　标准规划评估

八、数据标准的全面定义

数据标准是通过一整套的数据规范、管控流程和各种技术工具确保重要的数据是一致的

和准确的。例如，通过数据标准保证产品、客户、机构、账户等内容都是一致的、准确的。

1. 数据标准体系设计指导原则

数据标准体系设计指导原则包括唯一性、稳定性、前瞻性、准确性、可执行性和低风险性，如图 6-37 所示。

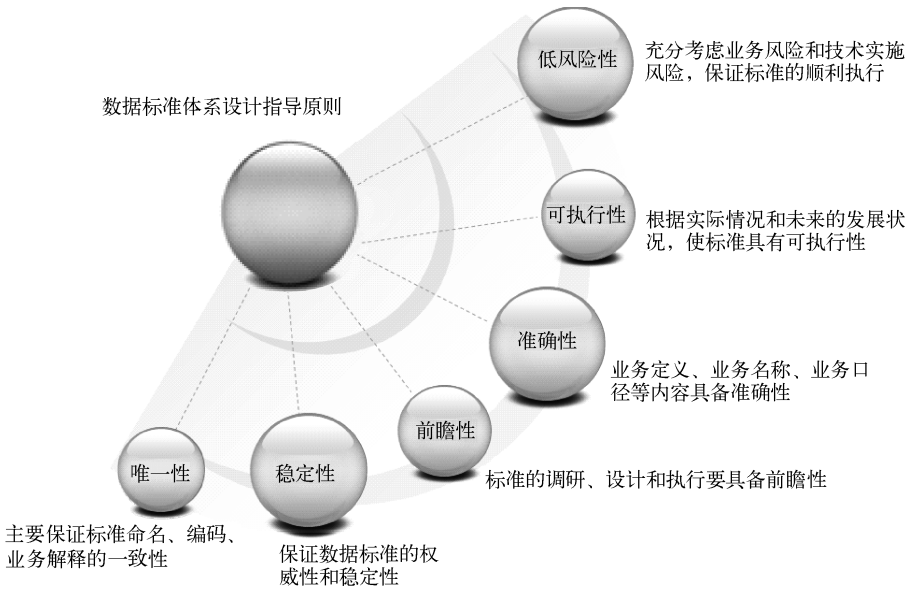


图 6-37　数据标准体系设计指导原则

2. 数据标准包含的内容

数据标准是企业级的数据定义，企业所有系统都应该遵守和执行数据标准。数据标准可以包括每个数据项的业务属性、技术属性和管控属性，如图 6-38 所示。

业务属性	标准编号、标准中文名称、标准英文名称、标准别名、业务定义、业务规则、相关标准关系、标准来源
技术属性	数据类别、数据格式、取值范围、编码规则
管控属性	标准定义部门、标准使用系统

图 6-38　数据标准包含的内容

（1）业务属性

主要包括标准编号、标准中文名称、标准英文名称、标准别名、业务定义、业务规则、相关标准关系、标准来源。

（2）技术属性

主要包括数据类别、数据格式、取值范围、编码规则。

（3）管控属性

主要包括标准定义部门、标准使用系统。

例如，客户张三、李四的年龄和性别分别为 40 岁、50 岁，男、女。此时，性别编码出现了不一致，见表 6-2 和表 6-3，这就需要针对两张表的内容制定统一的数据标准。

表 6-2　性别编码 1

客 户 姓 名	年　　龄	性　　别
张三	40	M
李四	50	F

表 6-3　性别编码 2

客 户 姓 名	年　　龄	性　　别
张三	40	00
李四	50	01

统一后的数据标准如图 6-39 和图 6-40 所示。

代码编号	CD000001
中文名称	姓名代码
英文名称	
代码描述	描述人的性别代码
定义原则	
引用标准代号及名称	
技术属性	Char（2）
版本日期	
标准类别	标准
⋮	

图 6-39　编码 1

低码值	低码描述	业务说明
01	未知性别	
02	男性	
03	女性	
99	未说明性别	

图 6-40　编码 2

上文提到数据标准包括每个数据项的业务属性、技术属性和管控属性，举例如图 6-41 所示。

九、数据标准的应用过程

1. 数据标准的应用过程

数据标准的应用过程如图 6-42 所示。

2. 数据标准项目建设过程

1）根据数据标准的实施路线图，可以有计划地进行数据标准主题定义工作，逐步实现

业务属性	标准大类	基本信息	标准子类	基本概况
	标准小类	证件信息	标准编号	1100
	中文名称	证件号码	英文名称	Identi_Code
	业务定义	描述个人客户某种证件的具体号码信息，如身份证号		
	业务规则	公民身份证号码是特征组合码，由 17 位数字本体码和一位校验码组成		
	行内制度或者外部标准			
技术属性	标准格式	文本		
	取值范围			
管控属性	标准拥有部门	零售部		
	信息使用部门	×× 银行中心		

图 6-41　编码表

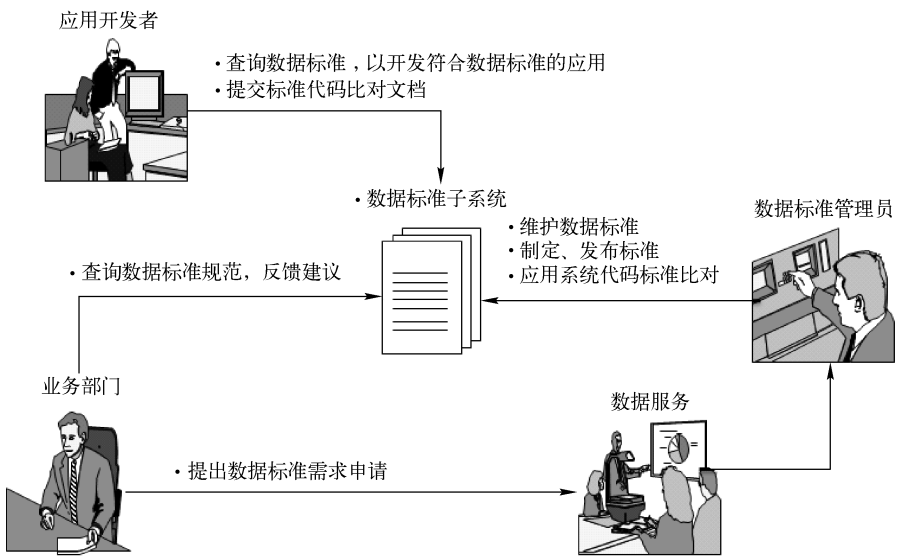


图 6-42　数据标准的应用过程

数据在业务和技术上的统一。

2）结合最佳实践，推动数据标准在系统建设中的执行和落地。

3）建立数据标准管理组织和流程，实现标准的维护、发布，同时实现数据标准的制度化，提高全企业的数据标准文化和标准的管理水平。

4）数据标准体系可以分成数据标准规划、数据标准定义、数据标准执行和数据标准管理等几个部分。其中数据标准规划是标准体系的总纲；数据标准定义是数据标准最重要的部分之一，是业务和技术进行沟通的桥梁；数据标准执行是业务价值体现的部分；数据标准管理是标准在系统中正常使用的保障。

数据标准规划是长期的和基础性的工作，通过对高层领导和业务部门的访谈，包括对数据标准实施的迫切性、难易度和业务部门对数据标准的关注度进行综合衡量，提出具体的数据标准实施路线图。

数据标准定义是在参考相关标准体系分类的基础上，确定数据标准的框架内容。对于商业银行来说，可以包括客户、产品、员工和机构、活动、交易等内容。数据标准不仅需要对核心的主题进行详细定义，而且还需要描述该主题的业务属性和技术属性。

数据标准执行是按照业务需求的紧迫程度，制定合理的执行方案。对于标准的管理工作是建立相应的管理组织，包括建立领导小组、标准管理办公室，设立数据标准管理员和相应的业务专家等。通过数据标准工作流程的制度化和工作化，提高数据标准的管理水平和管理效率。数据标准体系的产出物可以包括数据标准体系的规划、标准主题的定义、执行建议和管理制度等内容。

根据业务需求和对系统现状的理解，编写数据标准实施计划、数据标准实施路线图和各个阶段的实施内容等。其中对数据标准主题的定义可以包括：客户主题数据标准、产品主题数据标准、渠道主题数据标准、交易主题数据标准、内部机构主题数据标准等。

客户主题数据标准是根据现有的数据现状、客户信息的使用情况，细化对客户主题的标准定义，包括对客户主题的详细定义、数据项类别、业务标准和技术标准等内容。

产品主题数据标准同样是根据现有的数据现状、各业务部门对产品信息的需求，细化产品主题的数据标准定义，包括产品主题的定义、产品特征和属性、对产品属性的标准定义等内容。

3. 数据标准的主要应用

数据标准的应用以业务标准和技术标准为基础，是业务部门和技术部门沟通的桥梁，同时为 IT 系统的建设提供重要参考。随着标准体系建设的不断深入，可以逐步实现数据标准对各个应用系统的指导作用，可以促成系统的集成和数据的共享，真正实现业务价值。

对于商业银行来说，可以借助企业级客户管理项目（ECIF）的建设和实施，实现客户主题标准在 ECIF 中的全面落地。在 ECIF 项目的需求和设计阶段，数据标准小组可以提出需求，同时进行数据标准解释工作等。

通过数据标准体系在相关系统中的应用，为业务部门和技术部门产生价值。

如图 6-43 所示，数据标准的主要应用包括：数据标准定义分析；通过数据标准的建设，优化业务流程和提高业务价值；利用数据标准，解决业务需求统一的问题；在数据标准的定义中，数据标准与源系统的映射关系反映了现有系统和数据标准之间的关系；数据标准的完善是一个闭环的过程等。

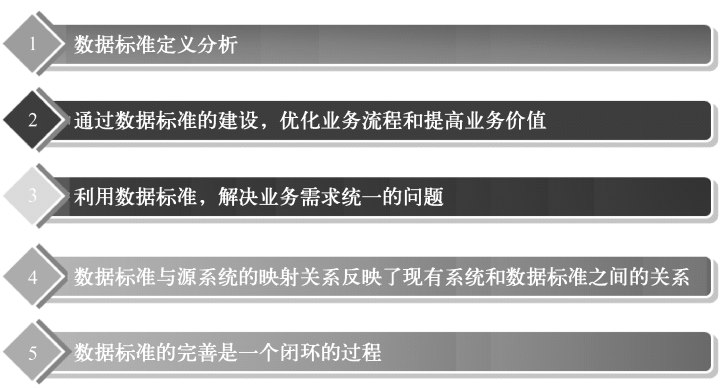


图 6-43　数据标准的应用

1）数据标准定义分析。通过数据标准对各个信息项的标准定义，包括业务定义和技术定义，使管理人员和业务人员通过数据标准了解统一的标准口径、业务定义和每个信息项的业务含义，提高数据的一致性和共享性。

2）通过数据标准的建设，优化业务流程和提高业务价值。数据标准可以对业务流程进行优化和改进。例如，证件类型的数据标准化可以实现对居民身份证的有效支持，提高客户的服务能力，特别是在 ECIF 项目中，对数据标准的实施有利于优化客户归并的业务流程，同时提高数据质量。

3）利用数据标准，解决业务需求统一的问题。数据标准的定义是基于业务部门和技术部门的讨论和确定后得到的。例如，ECIF 系统代码采用公共代码数据标准，可以减少业务需求统一的工作量，满足业务需求。

4）在数据标准的定义中，数据标准与源系统的映射关系反映了现有系统和数据标准之间的关系。

5）数据标准的完善是一个闭环的过程。例如，数据标准的执行为相关业务与技术的规划提供参考，同时业务需求的变化促使对数据标准的修订，然后逐步完善数据标准。

如何保证数据标准的可持续发展和不断完善？可参考如下内容。

1）通过遵循业务需求，推动数据标准在全企业的落地实施。

2）数据标准需要结合企业战略和业务需求，这样才能体现业务的价值。在这种思路下，开展数据标准的定义、执行工作，形成数据标准、业务需求和系统设计开发三者之间的融合。

3）可以借助数据标准的评审工作，以及对数据标准管理系统的建设，促进数据标准的执行和落地。

6.2.4 数据标准项目总结

数据标准建设是长期性的工作，对于企业或者商业银行来说，数据标准体系建设的好坏直接影响企业内部管理水平和对外服务的能力。

数据标准建设可以引入先进的行业经验和方法论，从数据标准的规划、定义、执行和管理等各个方面进行标准体系的建设，提高全企业的数据标准文化水平。标准体系的建设依赖业务需求，它也是一个长期的过程。

6.3 数据质量管理

6.3.1 数据质量管理概况

1. 数据质量管理概念

数据质量管理可以通过提高管理水平，严格执行相关的政策和规范，或者使用一些技术工具，使得数据质量得到进一步的提升。对于数据质量管理来说，它是一个闭环的管理过程，经过不断循环、改善，逐步提高数据的质量，并最终为企业赢得经济效益。

数据质量管理的目的是提升系统的数据质量，业务人员通过数据质量管理系统发现数据在流转过程中存在哪些问题，经过不断修正和完善，使数据质量得到不断提升。

数据质量管理目标是提升数据的正确性、一致性和完整性。通过数据质量管理办法、组织、流程，发现数据质量问题并且及时得到解决，从而最大限度地提升业务价值。

数据质量存在问题的原因归为以下几类：如图 6-44 所示。

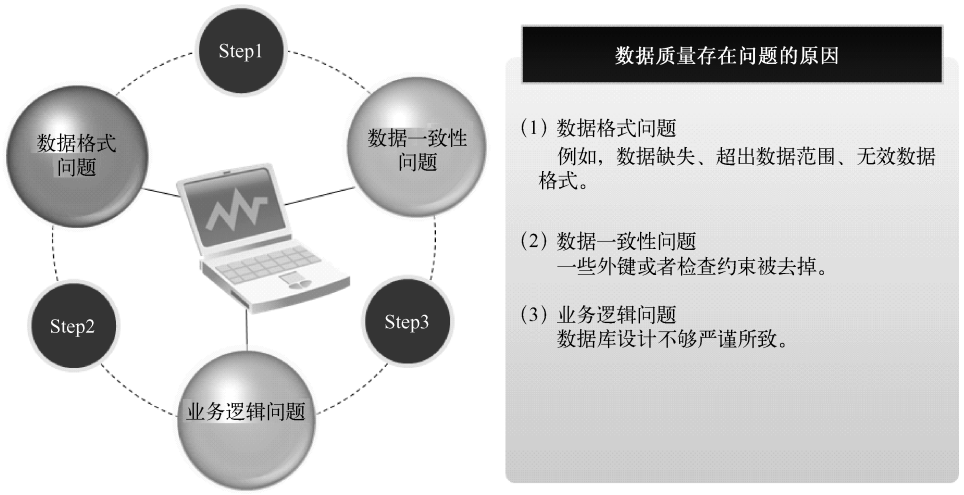


图 6-44　数据质量存在问题的原因

2. 数据质量管理办法和实施细则

数据质量管理办法包括质量管理的工作方向和工作思路，例如数据质量问题的识别、评估与处理。明确参与的部门、人员，包括在数据质量管理工作中承担的角色和职责。

数据质量实施细则包括质量检查规范管理办法，明确质量检查中的参与部门以及具体的流程，例如问题的收集、更新和终止。

3. 数据质量管理范畴

技术检查指标主要包括空值检查、空格检查、日期字段检查、唯一性检查和编码检查如图 6-45 所示等。

- 空值检查

判断字段值是否为空，是否需要赋默认值。

- 空格检查

判断字段值是否为空格，是否需要赋默认值。

- 日期字段检查

判断该字段是否为合法的日期，是否需要赋默认值。

- 唯一性检查

唯一性检查主要是针对业务唯一性的检查。

- 编码检查

检查编码的合法性。

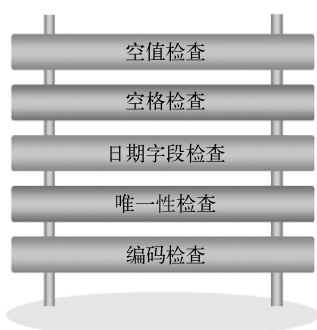


图 6-45　技术检查指标

4. 数据质量管理框架

数据质量管理框架如图 6-46 所示，主要包括关于数据质量管理政策、组织、流程和技术工具。其中管理政策包括数据质量管理方法、数据质量实施细则，组织包括数据质量角色定义、数据质量职责划分，流程包括数据质量事前防范、数据质量事中监控、数据质量事后

改进，技术工具主要包括数据质量管理系统。

图 6-46　数据质量管理框架

综上所述，数据质量概况如下：

（1）数据质量管理的定义

数据质量管理的目的是提升数据的质量。业务人员通过数据质量管理系统发现数据在流转过程中存在哪些数据质量问题，经过不断修正，使数据质量不断得到提升。

（2）数据质量的管理目标

数据质量的管理目标是提升数据的正确性、一致性和完整性。通过数据质量管理办法、组织、流程，发现数据质量问题并且及时得到解决，从而最大限度地提升业务价值。

（3）产生数据质量问题的原因

数据质量问题的原因包括数据格式问题、数据一致性问题和业务逻辑问题等。

6.3.2　数据质量管理的设计方法和流程

数据质量管理的设计方法和流程包括：数据质量管理总体规划、数据质量管理的解决办法和数据质量管理的执行等。

（1）数据质量管理总体规划

数据质量管理总体规划主要包括总体规划的指导原则、数据质量管理基本制度及规范、数据质量管理规范和管理办法、数据质量管理组织和数据质量管控流程等内容。

（2）数据质量管理的解决办法

数据质量管理的解决办法主要包括定义、发现、分析、反馈、整改和监控。

（3）数据质量管理的执行

数据质量管理的执行主要包括提供考核指标问题查询、相关 IT 部门进行数据提升和数据质量管理人员进行管理操作等内容。

一、数据质量管理总体规划

1. 数据质量管理总体规划的指导原则

数据质量管理总体规划的指导原则主要包括：完整性原则、正确性原则、一致性原则、及时性原则和适当性原则。

- 完整性原则

所有的信息、属性是否按照系统和业务规则完整填写。

- 正确性原则

是否准确地收集到相关信息，并如实在系统中进行录入和处理。

- 一致性原则

不同系统、业务之间关联的数据是否一致，包括一致的定义、含义、取值及操作规则等。

- 及时性原则

数据是否能够及时地被获取，是否能够反映当前业务运营状况，以满足对数据进行加工、查询和分析的业务需求。

- 适当性原则

数据是否适当地进行了发布和使用，以确保数据的安全性。

2. 数据质量管理基本制度及规范

无论是事前防范、事中监控还是事后改进，必须遵循数据质量管理制度和规范。

3. 数据质量管控规范和管理办法

对数据质量的管控包括以业务需求为导向，选取对数据质量要求最为紧迫的数据，并且设定相应的数据质量指标。然后制定数据质量的管控规范和管理办法。

4. 数据质量管理组织

数据质量管理组织主要落实管理的组织架构和相应的岗位职责，从而保证事前防范、事中监控和事后改进的落地执行。当这三个流程发生变化的时候，可能会调整相应的管理组织架构。

数据质量管理组织举例如下：

- 数据质量管控委员会
- 数据质量主管

主持数据质量管理全面工作，并对数据质量管理的各项工作结果负责。

- 数据质量管理员

指导相关业务部门和技术部门对数据质量管理的执行；组织和协调相关部门对于数据质量检查规则的制定；保证数据质量管理建设方法顺利执行，同时进行日常的监督和管理

数据质量管理组织包括数据管控办公室、数据责任人和系统责任人三个角色。

- 数据管控办公室：数据质量主管和数据质量管理员。
- 数据责任人：数据质量负责人、数据录入人员和数据报送机构。
- 系统责任人：数据质量负责人、系统负责人和系统运维人员。

5. 数据质量管控流程

数据质量管控流程主要包括事前防范、事中监控、事后改进。

（1）事前防范

事前防范数据质量问题主要包括数据质量问题的总结、数据质量问题的分析和汇总、数据质量防范方案规划、数据质量防范方案评审、数据质量防范方案实施和数据质量防范方案最后评估等内容。

事前防范是对数据质量问题尽可能地进行规避和防范。数据质量事前防范流程如图 6–47 所示。

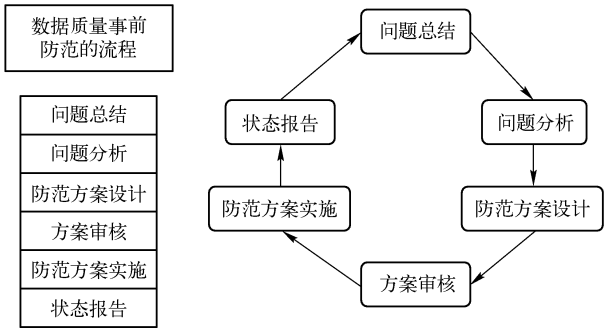


图 6-47　数据质量事前防范

（2）事中监控

数据质量事中监控主要包括监控数据质量的问题、问题分析、数据质量问题的解决、重新分析数据质量问题、生成质量分析报告，关于事中监控的流程，如图 6-48 所示。

事中监控的主要目的是为了快速地解决数据质量问题。

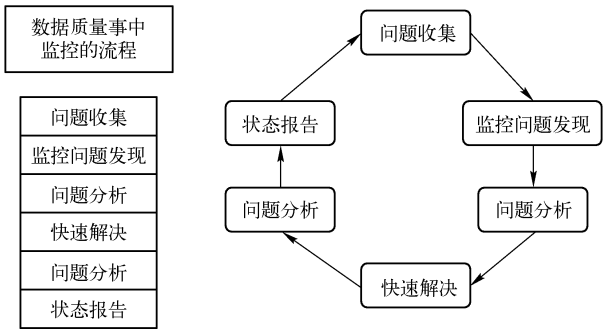


图 6-48　数据质量事中监控

（3）事后改进

数据质量事后改进包括问题的收集、质量问题分析、质量改进方案设计、方案审核、方案实施、方案效果评估。事后处理数据质量问题是对已经存在的质量问题进行优化和改进。相关人员可以包括数据质量管理员、业务数据质量负责人、技术数据质量负责人、数据质量主管，数据质量事后改进流程，如图 6-49 所示。

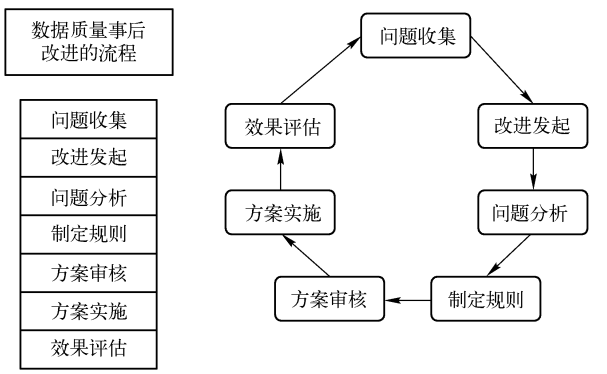


图 6-49　数据质量事后改进

二、数据质量管理的解决办法

数据质量管理的解决办法主要包括定义、发现、分析、反馈、整改、监控等。

- 定义

对数据质量问题进行分类，制定相关的检查规则。

- 发现

可以使用相关质量管理工具，根据检查规则去配置检查任务，从而发现问题。

- 分析

当发现问题后，对问题进行分析，判断是自身问题还是数据源的问题。

- 反馈

根据处理流程，由负责人将质量问题反馈至相关系统。

- 整改

由数据质量管理小组负责对问题的修改。

- 监控

数据质量管理系统对质量问题持续监控，保证数据的正确性，形成一个闭环结构，经过不断修正、循环，逐步提高数据的质量，如图 6-50 所示。

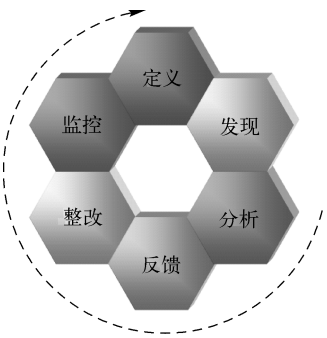


图 6-50　数据质量管理的解决办法

三、数据质量管理的执行

举例来说，某日，某银行员工张三发现数据质量问题，并把该问题记录下来，同时反馈到该银行的 IT 部门进行数据提升。

处理过程：由各机构进行结果反馈，对于不能提升的数据，由数据质量管理人员进行数据忽略，结果体现在系统中，如图 6-51 所示。

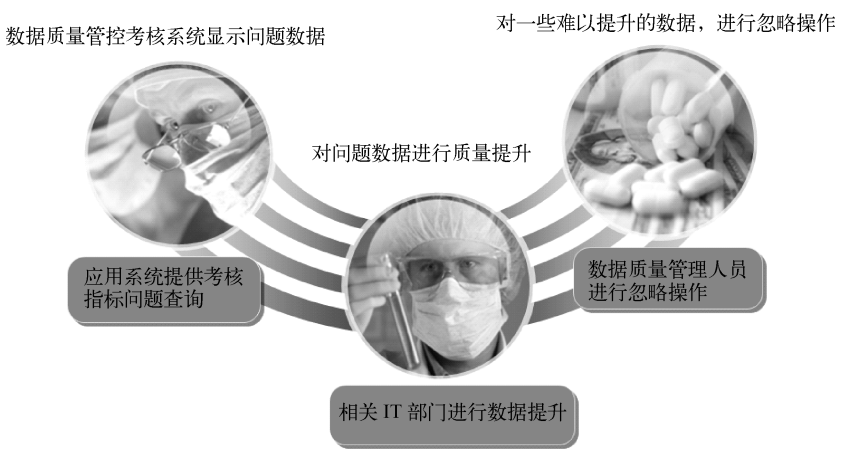


图 6-51　数据质量管理的执行

总结来说，在数据质量管理平台中建立数据质量监测体系，使得数据质量问题得到根本解决，最终形成数据质量闭环的提升流程，如图 6-52 所示。

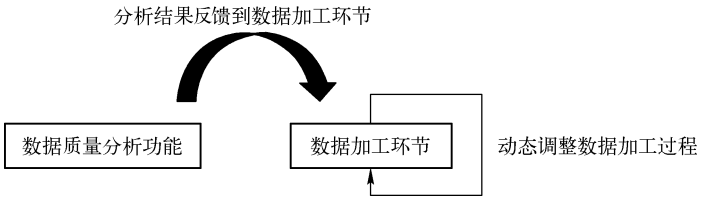


图 6-52　建立数据质量监测体系

6.4　元数据管理

6.4.1　元数据管理概况

元数据管理是管理企业的元数据信息，包括技术元数据、业务元数据和管理元数据。元数据管理的目标是获取、理解和共享企业的信息资产。

1. 元数据管理功能

元数据管理功能主要包括：

- 元数据采集

实现业务元数据的自动采集，完成技术元数据的自动匹配，实现对无法自动采集元数据信息的补录。

- 元数据查询

提供技术元数据、业务元数据和管理元数据的信息查询，支持对元数据的统计。

- 元数据版本管理

自动对元数据版本进行匹配，提供对元数据历史版本的查询和对比。

- 元数据分析

实现数据管理相关的数据分布地图、数据血缘分析和影响性分析等。

2. 元数据管理功能主要体现在以下几个方面：

（1）元数据采集

1）配置元数据采集器。

2）实现对业务元数据的自动采集。

3）自动匹配技术元数据的关联性。

（2）元数据自动补录

补录无法自动获取的元数据，对元数据信息进行修改和完善。

（3）元数据版本管理

1）对采集到的元数据信息进行版本比对。

2）对发生变化的元数据进行提醒。

3）支持历史各个版本元数据的查询。

（4）元数据查询

元数据查询包括数据映射、加工规则、数据标准信息、数据指标口径、数据分布等信息，如图 6-53 所示。

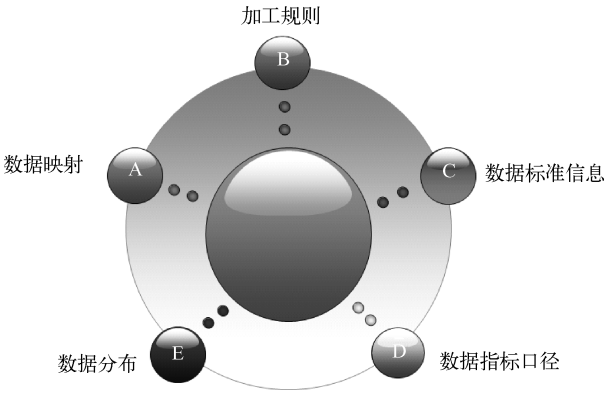


图 6-53　元数据查询

（5）元数据分析

元数据分析主要包括元数据血缘分析和元数据影响性分析。

1）元数据血缘分析。当发现报表中的指标有问题的时候，可以通过元数据的血缘分析追溯到该指标的下游系统到上游系统的转换流程中，帮助分析人员了解该指标的处理流程，为进一步定位问题提供帮助。血缘分析和影响性分析类似，但是方向相反。

触发血缘分析的方式：

① 通过查询找到变化的目标表，经过血缘分析，发现变化的表是由上游哪些源引发的。

② 如果在报表中发现某个指标有问题，可以进行血缘分析，分析该指标的数据加工过程，了解该指标出现问题的原因。

2）元数据影响性分析。在数据处理过程中，如果源系统的表结构或者属性发生变化，需要通过元数据的影响性分析，了解这些变化会影响数据处理流程中下游的哪些表结构或者属性。

3. 元数据管理的几个角色

● 元数据管理者

主要负责元数据收集、维护、录入，以及元数据版本管理、信息发布等工作。

● 元数据消费者

包括对元数据基础信息的查询、信息分析等。

元数据管理框架主要包括管理政策、组织、流程和技术工具。其中管理政策包括元数据管理方法、元数据管理实施细则；组织包括元数据管理组织架构、元数据管理岗位职责；流程包括元数据申请、元数据审批、元数据实施和推广，以及元数据维护；技术工具主要是元数据管理系统，如图 6-54 所示。

4. 数据标准和元数据的对比说明

1）从定义上来说，数据标准是经过相关机构确认和批准的规范性的文件，标准可以保障核心数据在使用和交换过程中的一致性和准确性。元数据是描述关于数据的数据，包括这些数据的定义、数据之间的关系等信息，可以分成业务元数据、技术元数据和管理元数据。

2）数据标准是为了更好地保障各个部门之间的数据共享。而元数据是对数据进行管理，方便数据检索，通过元数据的分析，更好地为数据分析人员服务。

图 6-54　元数据管理框架

3）数据标准的实施和落地需要业务部门和技术部门之间的合作，业务部门参考数据标准规范文档制定业务规则，技术部门在系统的建设过程中参考该规范文档进行设计和开发。元数据的实施和落地通过元数据管理平台对数据进行血缘分析和影响性分析。

4）数据标准主要是统一业务和技术定义，目的是消除企业内部人员对于业务和技术术语的分歧，它是一种规范性的文档。元数据管理是对数据结构的描述，并且提供数据管理和分析的功能。

6.4.2　元数据管理的设计方法和流程

元数据管理的设计方法和流程主要包括元数据管理总体规划、元数据管理的解决办法和元数据管理的执行，如图 6-55 所示。

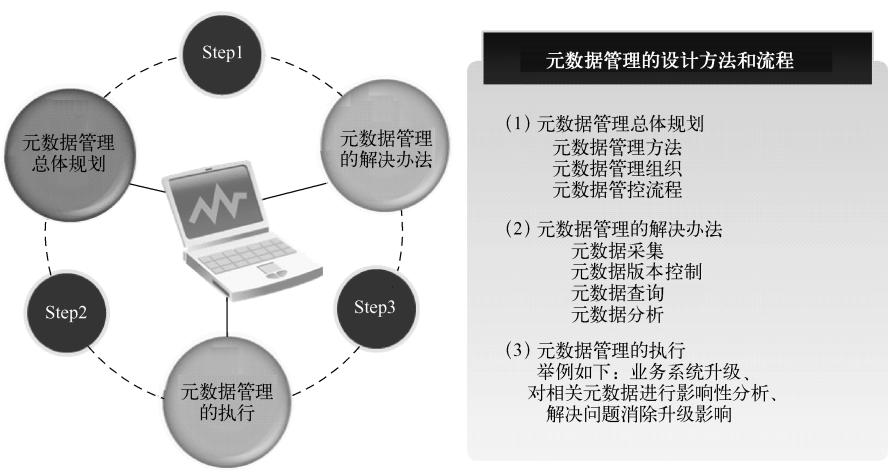


图 6-55　元数据管理的设计方法和流程

1. 元数据管理总体规划

（1）元数据管理方法

元数据管理方法主要是明确元数据管理的工作方向和参与元数据管理的部门。

（2）元数据管理组织

元数据管理组织主要包括数据管控办公室、数据责任人和系统责任人，如图 6-56 所示。

图 6-56　元数据管理组织

1）数据管控办公室包括元数据主管、元数据管理员。

2）数据责任人包括元数据负责人、数据录入人员和数据使用人员。

3）系统责任人包括元数据负责人和系统负责人。

（3）元数据管控流程

元数据管控流程包括元数据申请、审批与发布、实施与推广以及维护，如图 6-57 所示。

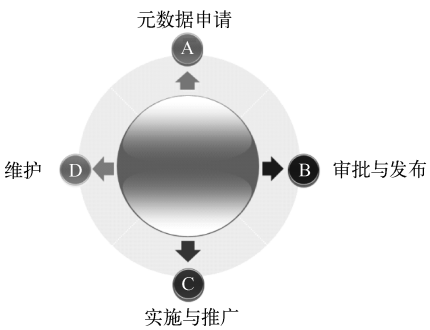


图 6-57　元数据管控流程

1）元数据申请。首先对业务元数据、技术元数据进行统一定义，形成版本。然后进行元数据新增、修改或者删除的申请，形成元数据的初稿。相关人员可以是业务元数据负责人、技术元数据负责人等。元数据申请流程主要包括元数据定义、提交申请、元数据导入，如图 6-58 所示。

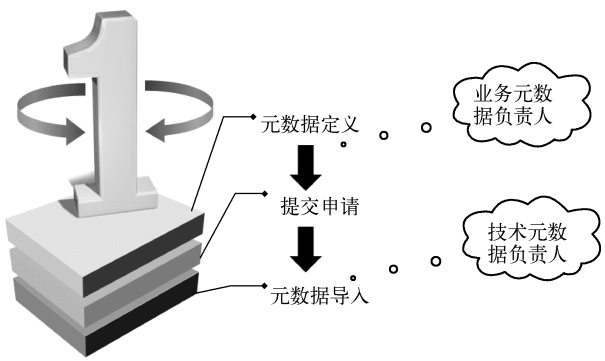


图 6-58　元数据申请流程

2）审批与发布。审批与发布的流程包括：对元数据进行分析、元数据审核和元数据发布。通过技术元数据获取数据的血缘关系，通过业务元数据获取相关业务文档，通过对元数

据的审核与发布确保元数据的真实性和完整性。一般来说，元数据主管进行审核，元数据管理员发布元数据。

元数据审批与发布流程如图 6–59 所示。

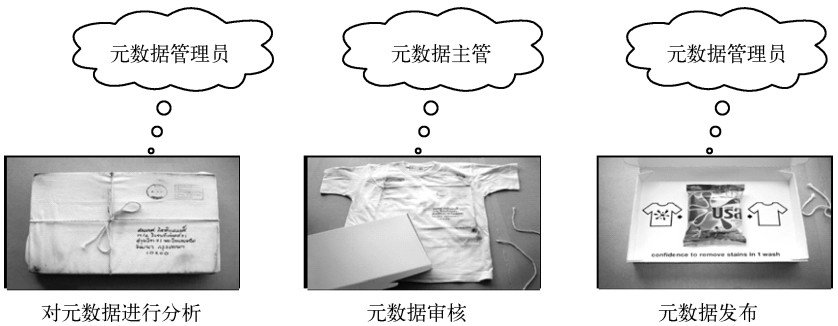


图 6–59　元数据管理审批与发布流程

● 对元数据进行分析

角色是元数据管理员。

● 元数据审核

角色是元数据主管。

● 元数据发布

角色是元数据管理员。

3）实施与推广。元数据的实施与推广是将元数据录入到管理平台进行实施和推广。相关人员可以包括元数据管理员、数据使用人员、系统负责人、数据录入人员等。

元数据实施与推广流程主要包括元数据查询、元数据使用和元数据反馈。

● 元数据查询

包括数据使用人员、系统负责人、数据录入人员、数据报送机构。

● 元数据使用

包括数据使用人员、系统负责人、数据录入人员、数据报送机构。

● 元数据反馈

包括元数据管理员。

4）维护。元数据维护流程包括：元数据的评估规划、元数据评估和元数据的变更。

例如，首先对元数据的使用情况进行评估，监测元数据在系统中的使用情况，考察相关的实施结果，提交使用分析报告，对相关情况进行总结。

然后提出更正或者注销申请，及时更正元数据内容，最后形成闭环的元数据管理流程。相关人员可以包括元数据管理员、业务元数据负责人和技术元数据负责人。

● 元数据的评估规划

包括元数据管理员。

● 元数据评估

包括业务元数据负责人和技术元数据负责人。

● 元数据的变更

包括业务元数据负责人和技术元数据负责人。

2. 元数据管理的解决方法

元数据管理的主要目的是为数据的有效利用提供全面的指导。通过元数据管理，可以建立数据的统一视图和统一口径，确保数据的完整性、准确性、一致性。

元数据管理功能包括元数据采集、元数据版本控制、元数据查询、元数据分析，如图 6-60 所示。

1）元数据的采集。采集的内容包括：技术元数据，如 ETL 映射关系，数据结构，数据字典等内容；业务元数据，如代码标准、指标标准等信息。如果无法自动采集元数据信息，则进行信息的补录，或者对元数据信息进行修改和调整。

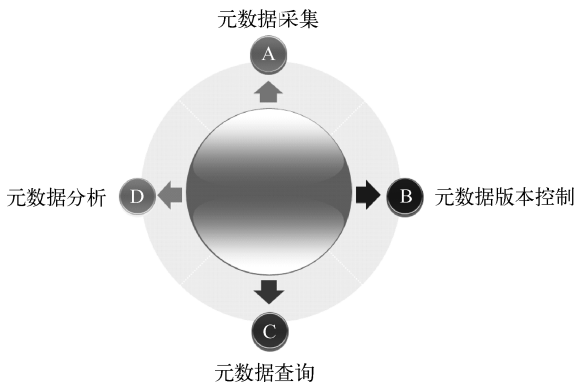


图 6-60　元数据管理功能

2）对元数据版本进行管理。对于采集的元数据信息进行版本对比，对于发生变化的元数据进行提醒，并且保留每个历史版本的元数据信息。

3）元数据查询可以提供对技术元数据和业务元数据的信息查询。

4）支持对元数据的统计分析。例如，实现数据血缘分析和影响性分析。

3. 元数据管理的执行

对于元数据管理的业务场景之一，举例如下：因业务升级，在“XXX 表”中增加了科目 YYY，需要找到本次升级后对相关系统的影响。

处理结果如下：例如，对元数据“XXX 表”进行影响性分析，发现对 ODS、报表指标都有影响。

它的过程如图 6-61 所示，在业务系统升级的时候，对相关元数据进行影响性分析，最后解决问题，消除升级影响。

图 6-61　元数据管理的执行过程一

对于元数据管理的业务场景之二，举例如下：某报表系统运维人员李四发现余额中的“金额”结果异常，因此，把问题反馈给元数据管理系统的高级分析员。

处理过程：由高级分析员登录到元数据管理系统，对报表系统余额中的“金额”进行血缘分析，然后再对问题进行定位。

总结来说，元数据管理平台在业务层面上帮助业务人员了解数据的定义，辅助数据标准的建设，解决业务定义不一致的问题，同时也帮助技术人员了解数据来源和数据加工规则，从而有效地提升开发效率，降低数据的复杂性，解决数据的冲突问题。通过分析数据的血缘和影响，找出问题产生的原因和影响范围。

6.5 数据生命周期管理

6.5.1 数据生命周期管理概况

一、什么是数据生命周期管理

数据生命周期管理是对数据进行统一管理，目的是降低数据的存储压力。一般来说，数据生命周期管理包括数据创建、数据使用、数据归档和数据销毁，如图 6-62 所示。

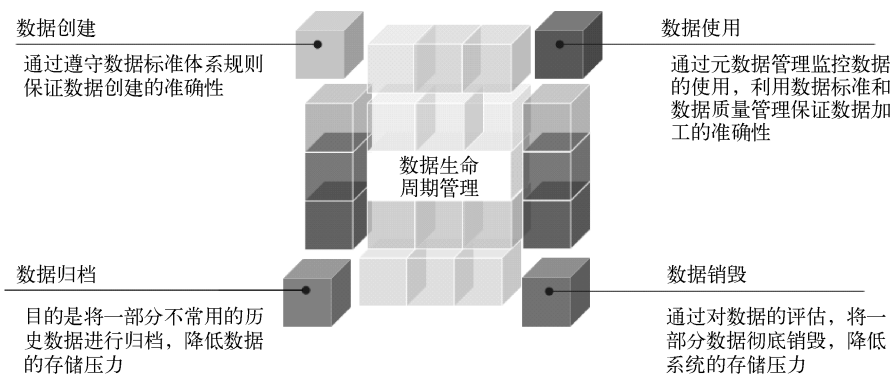


图 6-62　数据生命周期包含的内容

- 数据创建

通过遵守数据标准体系规则保证数据创建的准确性。

- 数据使用

通过元数据管理监控数据的使用，利用数据标准和数据质量管理保证数据加工的准确性。

- 数据归档

目的是将一部分不常用的历史数据进行归档，降低数据的存储压力。

- 数据销毁

通过对数据的评估，将一部分数据彻底销毁，降低系统的存储压力。

二、数据生命周期管理框架

数据生命周期管理框架主要包括数据生命周期的管理政策、组织、流程和技术工具，如图 6-63 所示。

- 管理政策

包括数据生命周期管理办法、数据生命周期管理实施细则。

- 组织

包括数据生命周期管理角色定义、数据生命周期管理角色责任。

- 流程

包括数据生命周期管理方案规则、数据生命周期管理方案实施和对具体问题的解决。

- 技术工具

包括数据生命周期管理系统。

图 6-63　数据生命周期管理框架

6.5.2　数据生命周期管理的设计方法和流程

数据生命周期管理的设计方法和流程包括：数据生命周期管理总体规划、数据生命周期管理的解决办法和数据生命周期管理的执行，如图 6-64 所示。

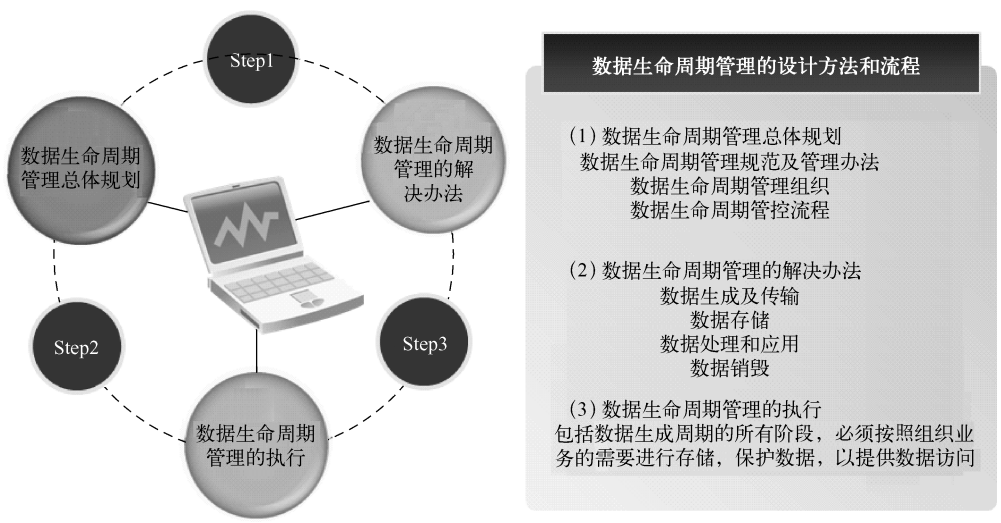


图 6-64　数据生命周期管理的设计方法和流程

1. 数据生命周期管理总体规划

（1）数据生命周期管理规范及管理办法

数据生命周期管理规范及管理办法包括相关制度规范和管理办法。

- 规范制度

明确数据生命周期管理的组织体系。明确各组织在数据生命周期管理工作中应该承担的角色与职责。明确数据生命周期划分阶段。

- 管理办法

确定数据生命周期的组织机构和各组织应该承担的工作职责。

（2）数据生命周期管理组织

数据生命周期管理组织包括数据管控办公室、数据责任人和系统责任人，如图 6-65 所示。

图 6-65　数据生命周期管理组织

- 数据管控办公室

主要包括数据生命周期主管、数据生命周期管理员。

- 数据责任人

主要包括数据生命周期负责人。

- 系统责任人

主要包括数据生命周期负责人、系统负责人和系统运维人员。

其中数据生命周期管理组织的角色主要有两种：数据生命周期主管和数据生命周期管理员，如图 6-66 所示。

（3）数据生命周期管控流程

数据生命周期管控流程包括数据生命周期管理方案规划、数据生命周期管理方法实施和落地、对具体问题的解决，如图 6-67 所示。

- 数据生命周期管理方案规划

图 6-66　数据生命周期管理组织角色

图 6-67　数据生命周期管控流程

数据生命周期管理规划是由企业的业务人员和相关系统运维人员提交数据生命周期的需求，再由管理人员设计并制定数据生命周期的管理策略。与数据生命周期管理相关的人员包括数据生命周期负责人、系统运维人员、数据生命周期管理员、数据生命周期主管等。

● 数据生命周期管理方法实施和落地

数据生命周期管理方法实施和落地的相关人员主要包括：系统负责人、系统运维人员、数据生命周期管理员等，职责是实施数据生命周期管理方案，评估方案的优劣等内容。

● 对具体问题的解决

具体问题的解决流程是针对出现的问题进行分析，如存储问题，然后提出具体解决办法，制定数据生命周期管理方案，当审核通过后执行该管理方案。例如，由数据生命周期管理员组织相关人员实施或评估数据生命周期的管理策略，人员可以包括系统运维人员、业务数据生命周期负责人、数据生命周期管理员、技术数据生命周期负责人、数据生命周期主管等。

数据生命周期管理具体问题的解决流程主要包括申请、制定方案、审核、执行和变更管理方案。

2. 数据生命周期管理的解决方法

数据生命周期管理涉及数据从开始创建一直到失去商业价值，最后按照规定被删除的过程。一般来说，它有如下几个过程：数据创建、数据使用、数据归档和数据销毁。

3. 数据生命周期管理的执行

数据生命周期的所有阶段，按照业务需求存储数据，以及提供数据的访问。数据生命周期的每一阶段，根据数据的价值，存在不同层次的性能、可用性、保护和处理。这样保证数据的可用性，也充分满足业务的需求。

在数据生命周期的各个阶段，初始数据的生成需要高速地进行存储，并且提供相应的保护措施，已达到高可用性。但是随着时间的推移，数据的重要性会逐渐降低，使用频率也会下降。根据这些变化，数据的存储、可用性、性能和保护措施在力度上也随着发生变化。

总结来说，通过建设数据生命周期管理系统，规范数据存储以及数据生命周期管理，提高系统运行的效率，为生命周期规范提供支撑。协助各个业务系统实现统一的数据归档管理，降低业务系统的复杂度，缩短业务系统建设的周期，避免因为系统重复开发而造成的资源浪费，节省各个业务系统的资源。实现更快、更方便的数据备份、恢复和升级，减少系统停机的时间。将不经常使用的数据转移到存档的基础设施中，以降低物理存储总量，节省硬件和维护成本。

小结

- 一般来说，数据是企业的原始材料，也是金融、电信、互联网等行业最大的价值来源之一，如何利用这些数据，以及如何更好地对数据进行挖掘和利用，已经成为提高企业竞争力最重要的手段之一。
- 数据治理是一套包含策略、原则、组织结构、管理制度、流程以及各种相关技术工具的管理框架。数据治理是对数据管理与应用行使权力和控制的活动集合，在数据管理与应用层面上进行规划、监督和控制，数据治理为数据管理、数据应用与服务提供保障。
- 数据治理可以看做是一门新的学科，能够把企业的独立系统结合起来，重新定义数据的价值和保护机制。从技术上来讲，数据治理是从 OLTP 系统到后台业务数据库，再回到前端的一个闭环的过程。
- 一般来说，数据治理可以分成两个部分：

1）数据的保障机制，包括政策的制定，考虑使用何种机制、流程和工具去保障数据的规范性。

2）需要考虑数据的质量标准和数据质量的任责体系。数据治理是企业的责任，需要统一的解决方案和治理模型来保护及共享不同层面的数据。

- 数据治理建设的关键要素：以数据标准为基础、以提高数据质量为核心、明确数据治理的职责。
- 对于数据治理体系的框架结构，可以包括规划、机制、治理对象和实现 4 个部分。
- 数据治理是保障企业和商业银行安全、稳定运营的必要条件，特别是对商业银行来说，如何避免数据的泄露、篡改，保证数据的一致性和完整性，这才是实现商业银行业务连续性的关键。
- 数据治理对商业银行等金融机构尤为重要：数据作为商业银行或者企业的重要资产，相当于人体的血液一样，是非常重要的。高质量的数据，有利于管理决策层做出准确

的分析。数据治理有利于保护核心的业务数据。

- 数据标准是一套完整的数据规范，是数据在使用和交换过程中，为了保持数据一致性和准确性而制定的规范，它主要包括数据分类、业务标准和技术标准的详细定义。数据标准是数据治理中基本的业务和技术层面的保障。
- 数据标准的体系框架可以包括：文化和战略，数据标准内容，数据标准制度和流程，数据标准的组织和角色，数据标准工具。
- 数据标准体系建设的规划方法可以遵循业界先进的方法，通过调研、规划访谈、数据标准现状分析，了解业务部门对数据标准的期待，确认业务部门对数据标准的想法，将对数据标准的需求转化成业务人员可以理解的文档，建立数据标准管理相关的治理架构和管理流程，同时建立企业对数据标准管理的共识和实施路线图。
- 数据标准规划的过程如下所示：对现有系统的数据标准进行梳理；建立公共代码的数据标准；通过公共代码数据标准的建立，为系统提供服务。
- 数据质量管理可以通过提高管理水平，严格执行相关的政策和规范，或者使用一些技术工具，使得数据质量得到进一步的提升。对于质量管理来说，它是一个闭环的管理过程，经过不断循环、改善，逐步提高数据的质量，并最终为企业赢得经济效益。
- 数据质量管理的设计方法和流程包括数据质量管理总体规划、数据质量管理的解决办法、数据质量管理的执行等。
- 在数据质量管理平台中建立数据质量监测体系，使得数据质量问题得到根本解决，最终形成数据质量闭环的提升流程。
- 元数据管理是管理企业的元数据信息，包括技术元数据、业务元数据和管理元数据。元数据管理的目标是获取、理解和共享企业的信息资产。
- 元数据管理功能主要体现以下几个方面：

1）元数据采集（配置元数据采集器；实现对业务元数据的自动采集；自动匹配技术元数据的关联性）。

2）元数据自动补录（补录无法自动获取的元数据，对元数据信息进行修改和完善）。

3）元数据版本管理（对采集到的元数据信息进行版本比对；对发生变化的元数据进行提醒；支持历史各个版本元数据的查询）。

4）元数据查询，包括数据映射、加工规则、数据标准信息、数据指标口径、数据分布等信息。

5）元数据分析，包括元数据血缘分析和元数据影响性分析。

- 元数据管理平台在业务层面上帮助业务人员了解数据的定义，辅助数据标准的建设，解决业务定义不一致的问题，同时也帮助技术人员了解数据来源和数据加工规则，从而有效地提升开发效率，降低数据的复杂性，解决数据的冲突问题。通过分析数据的血缘和影响，找出问题产生的原因和影响范围。
- 数据生命周期管理是对数据进行统一管理，目的是降低数据的存储压力。一般来说，数据生命周期管理包括：数据创建、数据使用、数据归档和数据销毁。
- 数据生命周期管理总体规划包括：

1）数据生命周期管理规范及管理办法。

2）数据生命周期管理组织。

3）数据生命周期管控流程。

- 数据生命周期的所有阶段，企业按照业务需求存储数据，以及提供数据的访问。数据生命周期的每一阶段，根据数据的价值，存在不同层次的性能、可用性、保护和处理。这样才能保证数据的可用性，也充分满足业务的需求。
- 通过建设数据生命周期管理系统，规范数据存储以及数据生命周期管理，提高系统运行的效率，为生命周期规范提供支撑。协助各个业务系统实现统一的数据归档管理，降低业务系统的复杂度，缩短业务系统建设的周期，避免因为系统重复开发而造成的资源浪费，节省各个业务系统的资源。实现更快、更方便的数据备份、恢复和升级，减少系统停机的时间。将不经常使用的数据转移到存档的基础设施中，以降低物理存储总量，节省硬件和维护成本。

第 7 章　商业智能架构理论

本章目标

通过前几章的学习，我们了解了数据架构、大数据和数据治理相关的知识和案例。很多企业已经充分认识到数据是核心资产和竞争力。同时为了提高企业的运营效率，增加企业的竞争力和领导者的决策能力，系统应该适应多渠道数据采集的能力，形成汇总功能型的视图。增强历史与趋势分析能力，这就需要 IT 人员理解商业智能方面的知识。

学习本章后，读者将掌握：

- 商业智能的历史
- 商业智能的定义
- 商业智能的功能
- 商业智能的发展趋势
- 商业智能的实施方法和步骤
- 关于商业智能的核心技术
- 数据仓库理论
- 数据仓库的特点
- 数据挖掘和分析
- ETL 处理技术
- 数据集市理论
- 数据集市产生原因
- 数据集市的定义
- 数据集市和数据仓库的联系和区别
- 可视化分析
- 大数据技术
- ODS 理论
- OLAP 系统与 OLTP 系统的区别
- OLAP 的实现方法
- OLAP 模型的设计与实现

7.1　商业智能概述

7.1.1　商业智能的历史

- 1970 年，IBM 公司的研究员埃德加·科德发明了关系型数据库。
- 1979 年，Teradata 公司诞生。1983 年，该公司利用并行处理技术为美国富国银行建立了第一个决策支持系统。

- 1988 年，IBM 公司的研究员提出一个新的概念：数据仓库。
- 1992 年，比尔·恩门出版了《如何构建数据仓库》一书，数据仓库真正拉开了应用的序幕。
- 1993 年，拉尔夫·金博尔出版了《数据仓库的工具》一书，并把部门的数据仓库叫做“数据集市”。

7.1.2 商业智能的定义

从全球范围来看，商业智能已经成为目前最具有发展前景的 IT 领域之一。

曾经看过这样一个例子，美国某超市有一个系统：当你采购了一车的物品准备结账时，美丽的收银员小姐扫完了你的所有物品后，计算机会显示出一些信息，然后收银员小姐会友好地问你：“我们有一种一次性纸杯正在促销，位于 ×× 货架上，您要购买吗？”结果你非常惊奇地说：“啊，谢谢你！我刚才一直没有找到纸杯。”那么计算机系统如何知道的？秘密在于当系统知道你的购物车里面有餐巾纸、大瓶可乐和沙拉的时候，则会计算出你买一次性纸杯的可能性在 80% 以上。这就是商业智能的一个简单应用。

再举一个例子，智能手机可以内嵌全球卫星定位系统，通过该系统，我们可以找到最近的银行网点，并且可以预约排队。同时银行可以分析出客户的喜好，向客户推送附近可以刷卡打折购物的信息，客户也可以享受到各种实用的银行服务。换句话说，银行可以利用商业智能为我们提供各种智能化和个性化的服务，如图 7-1 所示。

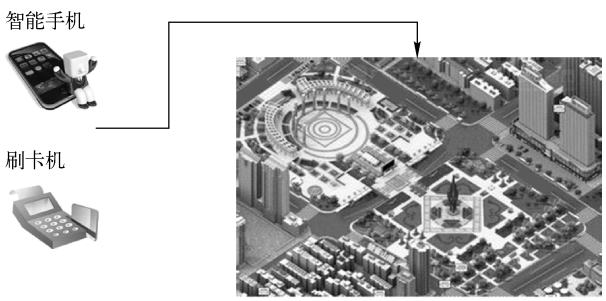


图 7-1　商业智能提供各种智能化和个性化的服务

当然，商业智能的作用绝不仅限于此。从小型的超市系统到国家银行、航空、水利、电力、铁路运输等大型系统，商业智能的应用无处不在。如果我们对商业智能做一个简单的定义，那就是：帮助用户把一些数据转化成具有商业价值的，而且可以获取的信息和知识，同时在最恰当的时候，通过某种方式把信息传递给需要的人。从专业的角度来说，商业智能就是利用数据仓库、数据分析和挖掘技术，以抽取、转换、查询、分析和预测为主的技术手段，帮助企业完成决策分析的一套解决方案。

在上面的例子中，计算机系统把餐巾纸、大瓶可乐、沙拉等商品信息转化成具有商业价值的信息（知识），同时在恰当的时候把顾客需要一次性纸杯的信息告诉收银员。商业智能的价值体现在将数据转化成信息和知识，最后转化成利润，如图 7-2 所示。

很多企业在经过多年的业务系统的运行之后，已经拥有了大量的经营数据，那么如何将这些宝贵的数据财富转化成信息、知识并传递给企业管理者呢？这就是商业智能需要研究和完成的工作。

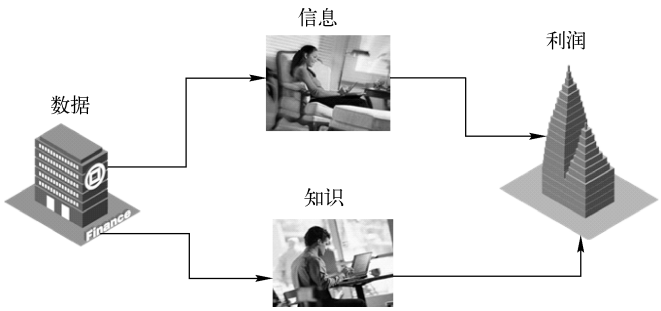


图 7-2　商业智能的价值

商业智能好像一个采矿加工场，它负责采集大量的矿石，然后经过进一步的分离、加工等操作，最后提炼出高纯度的精矿，如图 7-3 所示。其实企业经营和管理的数据就是这些“矿石”，而商业智能的作用就是将这些“矿石”转化成“精矿”。

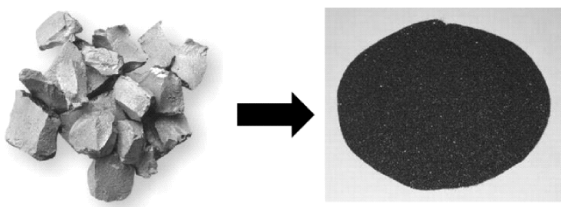


图 7-3　采矿加工场

商业智能对企业的重要性是不言而喻的，它可以提高企业的运营效率，增加竞争力和领导决策能力，从而获得更大的市场，提高企业的利润。同时也为公司的管理人员提供一种全新的思维方式，通过使用这些宝贵的数据资产进行挖掘和分析，发现内部潜在的规律和趋势，这样才能做出准确的判断，制定出正确的决策方针。此外，还优化了企业内部组织结构，增强了企业资源的合理配置，使企业在竞争中处于不败之地。

7.1.3　商业智能的功能介绍

商业智能最早出现在 20 世纪 90 年代，当时的主要功能是查询报表、数据分析、数据备份和恢复等，但随着技术的发展和应用的拓展，商业智能已经扩展了其他的功能，如图 7-4 所示。

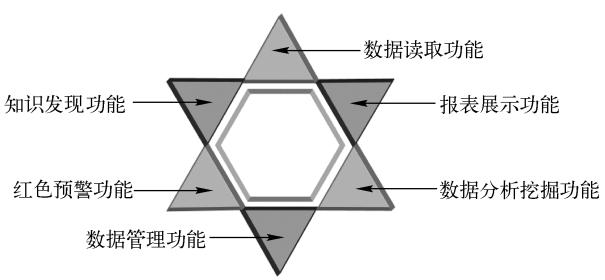


图 7-4　商业智能的功能介绍

（1）数据读取功能

除了读取结构化数据，还可以读取非结构化数据和半结构化数据。

（2）报表展示功能

例如，利用报表工具（Cognos、BO 等）的可视化功能将数据呈现给用户，呈现的形式包括：交叉报表、饼图、柱状图、散点图、线图、直方图。其中柱状图示例如图 7-5 所示。还可以通过向下钻取、数据切片和旋转以及交互式的图形分析能力，使用户能够从任何角度去观察业务。

（3）数据分析挖掘功能

通过业务之间的关联关系，去探究事物发生的概率。

（4）知识发现功能

知识发现是从大量的数据中提取人们感兴趣的知识的能力，这些知识可以是隐含的、事先未知的或者潜在有用的信息，提取的知识表示为概念、规则、规律和模式等形式。

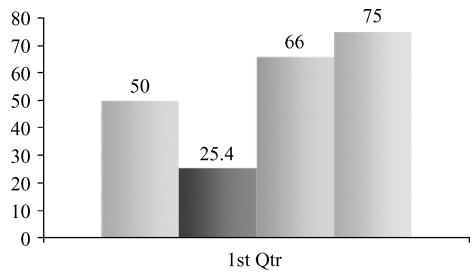


图 7-5　柱状图

（5）红色预警功能

可以基于数据仓库提供预警的功能。

（6）数据管理功能

管理功能是从多个数据源抽取、转换和加载，以及清理和集成数据的能力，包括高效的存储与维护的能力。

7.1.4　商业智能的发展趋势

随着互联网的普及和技术的进步，商业智能的发展也成了不可阻挡的趋势，企业可以通过互联网信息的收集，获取更多的客户信息和交易信息，通过商业智能技术的应用，将这些数据转化成更有价值的信息，帮助企业的高层做出准确的分析和决策。

商业智能除了帮助企业管理人员做出准确的分析和决策，还可以为客户提供各种个性化的服务。例如，通过客户的特征和以往的交易情况，分析出客户的购买力和喜好，从而进行有针对性的营销。这不仅给商家带来直接的经济利益，同时也可以帮助客户在最短的时间内购买到最需要的商品。商业智能的发展必然通过 Web 和局域网的交互，实现信息和知识的共享。

目前随着商业智能技术的发展，增强了对非结构化数据的处理能力。以前商业智能处理的数据还是以结构化的信息为主，也就是存储在内部数据库中的数据和文本。而现在，越来越多的企业已经将各种非结构化数据当做主要的数据源，例如各种客户的呼叫记录、影像资料、音频资料、文本、图片和各种电子邮件等。

随着移动互联网的发展，大大提高了对金融数据的收集能力，包括用户的交易数据和行为数据。金融服务的多样化和市场规模的不断扩大，需要对这些数据进行深度挖掘和分析，匹配金融产品的交易需求，发现隐藏的趋势信息，让金融机构发现商机。

为了实现经济快速发展的目标，很多制造、能源企业必将大力发展商业智能技术，加大对商业智能解决方案的投入，从而降低生产成本，提高资源利用率和市场占有率，使其生产运营能够健康平稳的发展。

7.1.5　商业智能的实施方法和步骤

1. 商业智能的实施方法

商业智能的实施方法包括项目规划、系统设计与实现、系统调优以及系统运行及维护，如图 7-6 所示。

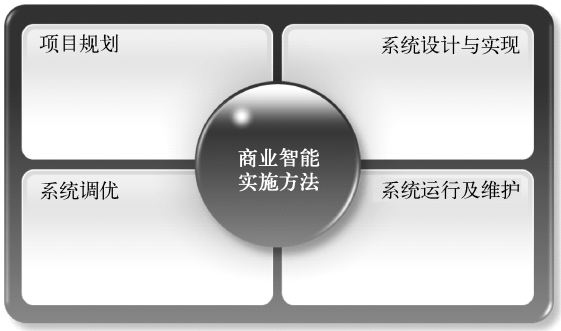


图 7-6　商业智能的实施方法

（1）项目规划

项目规划主要包括项目前期的准备、业务现状的调研、目前系统的现状分析。分析内容包括业务需求的定义和系统实现的目标，系统运行环境的定义，系统的框架结构定义，逻辑模型的设计等。

（2）系统设计与实现

系统设计与实现主要包括系统体系结构的设计，物理数据库的设计，数据抽取、转换和加载的实现，前端应用的开发，元数据的管理等内容。

（3）系统调优

系统调优主要指逻辑、物理模型的调整，系统性能的调优。

（4）系统运行及维护

系统运行及维护主要指编写系统运行及维护手册，以及用户操作手册、培训教材等文档。

2. 商业智能的实施步骤

商业智能的实施步骤包括定义需求，数据仓库模型的建设，数据抽取、清洗、转换、加载，建立商业智能分析报表，如图 7-7 所示。

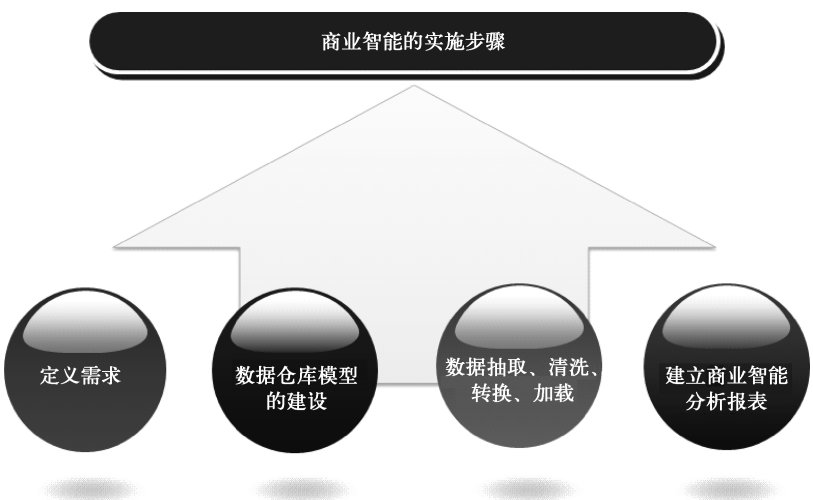


图 7-7　商业智能的实施步骤

（1）定义需求

需求分析是商业智能项目重要的一步，需要描述项目背景与目的、业务范围、业务目标、业务需求和功能需求等内容，明确企业对商业智能的期望和需要分析哪些主题等方面。其中项目背景主要描述已有系统的当前现状是什么，以及不同的历史时期，它的业务需求分别是什么。这些独立的信息系统特点一般是缺乏统一的整体规划和标准，数据分散，每个业务之间不能共享信息，报表展示功能单一，各业务系统之间存在数据不一致的现象，企业领导层无法从全局的角度对业务进行综合分析。

商业智能项目最重要的目的之一是解决各个业务系统之间数据集中整合的问题，为企业管理人员提供高效的数据查询和强大的报表展示功能，同时能够进行多维度的深入分析和数据挖掘，为企业未来的经营状况做出准确的预测。

业务范围是对项目团队所有人员工作范围的界定。

业务需求是描述客户对于系统实现的总体性要求，商业智能项目的特点是从不同的维度去分析各个主题，以报表的形式对业务进行阐述。功能需求可以包含：各个业务专题分析、关键性指标查询和监控、报表查询、高级分析和数据挖掘等内容。

商业智能的功能框架如图 7-8 所示。

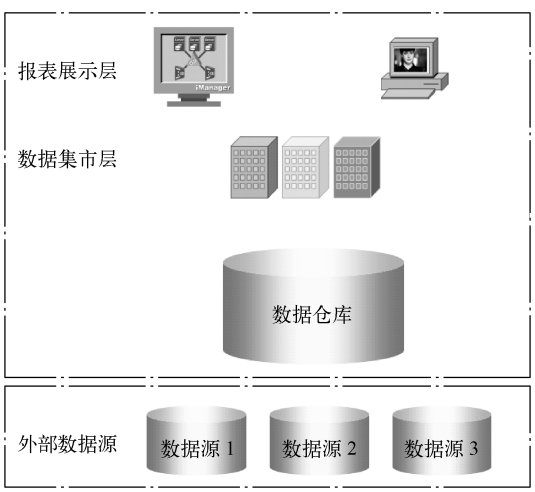


图 7-8　商业智能功能框架图

（2）数据仓库模型的建设

模型是对现实世界的抽象。数据仓库模型是在需求分析的基础上建立起来的。数据模型的设计流程是：在系统设计、开发之前，业务人员和设计人员共同参与概念模型的设计，核心的业务概念在业务人员和设计人员之间达成一致。在系统设计开发时，业务人员和系统设计人员共同参与逻辑模型的设计。最后，设计开发人员以逻辑模型为基础进行物理模型的设计。

（3）数据抽取、清洗、转换、加载

● 数据抽取

抽取主要负责将数据仓库需要的数据从各个业务系统中抽取出来。如果每个业务系统的数据情况各不相同，可能对每个数据源都需要建立独立的抽取流程，每个流程都需要使用接

口将源数据传送给下一环节，即清洗与转换阶段。通过数据抽取程序，可以从业务源系统中不断地将数据抽取出来，抽取周期可以设定为某个固定时间，例如每天中午 12 点对源数据进行抽取，也可以设定为某个时间间隔，例如每 6 个小时抽取源数据一次。

● 数据清洗

清洗阶段是对业务源数据的清洗和确认，检查抽取的源数据质量是否达到数据仓库的规定标准。数据清洗大致有两种方式：① 不同业务系统间各自专用的清洗程序；② 不同业务系统间有满足数据仓库清洗需求的通用程序，从不同业务系统抽取的数据有可能存在数据不一致的情况，可以使用相关规则和标准检查业务源数据的质量。

● 数据转换

转换是对源系统的数据在最后一步进行的修改，包括对源数据的聚合以及各种计算，是整个 ETL 过程的核心部分。

● 数据加载

加载是将数据加载到最后的目标表中，其复杂度没有转换高，一般采用批量装载的形式。

（4）建立商业智能分析报表

商业智能分析报表通过对数据仓库的数据分析，使企业的高层领导可以多角度地查看企业的运营情况，并且按照不同的方式去探查企业内部的核心数据，从而更好地帮助企业决策人员对公司未来经营状况进行预测和判断。

7.1.6 商业智能项目成功的关键

商业智能项目成功的关键因素如下。

1）企业高级领导层对商业智能项目的支持和雄厚的资金是项目成功的关键因素之一。

2）拥有实力雄厚的技术团队。技术团队成员不仅精通商业智能相关技术，同时也熟悉相关的业务规则和开发流程。

3）商业智能项目团队的协同合作能力。项目的管理者需要保证团队中每个成员分工明确，沟通及时，并且需要各部门之间有良好的合作能力。总之，商业智能项目的实施是一个长期的不断完善的过程。

7.1.7 关于商业智能的核心技术

商业智能实质上是数据转化成信息和知识的过程。构建一个完整的商业智能系统需要以下几种核心的技术：数据仓库、数据挖掘和分析、ETL 处理技术、联机分析处理（OLAP）技术、可视化分析、大数据技术、商业智能元数据管理，如图 7-9 所示。

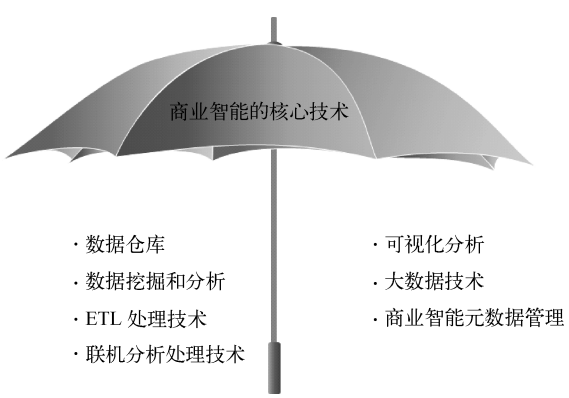


图 7-9 商业智能的核心技术

1. 数据仓库

数据仓库之父——比尔 · 恩门在《如何构建数据仓库》一书中将数据仓库

定义为：“数据仓库是在企业管理和决策中面向主题的、集成的、时变的 、非易失的（不可修改的）数据集合”。实质上，数据仓库是对数据处理技术的集成，它是为了进一步挖掘数据资源，为了决策分析而产生的。数据仓库的目的是为了前端报表查询和决策分析。

数据仓库与传统数据库的区别是：传统数据库主要用于企业日常的事务处理，而数据仓库主要用于商业分析，在不影响日常业务处理的前提下，辅助企业高层进行商业决策。

最终用户对数据仓库的访问方式包括：即席查询、报表、联机分析处理（OLAP）、数据挖掘，如图 7-10 所示。

图 7-10　数据仓库的访问方式

2. 数据挖掘和分析

数据挖掘（DataMining）起源于 1989 年 8 月，出自在美国底特律举办的第 11 届国际联合人工智能学术会议中 Piatetsky - Shapiro 提出的 KDD（Knowledge Discovery and DataMining）。数据挖掘是指从海量的数据中抽取出有意义的、重要的和潜在有用的信息和知识的过程。从技术上来说，数据挖掘是一门交叉学科，融合了统计学、人工智能、模式识别、机器学习等内容。

数据挖掘的工作过程可以包括数据的抽取、存储管理、挖掘和展现等几个部分，如图 7-11 所示。

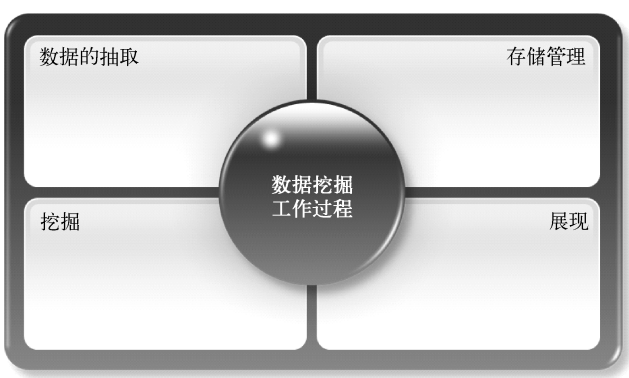


图 7-11　数据挖掘的工作过程

● 数据的抽取

所谓抽取就是将数据从外部数据源或者其他联机事物处理系统中导入到数据仓库或者其他数据库中。

● 存储管理

存储管理主要针对如何管理海量的数据、优化查询效率和处理各种并发数据等。

● 挖掘

挖掘就是利用各种的挖掘算法得到相应知识的过程。

● 展现

最后的数据展现就是实现各种的预定义查询、动态报表查询等内容，展示的方式包括各种的直方图、动态模拟和饼图等。简单地说，数据挖掘就是将对数据的简单查询提升到挖掘信息和知识的过程。

数据挖掘和分析主要用于从大量的数据中发现背后隐藏的规律和数据间的关系。采用数据挖掘技术，可以为用户提供自动化和智能的辅助决策分析。特别是在金融行业、零售业和医疗卫生领域，都有大量的应用。

在数据挖掘技术中，常用的模型有：分类模型、关联模型、顺序模型和聚簇模型，如图 7–12 所示。

（1）分类模型

根据商业数据的属性将数据分配到不同的组中。

（2）关联模型

主要描述一组数据项目的密切度和关系。

（3）顺序模型

主要用于汇总数据中的常见顺序或事件。顺序模型可以看成是一种特殊的关联模型，它在关联模型中增加了时间属性。

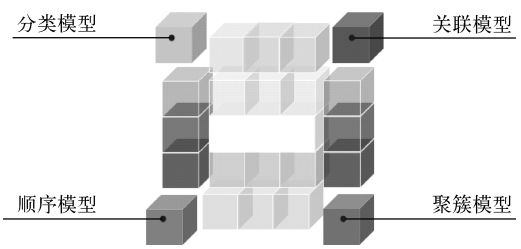


图 7–12　数据挖掘技术中常用的模型

（4）聚簇模型

按照某种相近程度将数据分成一些组。组中的数据相近，组之间的数据相差较大。

数据挖掘是一个闭环的、反复循环的过程，需要业务分析人员、IT 工程师共同完成。一般来说，它有以下几个步骤：

1）对业务范围的定义，在这个阶段需要明确对数据挖掘的目标和定位，制定数据挖掘的计划。

2）选择合适的数据，定义相关的训练数据集和验证数据集等内容。

3）对数据进行探索分析，使数据集尽可能满足建模算法的要求。

4）分析并且确定数据挖掘模型。建模人员需要不断地测试模型性能，从而选择最佳的数据模型。

5）模型实施和评价。通过模型实施的结果帮助相关人员做出战略决策。同时收集结果反馈，判断是否需要改进模型。

我们可以引用商业智能的概念。决策人员以企业级数据仓库为基础，利用联机分析处理工具、数据挖掘工具，加上决策人员的专业知识，从数据中获得有用的信息和知识，帮助企业获取利润，而数据挖掘就是建立在数据仓库基础上的增值技术。

数据仓库和数据挖掘之间的关系如图 7–13 所示。

数据仓库是为了支持企业决策分析的数据集合。它是面向主题的、集成的、稳定的，并

且随时间发生变化。它的关键技术包括数据的抽取、清洗、转换、加载和维护技术。

数据挖掘是从海量的数据中，抽取出有意义的、潜在有用的信息和知识的过程。数据挖掘的数据来源可以是数据仓库或者其他数据库。对于挖掘的数据需要进行选择，挖掘的结果需要进行评估，按照评估结果的不同，一般需要重新分析和计算。

数据挖掘可以对数据仓库中的历史数据进行提炼和挖掘，使得这些数据成为信息和知识。可以借助对历史数据的分析，发现数据内部有价值的规律。

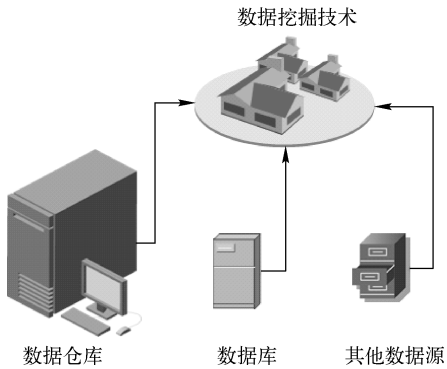


图 7-13　数据仓库和数据挖掘之间的关系

数据仓库是数据挖掘的基础。因为数据仓库的数据是完整的、集成的，所以它为数据挖掘提供了扎实的数据基础。数据仓库可以为数据挖掘提供需要的历史数据和全面的数据处理、分析等基础设施。

3. ETL 处理技术

ETL 即数据抽取（Extract）、转换（Transform）、装载（Load）的过程。它是构建数据仓库系统的关键环节。因为数据仓库主要存储面向主题的、集成的、稳定的并且随时间不断变化的数据集合，所以数据在进入到仓库之前，需要经过清洗、转化的过程，保证数据仓库的数据是准确的。ETL 的作用就是解决数据集成化的问题。

ETL 过程中包含一些灵活的计算、汇总、字段拆分、字段合并、数据比较、过滤、混合运算等内容，还包括对自定义函数的支持、复杂条件的过滤、数据的批量加载、时间类型的转换、多种数据类型支持、去重复记录等功能。

在数据仓库系统中，ETL 占有重要的地位。ETL 作为一种数据整合解决方案，已经上升到了一种理论的高度。ETL 在数据仓库中具有以下两个特点。

1）数据流动具有周期性。一般来说，商业智能 ETL 按照某种业务抽取规则周期性运行，每次运行都会加载新的数据到目标库中。

2）因为数据仓库中的数据量巨大，所以一般采用成熟的 ETL 工具去完成抽取、转换、加载工作，以降低设计开发和维护的复杂度，使设计开发人员有更多的时间专注于业务转化规则。

ETL 是数据仓库项目中最艰难且耗时最长的工作之一。ETL 系统的设计和开发工作对商业智能项目的成败产生至关重要的影响。如果把数据仓库项目看成一座大厦的话，那么数据模型就像图样，而 ETL 就是建造这座大厦的过程。而作为从事商业智能的专业人士，需要真正理解 ETL 理论方面的知识，而不仅仅停留在 ETL 工具的使用上，因为只有这样，才能更好地发挥它的作用。

4. 联机分析处理技术

联机分析处理（OLAP）技术主要通过多维的方式对数据进行分析、查询和报表处理。这种决策分析是基于多维的和历史数据的。

联机分析处理是数据仓库应用的前端工具，同时可以与数据挖掘工具配合使用，以增强

决策分析的功能。

OLAP 的基本目标就是支持决策分析和多维数据查询。OLAP 通过对信息的各种形式的存取，满足企业决策人员和管理人员对复杂查询的处理，并且将结果提供给决策分析人员，使他们对企业的运营状况有更深入的了解，能够制定出正确的决策方针。形象地说，OLAP 是引领企业发展的“灯塔”。

OLAP 系统的特点包含以下几个方面，如图 7-14 所示。

1）丰富的报表展示功能：OLAP 系统一般有丰富的报表展示功能，如柱形图、折线图、饼形图。

2）数据访问和多维分析的能力：提供给用户数据访问和多维分析的能力，并以用户希望的方式进行展示。

3）快速的数据分析能力：OLAP 系统有秒级的数据分析能力。

图 7-14　OLAP 系统的特点

5. 可视化分析

“一图胜千言”，虽然图形可以传达大量信息，但是图形一定要干净、清晰，同时传达出重要的信息。很多企业领导或者分析人员看到复杂的图形时，可能会非常苦恼。

数据可视化分析是指数据用各种图像处理技术，将数据转化成各种图表的方法和手段。例如，数据可以用饼图、散点图、直方图和柱状图等方式进行展示。它们是数据可视化的基础。但是面对复杂的数据集，比如财务报表、用户行为数据，可以用立体、多维或者动态实时的方式进行展示。数据可视化本身可以看做是一门艺术。

数据可视化分析的特点如图 7-15 所示。

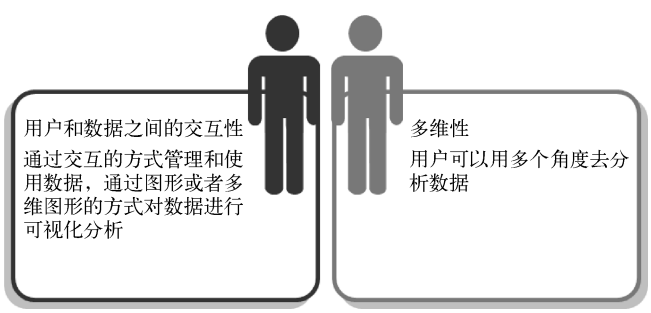


图 7-15　数据可视化分析的特点

数据可视化分析的功能如下：

1）通过可视化技术，辅助进行数据关联分析。

2）通过可视化技术，识别和预测活动，帮助高层人员做出及时和准确的决策。

● 数据可视化的过程

复杂的数据可视化包括数据的采集、数据分析和挖掘等一系列的过程，然后由技术人员以立体、多维或者实时动态的方式将数据展示出来。

● 数据可视化的目的

数据的可视化是为了观察和跟踪各种数据，生成实时的、可读性强的图表；分析数据，生成交互式的图表；发现数据之间的潜在关系，生成多维图表，以及多角度的分析数据，帮助用户深刻地理解数据含义和变化。

数据可视化可以有多种表现形式，如图 7-16 ~ 图 7-18 所示。

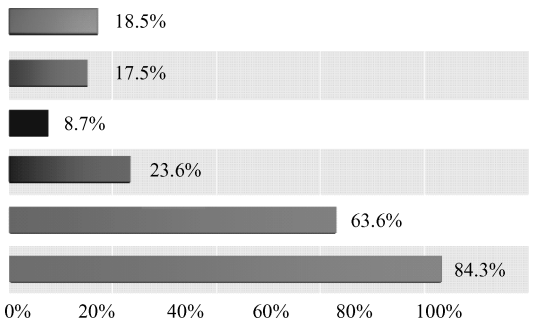


图 7-16　数据可视化的表现形式之一

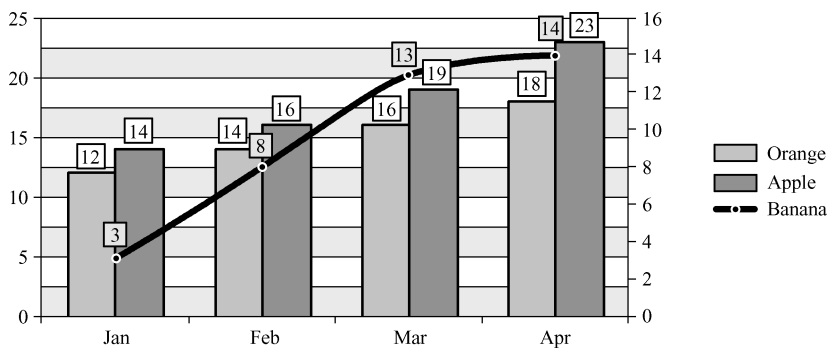


图 7-17　数据可视化的表现形式之二

6. 商业智能元数据管理

在商业智能领域中，元数据定义为：在数据仓库系统的建立、维护、管理和使用过程中，用以描述实际数据的信息，是关于数据的数据。在商业智能系统的建设过程中，元数据占有非常重要的地位，它不仅定义了数据仓库的许多对象，例如表结构、所有的字段列等属性，还包括对数据仓库内部数据流动和业务规则的描述。元数据的框架图如图 7-19 所示。

元数据管理是整个商业智能系统中最重要的环节之一。元数据管理贯穿于商业智能系统数据“流动”的全过程，主要包括数据源元数据、采集元数据、数据仓库元数据、数据集市元数据、应用服务层元数据等。

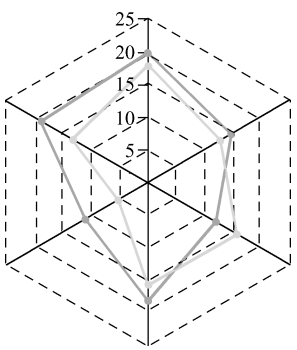


图 7-18　数据可视化的表现形式之三

元数据的分类主要包括业务元数据、技术元数据和管理元数据，如图 7-20 所示。

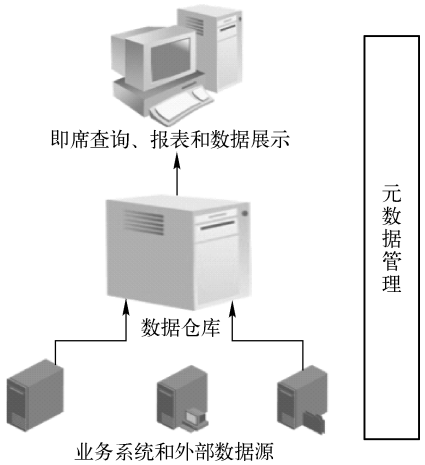


图 7-19　元数据的框架图

图 7-20　元数据的分类

● 业务元数据

业务元数据可以分成业务规则、业务指标、业务描述和业务术语等 4 个部分。这 4 个部分共同完成对业务信息的表述。

● 技术元数据

技术元数据包含关于商业智能系统技术层面的信息，描述了数据源接口、ETL 映射关系、数据仓库和数据集市等系统的特征。

● 管理元数据

管理元数据主要是指商业智能系统日常建设过程中涉及开发、运维管理各方面的基本信息，在此基础上对系统需求开发和日常运维管理提供支持。

元数据在商业智能项目中占有非常重要的地位，是数据仓库系统的灵魂和核心。数据仓库系统在建设的过程中产生的数据源定义、转换规则的定义、目标库的定义都存储在元数据库中。元数据还支持以下几种功能：

1）描述数据仓库系统存在哪些数据。

2）描述哪些数据是在数据仓库系统中产生的。

3）描述哪些数据将要抽取到数据仓库系统中。

4）评估数据质量的好坏。

5）记录数据抽取工作的执行情况。元数据为企业建设数据仓库系统提供了详细的记录，并且保证了数据的一致性和准确性。因此，元数据对于数据仓库系统的开发和管理是非常重要的，具有决定性的意义。

7.2　商业智能—数据仓库理论概述

7.2.1　数据仓库的概念

数据仓库是一个面向主题的、集成的、非易失的、反映历史变化的、随着时间的流逝发

生变化的数据集合，它主要用来支持企业管理人员的决策分析。

数据仓库中面向主题的特性是根据业务的不同而进行的内容划分。数据仓库的集成特性是因为不同的业务源数据具有不同的数据特点，当业务源数据进入到数据仓库时，需要采用统一的编码格式进行数据加载，从而保证数据仓库中数据的唯一性。数据仓库的非易失性是指数据仓库通常保存数据不同历史时期的各种状态，并不对数据进行任何更新操作。数据仓库的历史特性是指数据保留时间戳字段，记录每个数据在不同时间点的各种状态。

7.2.2 数据仓库的特点

数据仓库的主要特点如图 7-21 所示。

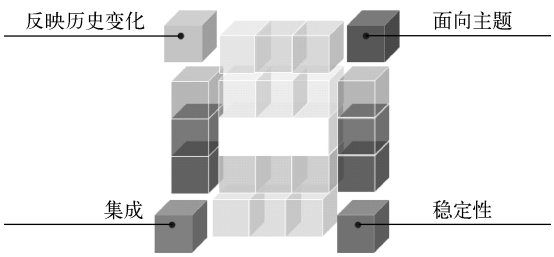


图 7-21 数据仓库的主要特点

1. 面向主题

普通的操作型数据库主要是面向事务性处理，而数据仓库中的所有数据一般按照主题进行划分，主题是对业务数据的一种抽象，是从较高层次上对信息系统中的数据进行归纳和整理。

面向主题的数据组织可以分成两部分：根据原系统业务数据的特点进行主题的抽取和确定每个主题所包含的数据内容是什么。典型的主题包括客户主题、产品主题、财务主题等，其中客户主题包括客户基本信息、客户信用信息、客户资产信息等内容。我们在分析数据仓库主题的时候，一般的方法是先确定几个基本的主题，然后将范围扩大，最后“逐步求精”。

2. 集成

数据集成是数据仓库的主要特点之一。

1）数据仓库是多个数据源的综合和汇总。

2）对于数据仓库来说，数据必须转换成统一的格式。

3）在数据仓库系统的建设过程中，数据集成工作占到系统建设的 80% 以上。

4）数据仓库中的数据经过源系统的抽取、清洗、转换、加载得到，为了保证数据不存在二义性，对源数据进行编码的统一和必要的汇总，以保证仓库内数据的一致性。数据仓库在经历集成阶段后，使得数据仓库中的数据遵循统一的编码规则。

集成一般有两种形式，如图 7-22 所示。

- 数据的集成

当数据从操作型数据库传向数据仓库时，数据就会被集成。

- 编码的集成

当数据仓库是从原有分散的源数据库抽取出来的时候，为了消除编码的不一致性，需要

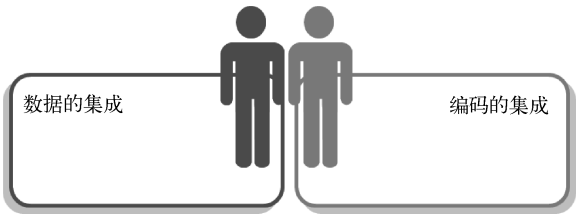


图 7-22　集成一般有两种形式

将这些来自不同数据源的数据编码集成起来，使之遵循统一的编码规则。

3. 稳定性

数据仓库中的数据反映的都是一段历史时期的数据内容，它的主要操作是查询、分析而不进行一般意义上的更新，一旦某个数据进入到数据仓库后，一般情况下，数据会被长期保留，当超过规定的期限时，才会被删除。通常数据仓库需要做的工作就是加载、查询和分析，一般不进行任何修改操作，是为了企业高层人员决策分析之用。

4. 反映历史变化

操作型数据库主要反映某一时间段内的数据，而数据仓库的目标就是对企业的发展趋势做出分析和预测。数据仓库不断地从 OLTP 数据库中获得变化的数据，从而形成分析和预测需要的历史数据，所以一般数据仓库中数据表的键码都含有时间项，以标明数据的历史时期信息，然后不断地增加新的数据内容。

通常来说，数据仓库包含的时间期限大概是 5 ~ 10 年，当超出规定的期限时，需要删除这些过时的数据。通过这些历史信息可以对企业的发展历程和趋势做出分析预测。同时我们要清楚，数据仓库的建设需要大量的业务数据作为积累，而将这些宝贵的历史信息经过加工、整理，最后提供给决策分析人员，这是数据仓库建设的根本目的。

7.2.3　数据仓库和数据库之间的区别

数据库生产系统主要是面向应用的、事务型的数据处理，一般来说，具有实时性较高、数据检索量较小、普通用户的数量较大等特点。而数据仓库系统主要面向主题的、分析型的数据处理，实时性要求不高，数据检索量较大，主要针对特殊的用户群体（一般是企业高层领导、决策分析人员等），用户的数量较小。

其中事务型处理数据和分析型处理数据是有区别的。

一般来说，事务型处理数据对性能的要求较为严格，数据是事务驱动的，主要面向应用，存储的一般都是具备即时性、细节性特点的数据，数据是可更新的。

对于分析型处理数据，一般来说，对性能的要求较高，数据是分析驱动的，主要面向决策分析，存储的一般都是历史、汇总性的数据，数据是不可更新的。

相比其他系统，数据仓库系统有哪些优势呢？有下面几种：

1）数据仓库系统可以获取生产系统综合的信息，作为科学决策分析的重要依据。

2）可以从宏观和微观的角度理解信息。

3）可以通过数据仓库系统建立企业各个部门之间的联系。

7.3 商业智能—数据集市理论概述

7.3.1 数据集市简介

1. 数据集市产生原因

1）数据仓库虽然能够满足所有最终用户的需求，但是各个部门业务不同，需求侧重点不同，且需求也是不断变化的，这就要求数据仓库存储的数据具有充分的灵活性，以适应各类用户的查询和分析。

2）最终用户对信息检索要求是高性能的，即越快越好。

对数据仓库而言，灵活性和性能是一对矛盾体。提高灵活性就要存储各种历史数据，但是一个特定查询就要关联很多表，性能就不能保证。为了解决这一矛盾，数据仓库中就增加了数据集市。数据集市存储为特定用户需求而预先计算好的数据，从而满足用户对性能的要求。

数据集市产生的另外一个原因是数据仓库开发周期较长，投入较大，规模较小的企业无法承担。数据集市能够快速解决某些问题，而且投资规模也比数据仓库小很多。

2. 数据集市的定义

比尔·盖茨说过：“如何收集、管理和利用信息将决定您的胜负。”商业智能正是在这种需求下诞生的，而数据集市是满足部分特殊用户群体用来收集、管理他们本部门、本专业信息的数据仓库。

大多数情况下，数据集市的数据来源于数据仓库，它是一种小型的部门级别的数据仓库。数据集市的重点就是它满足了某些用户的特殊业务需求，根据所属部门的需求，对历史数据进行必要的汇总和计算。那么什么是数据集市呢？

数据集市就是满足特定的部门或者用户的需求，按照多维的方式进行存储，包括定义维度、需要计算的指标、维度的层次等，生成面向决策分析需求的数据立方体。数据仓库体系结构中增加了数据集市，数据集市又可以看做部门级的小型数据仓库，如图 7-23 所示。

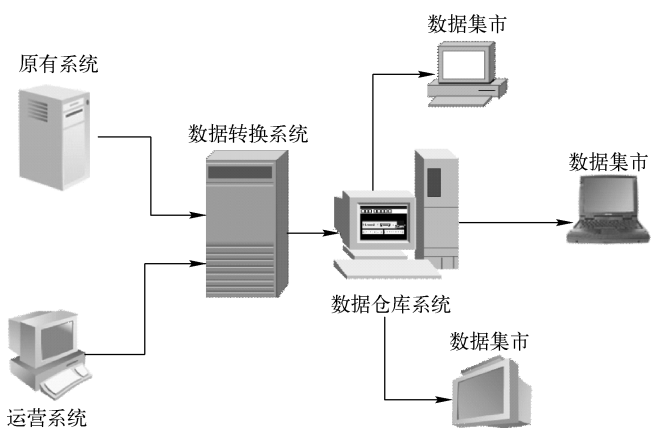


图 7-23　数据集市与各系统之间的关系

3. 数据集市的优点和缺点

数据集市的优点：

投资规模小，投资回收期相对较短、灵活，风险性较小，同时可以按照多种方式进行织，如部门、应用等。

数据集市的缺点：

1）建立数据集市的部门是相互隔离的，很多标准、流程和知识经验不能共享，这会导致大量的资源浪费和重复劳动。

2）数据集市在某种程度上会造成成本的增加，例如很多部门会选择不同的工具、软件和硬件，同时需要一定数量的技术人员。

3）不同的部门建设各自的数据集市，这些集市之间没有数据的集成，相互独立，因此可能会出现数据不一致的现象。

4. 数据集市分类

数据集市的分类包括：产品类数据集市、管理类数据集市和研发类数据集市。

（1）产品类数据集市

产品类数据集市的定位是通过数据挖掘、建模和其他方法，帮助企业发现重要的趋势和规律，以提高运营效率。产品类数据集市的对象主要是企业内部人员。

产品类数据集市主要包括：文本分析、模拟分析、数据挖掘、预测分析、优化分析和可视化分析，如图 7-24 所示。

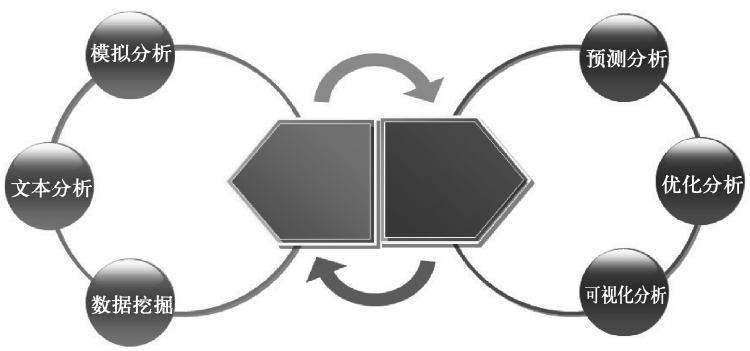


图 7-24　产品类数据集市

- 文本分析

对非结构化数据中的文本进行分析，以提高业务洞察力。

- 模拟分析

用先进的技术手段模拟流程、行为和业务，可以帮助企业分析未来业务的发展方向。

- 数据挖掘

数据挖掘是由专业人士根据不同的业务场景选择不同的挖掘算法，通过数据挖掘探索数据背后隐藏的规则，从而进行业务预测和归类。

- 预测分析

通过历史和当前交易数据去分析和预测未来的业务能力。

- 优化分析

利用先进的数学技术，帮助企业提高运营效率，同时提供强大的知识库。

- 可视化分析

通过图表、地图、日程表和图片等，利用专业的工具分析业务的趋势等。

（2）管理类集市

管理类集市是指为了运营管理的需要而进行的数据整合分析，从而更好地提高企业的运营水平。管理类集市主要面向企业的内部人员，一般来说，对于数据的实时性要求不高。

管理类集市应用包括管理驾驶舱、固定报表、OLAP 分析、关键绩效指标（KPI）和数据质量检查等，如图 7-25 所示。

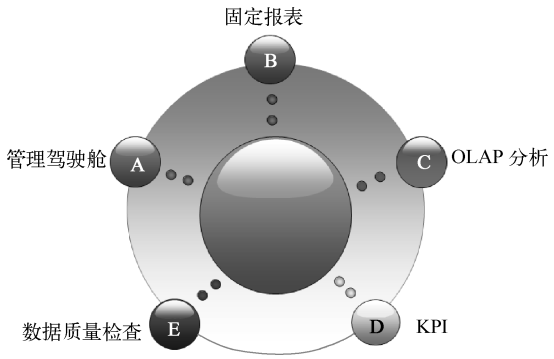


图 7-25　管理类集市应用

- 管理驾驶舱

管理驾驶舱对企业领导层所关注的经营活动的关键指标做定制化展示，并以直观的图表形式展示结果。

- 固定报表

以固化报表的形式将手工报表自动化。

- OLAP 分析

通过灵活的多维分析，帮助企业决策者发现问题，追溯问题根源，预测发展趋势等。同时为制定问题解决方案，改善企业经营状况提供帮助。

- KPI

KPI（Key Performance Indicator）即关键绩效指标。

- 数据质量检查

按照业务需求定义数据质量检查规则，按照规则定期得出数据质量分析报告，提供给业务部门敦促报送机构提供数据质量。

（3）研发类数据集市

研发类数据集市主要是支撑企业各部门的业务应用系统，提供业务需要的数据集合，主要用于支持数据研究分析工作。研发类数据集市同样也支持各部门的临时业务需求。研发类数据集市之间是相互独立的。

在架构中，数据集市是基于数据仓库进行产品加工的，数据集市的建设方式可以分成两种模式：库内数据集市和库外数据集市。

所谓库内数据集市是部署在企业数据仓库之内的，在数据仓库汇总数据的基础上构建特定应用的数据集市。库内集市可以共享仓库内的汇总数据。

库外数据集市是在数据仓库之外单独部署，具有专门的软硬件设备，数据来源可以是数据仓库的基础层数据，或者是汇总层的数据。

7.3.2 数据集市和数据仓库的联系和区别

（1）数据集市和数据仓库的联系

数据集市是一组特定的、针对某个主题域、某个部门或者某些特殊用户而进行分类的数据集合，也可以说是小型的数据仓库。用户可以在数据集市中快速地对数据进行访问和对报表进行展示，同时在数据结构的内部对数据进行必要的汇总和优化。

数据集市的存储通常按照划分主题的形式进行存放，其模型一般是星形结构或者雪花形结构。而数据仓库除了按照主题的形式进行存放外，其模型一般按照第三范式的形式进行设计。数据仓库到数据集市的过程是从数据规范化到多维建模的过程，包括数据仓库内的实体表转化成事实表、维表，以及将实体之间的关系转化成多维关系的映射。

在数据仓库项目中，数据集市通常按照地区、日期等维度对数据进行组织和汇总，因此数据仓库转化成数据集市也是按照轻量级汇总或者中度汇总和计算所完成的。简而言之，数据集市里的数据一般都是从数据仓库中经过转换、汇总计算获取的，直接支撑前端的应用需求，如图 7-26 所示。

数据集市的数据通常会作为 OLAP 服务和应用服务的数据输入。数据集市的数据一般不会从源数据系统中直接抽取，即一般不提倡建设独立型的数据集市。这是因为，如果数据集市从源数据系统中直接抽取数据，则可能导致数据的不一致性，同时也会增加多个额外的进程，这些进程在源系统中将大大消耗系统的 CPU 资源，从而造成资源上的浪费。数据集市和数据仓库的关系如图 7-27 所示。

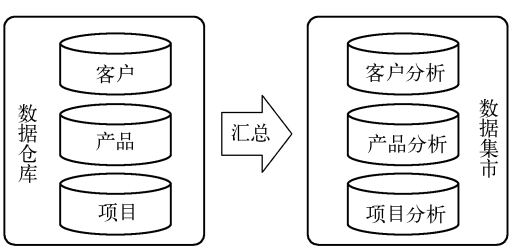


图 7-26　数据集市的数据来源

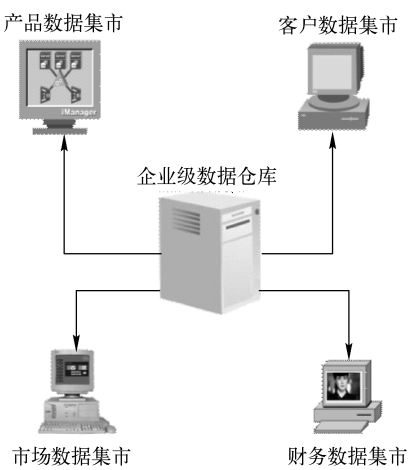


图 7-27　数据仓库和数据集市的关系

（2）数据集市和数据仓库的区别

数据仓库的数据是经过整合和清洗的，它能够提供统一的视图。当数据仓库建成之后，报表、OLAP 应用和数据分析挖掘都可以从数据仓库中获取数据。

对于数据集市来说，它主要是通过分析应用的特点，判断应该获取什么样的数据。例如，市场部的数据集市可能不需要人力资源的数据。一般来说，数据集市就是企业级数据仓库的一个子集，主要面向部门级的业务，或者某个特定的主题。

在数据结构上，数据仓库是面向主题的、集成的数据的集合。而数据集市通常定义为星形结构或者雪花形结构。数据集市一般是由一张事实表和几张维表组成。数据仓库和数据集市的数据结构如图 7-28 所示。

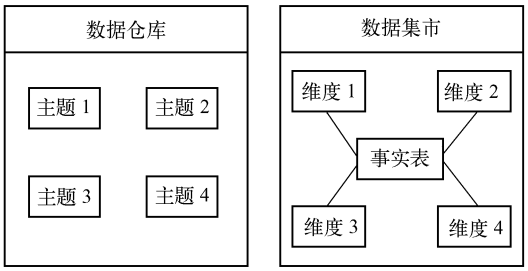


图 7-28　数据仓库和数据集市的数据结构

数据仓库和数据集市的区别见表 7-1。

表 7-1　数据仓库和数据集市的区别

比 较 对 象	数 据 仓 库	数 据 集 市
数据来源	ODS	数据仓库
数据范围	面向企业级	一般是部门级
数据结构	第三范式	雪花形或星形结构
历史数据	大量的历史数据	一部分历史数据
索引	高度索引	高度索引

7.3.3　数据集市的技术特性

数据集市是数据仓库体系中的一种小型的部门或工作组级别的数据仓库，从而满足用户对性能的需求。数据集市在一定程度上可以缓解访问数据仓库的瓶颈问题。根据数据集市应用的不同，可以分成库内集市或库外集市。数据集市技术路线的指导原则包括：

1）大规模并行处理能力。

2）数据高速加载和卸载。

3）存储压缩。

4）快速刷新。

5）海量数据处理能力。数据集市和数据仓库的区别在于数据的范围和主题，数据仓库是全局的整体的数据，数据集市主要服务于特定主题，在某些时候，数据集市的数据量很大，因此，集市需要具备处理大并发、复杂查询的能力。

6）线性扩展能力。数据集市平台应该具备线性扩展的能力，可以满足数据不断增长的需求。

7）工作负载管理能力。提供工作负载管理能力。

8）高可用性。数据集市平台可以提供高可用的方案，满足系统的高可用性要求。

9）数据压缩。必须提供良好的数据压缩能力，降低存储成本，多段备份和恢复时间，满足系统的时间要求。

10）高速数据加载和卸载能力。必须提供高速的数据加载和卸载能力，以保证数据加载和卸载能够在较短的时间内完成，从而减轻运维压力。

11）星形模型/雪花形模型性能优化。一般来说，数据集市中的数据，通常按照星形和雪花形模型组织，数据集市平台必须提供针对性的优化，以满足用户响应时间的要求。

12）满足数据库平台需求的能力。数据集市平台本身是一个数据库平台，除了满足数据集市的特殊需求外，还必须满足数据库平台全部能力。

7.4 商业智能—ODS 概述

7.4.1 ODS 简介

1. ODS 的概念

通过前几章的学习，我们已经知道数据仓库是一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合，用于支持用户的辅助决策分析。

而 ODS（Operational Data Store，操作数据存储）则是一个面向主题的、集成的、可变的、反映当前细节的数据集合。它主要用于支持企业处理业务应用和存储面向主题的、即时性的集成数据，为企业决策者提供当前细节性的数据，通常作为数据仓库的过渡阶段。

2. ODS 建设原因

ODS 系统建设的原因有多种，主要原因有：

1）系统重复开发，造成资源极度浪费。不同应用之间，可能存在相同的数据抽取需求，经过多次抽取，浪费网络存储资源，造成不同应用系统之间数据的不一致性，同时也会给业务系统带来沉重的压力。

2）一般来说，业务部门需要的信息可能来自于多个系统，但是由于各个系统之间的数据可能会出现口径不一致，数据不规范的现象，因此大大增加了临时抽取数据的难度，同时很难保证数据的一致性和准确性。

综上所述，通过 ODS 系统的建设，既可以大大缩短应用系统的实施路径，减少重复性的设计和开发，又可以提高数据的响应速度和准确性，为以后的数据挖掘和分析打下基础。

3. ODS 的特点

业务数据经过 ETL 数据抽取、转换、加载，进入到 ODS 系统中，为企业提供了一种全局的、集成的和反映当前实时性的视角，在支持企业决策分析需求的同时，还能够在业务系统和数据仓库之间构建一个数据缓冲带，使得数据之间的传输和转换变得相对容易。

ODS 系统的主要功能就是将多个业务系统中不同的数据源进行数据集成，通过数据抽取、转换、加载，将数据放入到共享的存储区中，以保证数据的一致性。

ODS 具有以下特点：

1）数据是不断更新和易丢失的，当新的业务数据进入到 ODS 时，旧的数据会被新数据覆盖或者更新，一般不存储历史数据，只反映当前实时性的信息。

2）ODS 系统一般存储的都是细节性的信息，很少有汇总的数据，即 ODS 包含粒度级别最低的数据。

3）ODS 系统支持快速的数据更新操作，数据刷新频率很快，一般不保存过期的历史数据。

4）ODS 系统一般存储在关系数据库中，通过将各个业务系统的数据集成起来，组成企业的全局统一性视图，实现 ODS 的数据共享功能。

5）用户可以频繁访问 ODS 系统，因为它是基于操作型应用的。

4. ODS 设计原则

ODS 的设计原则包括可扩展性、高可用性、可重用性和高性能，如图 7-29 所示。

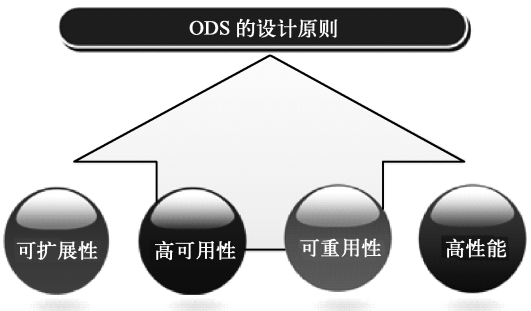


图 7-29　ODS 的设计原则

（1）可扩展性

可扩展性是指 ODS 系统可以支持业务系统扩展的需要。举例来说，设计 ODS 数据模型的时候，应该充分考虑新旧系统的业务数据模型能够扩展到 ODS 系统中。

（2）高可用性

高可用性是指系统发生变化的时候，可以依赖架构的灵活性，仍能保证系统的正常运行。例如，对于模型的设计，应该考虑业务源系统结构发生变化时对 ODS 系统带来的影响。也就是说，局部模型的扩展不会影响到 ODS 数据模型。

（3）可重用性

可重用性是指尽量避免重复的系统建设，尽可能考虑物理设备、系统软件、模型以及应用上的复用。举例来说，对于 ETL 处理流程，分析 ETL 任务的各个环节，找出公共的组件，进行封装，然后进行复用。

（4）高性能

高性能是指 ODS 系统可以承受峰值时的系统压力和更多的应用，保证系统可以正常运行。

5. ODS 的主要功能

ODS 的主要功能如图 7-30 所示。

（1）作为业务系统和数据仓库之间的隔离地带

一般来说，数据仓库系统的数据来源非常复杂，数据可能存储在不同的应用系统和业务数据库中，为了满足数据仓库对业务数据的抽取标准，需要在应用系统和数据仓库系统之间建立一个“隔离墙”，如图 7-31 所示。ODS 系统作为“隔离墙”的目的是临时存储多个业务源数据，经过一系列的清洗、转换并达到数据仓库对数据的要求后，再将数据加载到数据仓库中。

在业务系统中直接将数据抽取到数据仓库中并不容易，ODS 系统作为业务系统和数据仓库系统之间的隔离地带，用于存放从业务系统抽取出来的数据，为数据仓库提供了平整、可靠的数据源。

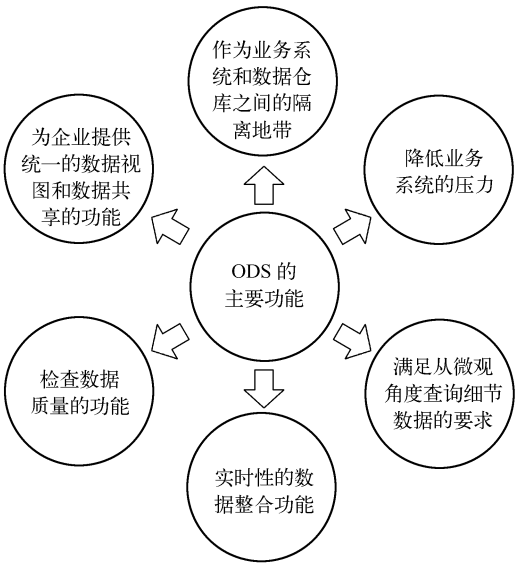


图 7-30　ODS 的主要功能

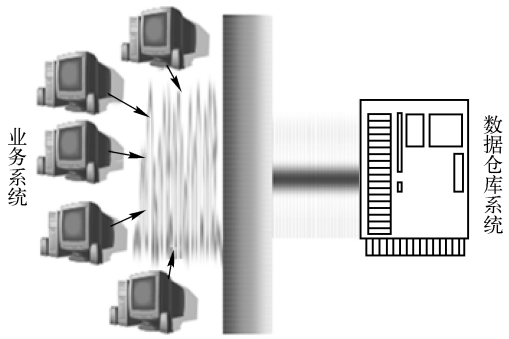


图 7-31　ODS 的“隔离墙”功能

（2）降低业务系统的压力

在数据仓库建立之前，大量的决策分析报表是由业务系统直接生成的，在报表生成过程中可能存在复杂的计算，对业务系统的运行产生非常大的压力。在建立 ODS 系统之后，原来由业务系统直接产生的复杂报表、对细节数据的查询都能够在 ODS 系统中进行，从而有效降低了业务系统的查询压力，提高了业务系统的运行效率。

（3）满足从微观角度查询细节数据的要求

一般来说，在数据仓库体系结构中，数据仓库层存储的数据都是经过轻度汇总的数据和历史数据，几乎不存储任何生产运营过程中产生的细节数据。但是，为了满足特殊用户群体的要求，可能需要对一些交易数据进行查询，这时需要把查询这些交易数据的功能让 ODS 系统来实现。通常，ODS 系统支持多维分析的功能，因为它也是面向主题的和集成的。数据仓库从宏观上支持多维决策分析，而 ODS 系统从微观角度描述细节性的数据查询。

（4）实时性的数据整合功能

ODS 系统具有实时性的数据整合功能。它通过 ETL 技术，实时地从各个业务系统中抽取企业的运营交易数据，通过数据转换、清洗、加载等操作最终形成共享数据，为企业提供统一的数据视图。这种数据整合功能有助于提高数据的一致性，为数据仓库提供优质的数据源。

（5）检查数据质量的功能

ODS 系统具有完善的数据质量检查功能。它通过对企业数据的质量检查和质量评估，完善企业内部的组织机构，支持对数据质量管理流程的监控，从而实现对源数据质量问题的发现和修正。

（6）为企业提供统一的数据视图和数据共享功能

ODS 系统为企业提供统一的数据视图和数据共享功能。它通过对各个业务系统运维数

据的集成，实现 ODS 的数据共享，同时为企业提供全局的统一数据视图。

6. ODS 的设计步骤

ODS 的设计步骤如图 7-32 所示。

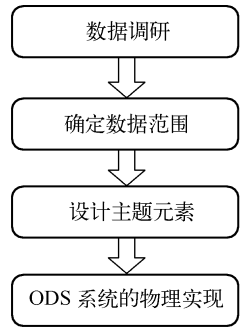


图 7-32　ODS 系统的设计步骤

（1）数据调研

数据调研主要是根据业务人员提供的需求意向，将业务系统划分成几个模块，并对各个模块所涉及的数据和数据源进行调研分析。数据调研分析可以分成编号、模块名称、数据来源（包括导入和输入）、备注等信息，见表 7-2。

表 7-2　数据调研分析表

编　　号	模块名称	数据来源		备　　注
		导　　入	输　　入	
1	生产数据	* * * * * *	* * * * * *	
		* * * * * *	* * * * * *	
2	基本数据	* * * * * *	* * * * * *	
		* * * * * *	* * * * * *	

（2）确定数据范围

确定数据范围是在需求调研的基础上进行的。为了保证所需的数据都能够从业务系统中得到，需要将应用需求与 ODS 的数据范围结合起来，也就是对 ODS 进行主题的划分。通常来说，ODS 主题的划分是以企业的业务模型为基础，通过参考各种业务系统信息模型，得到 ODS 数据主题的范围，根据该范围进行 ODS 主题的定义，从而确定 ODS 的数据范围。

（3）设计主题元素

ODS 系统的主题元素主要包括主题名称、维度、度量值、粒度、存储的时间，下面分别进行介绍。

- 主题名称：说明该主题主要包含哪些分析数据，用户重点关注的对象是什么。
- 维度：说明数据分析时的角度有哪些，如时间维度的年、季、月、日等。
- 度量值：说明用户关注的指标值，如工资额、销售量等。
- 粒度：是指对数据的细化程度。一般来说，细化程度越高，粒度级别就越低；细化程度越低，粒度级别就越高。
- 存储的时间：主要描述数据的存储周期和存储期限是多少。

（4）ODS 系统的物理实现

ODS 系统的物理实现主要包括数据库的物理实现、数据抽取的设计等内容。

7.4.2　ODS 系统与数据库系统、数据仓库系统的区别

ODS 系统是既不同于一般的数据库系统，又不同于数据仓库系统的一种特殊的数据存储系统。它与一般数据库有很多区别，它的数据组织方式是面向主题的、集成的，而数据库系统则是面向应用和事务处理的。

ODS 系统与数据仓库系统相比，它只存储当前的、细节性的信息或者接近当前的实时性数据，可以对数据进行增加、删除和修改等操作，而数据仓库系统虽然是面向主题和集成的，但是数据一般不进行修改，并且存储大量的历史数据。ODS 系统和数据仓库系统的主要区别体现在数据的时间性、稳定性、可修改性、细节性和用户访问频率上。

ODS 系统与数据库系统、数据仓库系统的区别如图 7-33 所示。

数据库系统	ODS 系统	数据仓库系统
• 面向应用、事务处理 • 实时性高 • 数据检索量小 • 只存储当前数据 • 访问频率高 • 响应时间控制到 1s 以下 • 用户数量大	• 面向主题、集成的 • 实时性要求高 • 数据检索量小 • 一般只保留当前数据 • 访问频率高 • 响应时间控制到 1s 以下 • 用户数量相对较小	• 面向主题、集成的 • 实时性要求不高 • 数据检索量大 • 存储大量历史数据和轻度汇总的数据 • 访问频率中、低 • 响应时间需几秒或者更长 • 用户数量相对较小

图 7-33　ODS 系统与数据库系统、数据仓库系统的区别

总结：

（1）ODS 系统与数据仓库系统的区别

1）ODS 系统是业务数据进入到数据仓库系统中的一段临时存储区域，存储当前或者接近当前的实时性数据，而数据仓库一般只存储历史数据。

2）ODS 系统对数据的更新是频繁的，而数据仓库中的数据是不能更新的，数据的任何变化都应该反映到数据仓库中。

3）ODS 系统主要存储细节性的数据，而数据仓库系统既包含细节性的历史数据，同时也包含轻度汇总的数据。

（2）ODS 系统与数据库系统的区别

1）数据库系统主要是面向事物处理和应用的，而 ODS 系统主要是面向主题的和集成的。

2）数据库系统的用户量相对较大，而 ODS 系统面对的用户数量相对较小。

7.4.3　基于 ODS 的即时 OLAP 应用

基于 ODS 系统的即时 OLAP 应用是建立决策分析的一种解决方案，通常应用于中、低级别的决策分析应用。基于数据仓库的 OLAP 应用是为了进行长期的趋势分析，但是一般运行较慢。如果企业决策者需要查看周期时间较短的一些指标情况，不需要太多的历史数据，这样就需要建立基于 ODS 的即时 OLAP 应用。基于 ODS 的 OLAP 和基于数据仓库的 OLAP 之间的关系如图 7-34 所示。

例如，查看一周之内的各地区销售情况，只需要参考当前时间内一周的历史数据，如果在数据仓库中建立即时 OLAP 应用，运行效率非常低，并且很难准确地反映当前时间的各地区销售情况。

基于 ODS 的即时 OLAP 应用	基于数据仓库的即时 OLAP 应用
• 是决策分析系统的一种解决方案 • 满足日常频繁的趋势分析 • 运行时间较短	• 是决策分析系统的一种解决方案 • 满足长期趋势的分析 • 运行时间较长

图 7-34　基于 ODS 的 OLAP 和基于数据仓库的 OLAP 之间的关系

7.4.4 ODS 系统的功能

一般来说，在数据仓库系统中，存储的数据都是轻度汇总的指标数据或者历史数据，很少有细节性的、当前的生产运营数据，但是在特殊的应用中，用户可能会对这些生产数据进行查询，然而数据仓库不支持这些特殊的查询，这部分功能可以由 ODS 系统来实现。

ODS 系统不仅可以支持多维分析等查询功能，还可以满足对细节性的交易数据或者粒度级别很低的数据进行查询的要求。ODS 系统是按照面向主题的方式进行数据存储，同时它又只存储当前时间段内的或者接近当前的细节性数据。ODS 系统的数据组织方式是基于主题的，它对所有业务系统的数据进行集成，组成全局共享的数据视图。ODS 系统的另一个重要功能就是数据共享的功能，它的数据存储量取决于对业务数据的抽取频率。

ODS 系统的数据具有交互功能，不仅提供企业的全局信息统一视图，满足对信息共享的需求，同时还可以在固定的周期内，实现决策分析系统与其他业务系统之间的交互。当 ODS 系统的数据有更新时，外围的业务系统数据也会发生相应的变化。ODS 系统的功能如图 7-35 所示。

图 7-35　ODS 系统的功能

7.4.5 ODS 系统的架构

ODS 系统是一个面向主题的、集成的、当前的、可更新的数据集合，用于细节性的查询和为决策分析系统提供当前时间段内的数据。ODS 系统是介于操作型数据库和数据仓库之间的一种存储方式，其中数据仓库存储的是概括性的数据和历史数据，而 ODS 系统存储的是细节性数据和当前时间段内的数据。

数据仓库系统和 ODS 系统之间的结合能够分析企业当前的运营情况，同时对未来企业的经营状况进行合理的规划和分析。ODS 系统中的数据可以进行增加、删除、修改等操作，但是数据仓库中的数据一般不能进行修改。数据仓库系统与业务系统相隔离，目的是减小数据仓库的处理和决策支持分析对业务系统造成的影响，减少业务系统的压力。

ODS 系统的架构如图 7-36 所示。

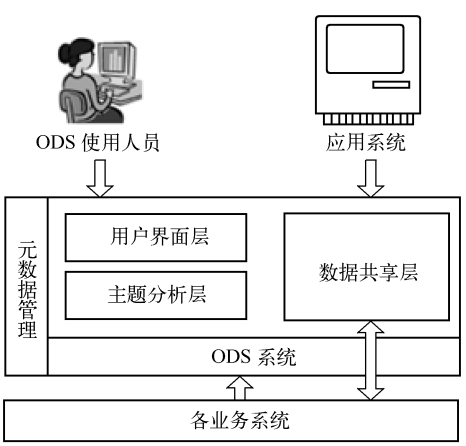


图 7-36　ODS 系统的架构

7.5 商业智能—ETL 概述

7.5.1 ETL 体系是商业智能核心的技术架构

在商业智能系统中，ETL（Extract – Transform – Load，数据抽取、转换、加载）占有重要的地位，ETL 作为一种数据整合解决方案，已经上升到了一种理论的高度。ETL 在商业智能系统中具有以下几个特点。

1）数据流动具有周期性。一般来说，商业智能 ETL 按照某种业务抽取规则周期性运行，每次运行都会加载新的数据到目标库中。

2）因为数据仓库中的数据量巨大，所以一般采用成熟的 ETL 工具去完成抽取、转换、加载，以降低设计开发和维护的复杂度，使设计开发人员有更多的时间去专注于业务转化规则。ETL 是数据仓库项目中最艰难且耗时最长的工作之一。ETL 系统的设计和开发工作对商业智能项目的成败产生至关重要的影响。如果把数据仓库项目看成一座大厦的话，那么数据模型就像图样，而 ETL 就是建造这座大厦的过程。而作为从事商业智能的专业人士，需要真正理解 ETL 理论方面的知识，而不仅仅停留在 ETL 工具的使用上，因为只有这样，才能更好地发挥它的作用。

例如，如图 7-37 所示，建筑图样的规划就是数据仓库模型的设计过程，根据图样建造房屋的过程就是 ETL 设计开发的过程，而那座美丽的房屋就是数据仓库的成果。可以看出，建造房屋的过程就是耗时较长和相对困难的工作，即 ETL 是整个数据仓库项目中难度最大、耗费时间最长的工作之一。

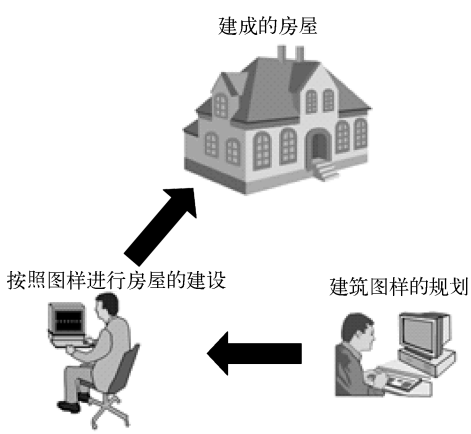


图 7-37　数据仓库模型的设计过程

7.5.2 ETL 的一般过程

ETL 是数据抽取（Extract）、转换（Transform）、加载（Load ）的英文简写。它的一般过程是指：首先访问源数据，连接数据源和目标仓库之间的数据流，然后经过数据的转换、传输和加载，最后加载到目标表中。整体流程中有相应的出错 处理，如图 7-38 所示。

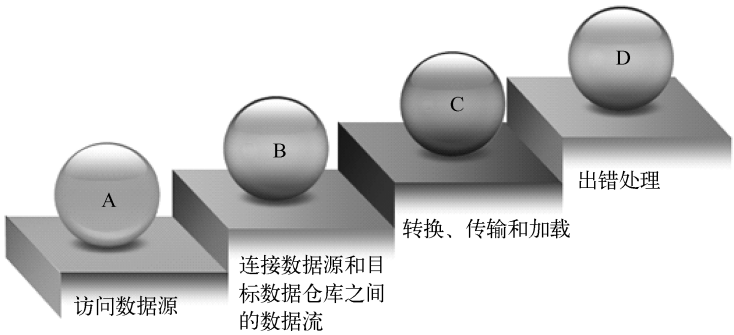


图 7-38　ETL 的一般过程

7.5.3　研究 ETL 的本质

为了更好地理解 ETL 的流程，有必要研究一下 ETL 的本质。

1）用户应该理解 ETL 本质上就是数据从源到目标的过程（即数据流动的过程）。

在数据仓库中历史数据是海量的，ETL 过程需要经历以下几个步骤：抽取、清洗、转换、加载。抽取和加载是转换过程的输入和输出部分，而数据转换是 ETL 过程的核心部分，也是难度最大的部分。可以把 ETL 分成静态单元和动态单元两个部分。所谓静态单元，就是业务数据转换的规则，而动态单元是 ETL 时间调度的最小单位。目前有很多成熟的工具都提供 ETL 功能，包括 Informatica、DataStage、Kettle 等。这些工具不但具有可视化的数据流动、转换编辑界面，还提供各种转换规则定义和数据转化的函数集。

2）多数 ETL 工具价格昂贵，虽然在宏观上一般都适合处理海量的数据，但是在微观上需要考虑 ETL 处理的不同情况。

在数据量和复杂度都不高的情况下，可以利用 ETL 工具提供的组件指定数据源和目标库，通过对图形的拖曳就可以设定需要转换的规则，操作非常方便。在处理大数据量和复杂数据转换时，一般采用编码的方式进行设计和开发，更直观地实现业务转换的规则。ETL 工具（如 Informatica、DataStage、Kettle）都是用图形界面去设置转换规则和编写代码程序，这需要 ETL 设计开发人员熟悉工具中的各种组件和规则转换函数。当然，因为这些 ETL 工具不可能提供所有的转换规则，所以一般 ETL 工具都提供特定语言环境（JavaScript 语言脚本和存储过程的调用功能）来实现高级转换功能。

3）元数据是 ETL 过程的重要体现，描述了数据源的属性、数据源到目标库的转换规则、数据抽取的历史记录等内容。

ETL 的所有过程一般都是依赖元数据去实现数据的清洗、转换，最后加载到目标数据仓库中，同时元数据也是数据仓库项目中不可或缺的部分。采用元数据方法，可以实现数据抽取流程的自动化，并且保证了数据抽取的及时、准确和完整。元数据的概念在数据仓库中非常重要，ETL 中存在大量的数据源定义和映射规则、转换规则，这些都是元数据需要管理和存储的。

4）如果构建一个商业智能系统，设计开发人员要完全理解业务数据源系统是非常困难的，需要花费大量的时间去整理数据源的属性，更多的人喜欢在 ETL 开始之前就将所有的业务转换规则弄清楚。

在 ETL 过程中，如果遇到质量有问题的源数据，一定要正面对待这些垃圾数据或者错误数据，是丢弃还是处理，这些问题都是无法回避的。如果这些数据不经过处理，那么在 ETL 过程中错误会逐渐放大。抛开数据源质量问题，我们再来看看 ETL 过程中哪些因素会对数据的准确性产生重大影响。

影响 ETL 数据质量的关键因素：

- 可能会有一部分数据因为客观或者人为的原因导致数据格式混乱。
- 源系统设计存在不合理性。
- 在开发过程中，因为开发人员的错误或者设计人员对业务规则描述的问题，同样会导致数据质量出现问题。

因为各种因素都有可能影响 ETL 数据的质量，所以保证数据质量的通常做法如下所示。

首先，用户必须遵守在数据仓库项目中数据源的质量要求，对业务源数据进行仔细分析，以便对数据源的任何错误或不规范的地方有相应的处理方法，如对错误数据舍弃或者修改。

然后，在保证数据源的质量之后，在设计 ETL 的过程中，对每一个步骤都应该有一个衡量数据质量的方法，需要重视 ETL 的每一个过程。对于有误差的数据，需要追溯到根本原因，并且将数据仓库的模型与数据质量的验证方法统一起来，实现每一步的 ETL 过程都有验证数据质量的脚本。

最后，就是规范业务流程，保证 ETL 的正确性，避免误删数据或者重复加载业务数据。

其中对质量的衡量有下面几种方式，如图 7-39 所示。

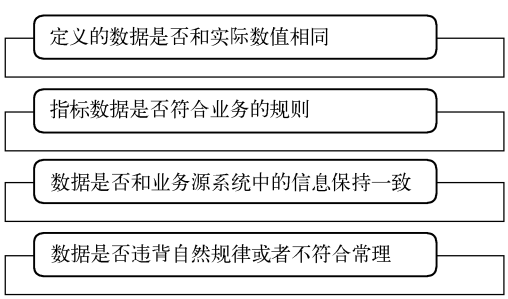


图 7-39　对质量的衡量方式

① 定义的数据是否和实际数值相同。

例如，一个数据项“客户来电等候率”是指在客户服务中，客人来电等候的时间超过 1 min 的次数与客人来电的总次数之比。随着客户服务质量的提高，客人来电等候率会发生变化，当客人来电等候的时间超过 1 min 的次数减少时，客户来电等候率也会相应降低，但是如果这个值没有被更新，那么该数值可能是不正确的。

② 指标数据是否符合业务的规则。

例如，“社会保险类别”是描述社会保险分类的信息，不包括劳动保障类别里的内容。这个指标如果出现劳动保障类别里的信息，就表明该指标违背了业务规则。

③ 数据是否和业务源系统中的信息保持一致。

该数据和源系统中公认的、权威性的信息必须保持一致，否则该数值可能是不正确的。例如，发票中的公司名称必须和公司注册的名称保持一致，公司涉及的所有票据名称必须和

公司合同里的名称保持一致，否则该数值可能不正确。

④ 数据是否违背自然规律或者不符合常理。

如果数据与业务源系统中公认的、权威性的信息保持一致，但是却违背了自然规律或者不符合常理，同样应该分析该数据是否正确。

7.5.4 主流的 ETL 工具

选择合适的 ETL 工具是实际数据仓库项目中必须要考虑的问题，选择的因素包括使用成本、技术人员对此工具的熟练程度、ETL 工具开发商业智能项目的成功案例和工具厂商强有力的技术支持。在实际项目中，常用的工具是 PowerCenter 和 DataStage，一些公司也会用开源的 ETL 工具，如 Kettle。

从本质上来说，ETL 工具的功能都是相同的，都提供了一个全面的数据集成解决方案。ETL 工具的功能如图 7-40 所示。

图 7-40 ETL 工具的功能

ETL 工具可以使用通用的接口（JDBC、ODBC）或者自己厂商的专用接口去抽取源数据，实现了 ETL 对不同数据源的支持。

数据转换是 ETL 工具提供的最强大的功能之一，也是 ETL 开发人员面临的难度最大的问题之一。一般来说，ETL 工具提供了各种组件来实现不同的转换功能，有行列转换、过滤、排序、汇总、分组、计算等常用的转换方式；同时可以实现代理主键的生成，Mapping 的调试功能，抽取远程源数据，各种数据增量加载方式；在转换过程中还可以支持数据比较、类型转换、字段拆分等功能，数据预览，数据的批量装载，性能监控，自动调度 ETL 程序，程序出错处理，按行、按列的聚合汇总等功能。

随着 ETL 工具的发展，ETL 的管理和调度功能得到了加强。管理功能包括 ETL 程序的备份与恢复，版本升级和管理。调度功能包括命令触发方式、事件触发方式和时间触发方式。目前很多公司都在拓展 ETL 的集成性；在原有的基础上嵌入了公共的 API，增加了 JavaScript 语言脚本和存储过程的调用功能，增强了 ETL 工具的灵活性。

7.5.5 ETL 的作用

商业智能数据仓库系统由数据仓库、数据集市、多维数据分析组成。ETL 的作用就是解决数据集成化的问题。ETL 过程中包含字段映射的自动匹配，字段的拆分和混合运算，去重复记录和记录间合并或计算，数据的批量加载，自定义函数，记录的行、列转换，复杂条件的过滤，数据预览和性能监控等内容，如图 7-41 所示。

商业智能系统的目的就是通过数据分析为企业管理者和决策者提供辅助决策支持。因为数据来源不统一，格式混乱、各种类型的“脏”数据都增加了对数据集成整合的难度，所以需要 ETL 提供一个完整的方案来解决数据一致性和集成性的问题。

ETL 的设计和实施是商业智能项目中工作量最大的部分之一，也是最重要的工作内容之一，可以说 ETL 是商业智能的核心和灵魂，如图 7-42 所示。

图 7-41　ETL 工具可以实现的转换要求

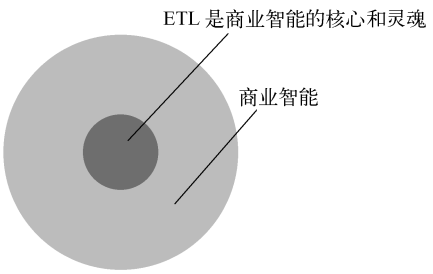


图 7-42　ETL 是商业智能的核心和灵魂

7.5.6　详解 ETL 过程

我们已经知道 ETL 是将业务系统的数据经过抽取、清洗、转换之后加载到数据仓库的过程，通常情况下，商业智能项目的 ETL 部分会占整个项目的 1/3 以上，ETL 的设计会直接决定商业智能项目的成败。下面详细介绍 ETL 中的抽取、清洗、转换、加载等各个部分的内容。

1. 数据抽取

数据抽取就是从源系统中获取业务数据的过程。数据的抽取需要充分满足商业智能系统的决策分析需要，为了保证不影响系统的性能，数据抽取时需要考虑很多因素，包括抽取方式、抽取时间和抽取周期等内容。

例如，抽取方式包括增量抽取、全量抽取。抽取时间应该尽量在系统使用的低谷时段，如夜间。抽取的周期是根据业务的需求制定的，如按小时抽取，或者按天、月、季度、年等抽取。在数据抽取之前，需要确定业务系统的数据情况，了解数据量的大小，以及业务系统中每张表的数据结构、字段含义、表之间的关系等信息，当收集完这些信息后，才能进行数据抽取的设计开发等工作。数据抽取有下面几种情况：

1）如果业务操作型数据库和数据仓库之间的数据库管理系统完全相同，那么只需要建立相应的连接关系就可以使用 ETL 工具直接访问，或者调用相应的 SQL 语句或者存储过程。

2）如果数据仓库系统和业务操作型数据库的数据库管理系统不相同，那么比较简单的方式是使用 ETL 工具导出成文本文件或者 Excel 文件，然后再进行统一的数据抽取。

3）如果需要抽取的数据量非常庞大，此时就必须考虑增量抽取。通常用标记位或者时间戳的形式，每次抽取前首先判断是否是抽取标记位或者是当前最近的时间，然后再将数据源的数据抽取出来。

2. 数据清洗

在一般情况下，数据清洗的目的就是选择出有缺陷的数据，然后再将它们正确化和规范化，从而达到用户要求的数据质量标准。其中数据“缺陷”可能包括以下几种情况：数值重复、数据缺失、数据错误、数据范围混淆、存在“脏”数据和数据不一致等几种情况，

如图 7-43 所示。其中数值重复是指标准不唯一，很多数值都代表着相同的含义。数据范围混淆是指相同的数值会应用到不同的场合中，代表着不同的含义。

第一步，需要跟业务部门进行沟通交流。为了提高数据的质量，得到标准的数据，应该首先过滤掉不符合业务要求的数据，这些数据都违背业务规则，数据清洗过程会根据业务规则去修正这些数据，每个业务规则都规定了数据必须满足的条件，然后通过 ETL 程序去修正这些不符合业务规则的数据。

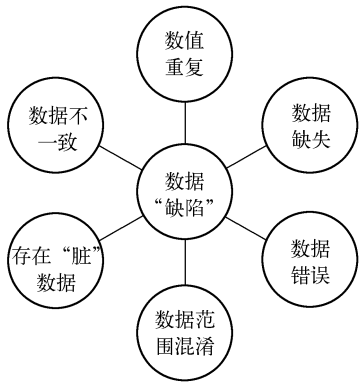


图 7-43　数据“缺陷”图

第二步，为了确保用于决策分析的数据质量，需要跟用户积极沟通，将缺失的数据补全，最后才能过滤到数据仓库中。而那些错误的数据，应该等用户完全修正后再抽取。重复的数据，同样应该等用户确认完毕后再进行抽取。我们应该理解数据清洗是一个非常费时、复杂的工程，需要多个业务部门的配合和技术开发人员对业务数据的理解，通过不断修正问题和解决问题才能完成。

数据清洗的流程包括以下几个方面，如图 7-44 所示。

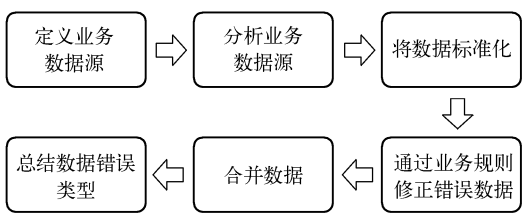


图 7-44　数据清洗的流程图

（1）定义业务数据源

标识出满足需求的数据源，并且决定什么时候进行数据清洗。

（2）分析业务数据源

分析数据源的数据是否符合业务的规则和定义，是否存在非正常的数据结构。

（3）将数据标准化

定义标准化格式的数据，并且加以转换。

（4）通过业务规则修正错误数据

定义是否为正确数据的标准，确定如何处理错误数据的方法。

（5）合并数据

将属于同一实体的多个数据进行合并，合并时应该有去重的功能。

（6）总结数据错误类型

通过总结数据出错的类型，提高清洗程序的完整性和正确性，从而降低数据出现重大问题的可能性。

3. 数据转换

数据转换是指从业务系统中抽取出源数据，然后根据数据仓库模型的需求，进行一系列

数据转换的过程。

我们已经知道数据转换是整个 ETL 过程中复杂程度相对较高的过程，包括对数据不一致性的转换，业务指标的计算和某些数据的汇总，为决策分析系统提供数据支持。其中对数据不一致性的转换就是依赖于编码表的设计，通过电压等级编码表（见表 7–3）将不同业务系统中相同类型的数据进行转换，即将各个省市的电力营销系统的电压等级编码标准化，例如将 110 kV 的编码统一设置成 1，220 kV 的编码设置成 2，380 kV 的编码设置成 3，500 kV 的编码设置成 4，1000 kV 的编码设置成 5，以消除数据仓库系统中数据存在不一致的可能。

表 7–3　电压等级编码表

电压等级/kV	电压等级编码
110	1
220	2
380	3
500	4
1000	5

通过建立程序代码编写规范，与模型设计小组共同制定编码规则，不仅可以提高数据模型的可靠性、可读性、可修改性、可维护性和一致性，而且还会提高数据模型的可继承性，促使每个人的成果可以互相共享。同时也应该建立公共的编码表作为数据转换的依据，可以根据编码表制定的业务规则进行数据的转换，保证数据仓库系统内部数据的一致性。例如，性别在客户关系表中用 1 和 0 分别代表男和女，而在单位员工表中可能使用 m 和 f 区分男和女，需要对不同业务表中相同类型的业务含义进行统一和规范。

在转换过程中，对粒度的分析也是工作的重要组成部分，因为存放到数据仓库中的数据对粒度的要求可能不相同，用户需要将低粒度的数据汇总形成决策分析型的数据，同时完成各种数据指标的计算，这都需要经过 ETL 转换过程。最后一步，将转换后得到的数据加载到数据仓库中，以供企业高层领导决策分析时使用。

ETL 转换过程可能包括以下几个方面，如图 7–45 所示。

1）对空值的处理：如果在转换过程中捕获到某些字段存在空值，那么在进行加载时需要将空值替换成某一数据或者直接进行加载，不做任何转换。

2）对数据格式的规范化：根据业务数据源中各个字段的数据类型，进行数据格式的规范和统一。例如，统一将数值类型转化成字符串类型。

3）根据业务需求进行字段的拆分或者合并。

4）对缺失数据的替换：根据业务需求对缺失数据进行替换。

5）根据业务规则对数据进行过滤。

6）根据编码表进行数据唯一性的转换：根据编码表制定的业务规范进行数据的转换，实现数据仓库系统内部数据的一致性。

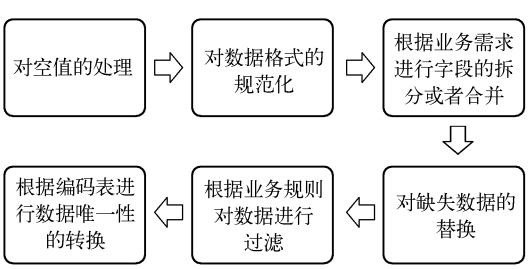


图 7–45　ETL 的转换过程

4. 数据加载

数据的加载过程就是将已经转换完成的数据存放到目标数据库的过程。这是 ETL 过程中的最后一步，需要保证加载工具必须具有高效的性能去完成数据加载，同时还需要考虑数据加载的周期和策略。数据加载策略包括时间戳的加载方式、全表对比的加载方式、通过读取日志表进行加载的方式、全表删除后再进行加载的方式，如图 7-46 所示。

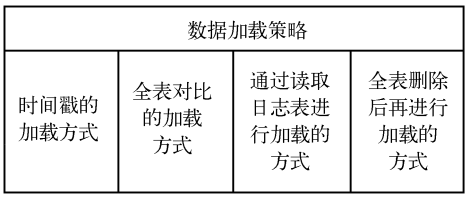


图 7-46　数据加载策略

时间戳的加载方式是通过对源系统的表添加时间戳字段，将系统当前时间和时间戳的值进行对比，决定哪些业务数据需要被抽取，可以实现数据的递增加载，是比较常见的一种加载方式。

全表对比的加载方式是在数据加载前，将每条数据都与目标表的所有记录进行全表对比，根据主键值是否相同，判断数据是更新还是插入。当数据量比较大的时候，有耗时长、效率低的缺点。通常也对全表对比进行改进，采用版本号、标记字段等缓慢变化维的形式进行增量的抽取。

通过读取日志表进行加载的方式是当源数据表发生变化时，不断更新日志表的信息，将日志表的信息作为数据加载的一个依据。日志表维护相对麻烦，会存在一定风险。

全表删除后再进行加载的方式是在数据加载前，先删除目标表的所有数据，然后去加载全部的数据，但是不能实现数据的递增加载，效率较低，实现方式却相对简单。

7.5.7　ETL 的日志

ETL 的日志功能非常重要，可以记录 ETL 执行过程中的每一步信息，包括运行的起始时间和结束时间，历史数据的抽取记录，数据抽取的行数和运行到某一步的出错信息，出错时间等内容。当然 ETL 工具是自动产生这些日志信息，帮助系统维护人员进行监控的。如果 ETL 过程中出现错误，将要形成错误日志，系统管理员可以通过邮件或者其他方式接收到该错误信息，然后对该错误及时进行处理。当然，我们已经知道 ETL 的日志信息也可以作为数据加载的一个策略，通过读取日志表的形式有计划地进行数据加载。

7.5.8　ETL 设计规范要点

ETL 设计需要遵循业务数据处理的要求，根据问题的多样性和不确定性，在设计过程中需要依照以下原则（见图 7-47）。

1）在 ETL 设计之前，需要根据业务的需求确定所要分析的主题和数据结构。

根据数据仓库的模型，考虑在 ETL 设计中是否增加预留字段和属性。

2）确定数据的粒度。可以通过粗粒度减少数据的总量，也可以根据细粒度追溯到最底层的数据，探寻原因。粒度的大小是业务需求和分析的主题所确定的。

3）确定 ETL 抽取的周期和时间。根据用户的需求，在设计 ETL 之前就应该确定抽取的时间、抽取的周期。

4）将增量抽取的方式作为 ETL 设计的重点，减少数据抽取的压力和抽取的时间。

5）通常数据的抽取和清洗可以分成许多步骤，根据不同的条件采用不同的处理逻辑。

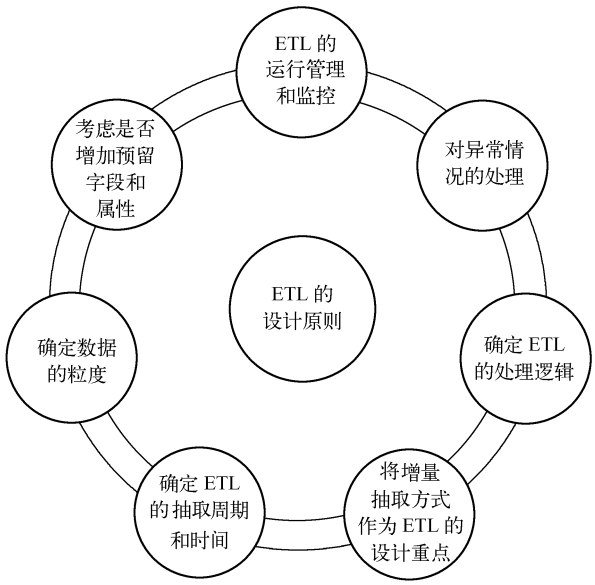


图 7-47　ETL 的设计原则

6）对异常情况的处理。网络的中断、数据流动过程中各种未知的错误，都需要通过相应的措施去解决，以保证数据的正确性。

7）对 ETL 的运行管理和监控措施。可以使用 ETL 工具中的管理监控组件对 ETL 进行设置，当 ETL 出现异常时可以进行人工干预，或者通过程序自动调度功能，对每一步的错误异常都调用相应的处理程序自动去解决，以保证数据的质量。

总结：按照以上设计原则，可以增加数据仓库系统的灵活性和扩展性，从而保证数据的正确性，降低维护成本。

7.5.9　ETL 的框架结构

ETL 的框架结构包括 ETL 调度、抽取策略、转换策略、加载策略等，如图 7-48 所示。它的每一步包括抽取、转换、加载的信息都记录到日志文件中，以便系统维护人员查看 ETL

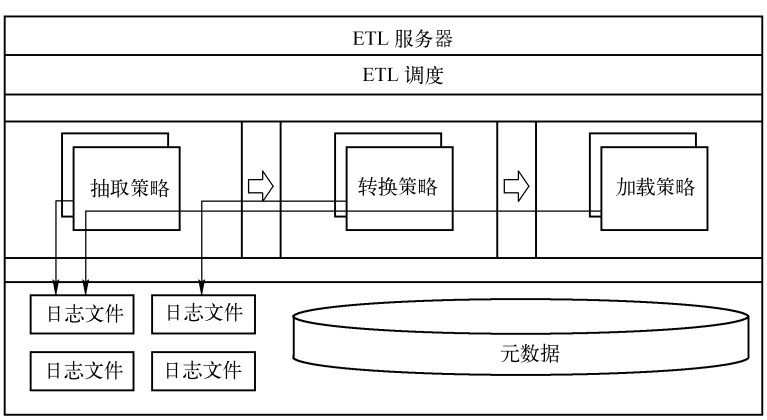


图 7-48　ETL 框架结构图

的运行信息，同时 ETL 又有异常处理的功能，对于每一步骤的异常都有相应的处理流程。

统一调度是 ETL 中较为重要的功能，通常有以下两种调度方式。

1）自动调度方式：可以使用 ETL 工具，每天定时启动后台程序，自动完成 ETL 的处理流程和加载过程。

2）手工方式：用户可以通过前台应用系统，使用它的监控功能对一些 ETL 处理程序进行手工调度。

当然，无论采用何种调度方式，都需要有报警和监控的功能，用来提醒管理人员在处理数据过程中是否出现错误。ETL 框架结构是整个商业智能系统的核心部分，占有重要地位。

7.5.10 ETL 数据加载

1. 日常增量处理

对于日常数据的增量处理有以下几种方法，如图 7-49 所示。

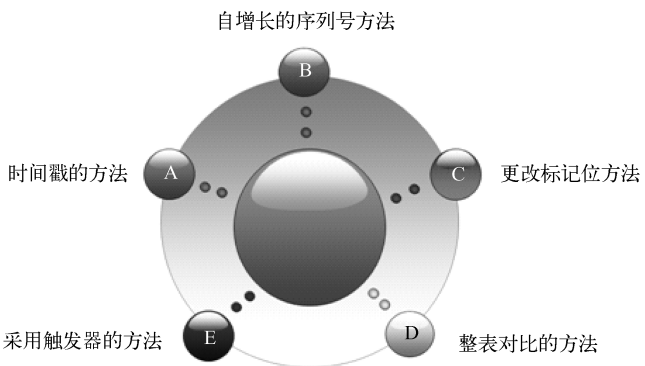


图 7-49　日常数据的增量处理方法

（1）时间戳的方法

对于交易流水信息，可以采用时间戳的方式获取增量数据。

（2）自增长的序列号方法

可以通过设置自增长序列号的方式生成唯一主键。

（3）更改标记位方法

通过定义一个字段作为数据被更改的标识。例如，设置 syn_flag 字段，初始化为 0，当记录被修改时，置为 1。

（4）整表对比的方法

对于没有时间戳的增量数据，同时数据量又不大，可以采用整表对比的方式找出增量数据，如编码表。

（5）采用触发器的方法

在源系统数据表上建立触发器，当数据项发生变化时，记录到表中，但是对业务系统会有一定的性能影响。

2. 数据初始化处理

从架构的角度来说，ETL 初始装载和日常增量加载的策略有所不同，需要考虑以下几个方面，如图 7-50 所示。

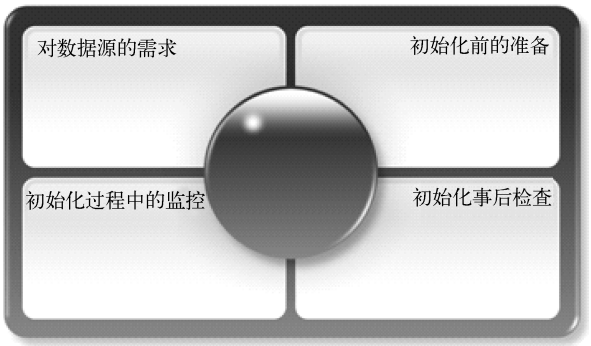


图 7-50　需考虑的因素

（1）对数据源的要求

对于 ODS 系统来说，面临的源系统可能很多，源系统需要提供初始化到上线时间点的完整信息。在 ETL 开发的同时，需要确保在规定的时间点内，ODS 系统可以完成初始化装载。ODS 系统在上线前需要将历史数据全部导入到 ODS 系统的物理表中。一些细节问题也需要考虑，例如在初始化装载前，应该先删除索引，再进行历史数据的加载，加载成功后，再重新创建索引。

（2）初始化前的准备

在初始化之前，应该对历史数据进入到物理磁盘时的容量进行估算，可以先预留较大的空间，当初始化完成后再进行缩减。

（3）初始化过程中的监控

在数据初始化过程中需要进行监控，以保证该过程能够正常运行和对错误的记录。对于一些拒绝掉的文件，应该通过事后分析，以决定是否应该重新加载，或者采用手工录入的方式。

（4）初始化事后检查

当初始化完成之后，需要对数据进行检查，以保证入库数据的准确性，可以进行自动化统计，或者由业务部门进行核对确认。

3. 错误处理与恢复

在 ETL 过程中，数据加载可能会出现各种错误，可以利用作业调度平台与监控系统对各种异常情况进行处理。

举例来说，可以在作业流程中设置异常条件，当错误记录超出一定阈值时，则需要转为人工处理。设计的原则是尽量采用自动的方式，同时根据实际情况，将自动化处理与人工处理相结合。

4. 异常情况处理策略

ETL 过程可能发生的异常包括如下几种：

1）因为硬件、操作系统或者网络等原因造成的异常。

2）目标物理模型的问题导致的异常。

3）因为人工干预导致的异常。

对 ETL 过程中的异常情况，我们应该采取哪些策略：

1）如果发生硬件、操作系统或者网络导致的异常，可以采取 ETL 中断处理，在系统运

维人员通知故障排除后，分析造成的影响，通过手工干预的方式调整 ETL 过程。

2）当物理模型发生变更时，ETL 将执行中断处理，当模型修改完成后，调整 ETL 程序，并重新进行处理。

3）生产环境应该建立合理的流程和规章制度，尽量减少人工干预的次数，降低因为人工干预造成的影响。

7.6 商业智能—OLAP 概述

OLAP（On－line Analytical Processing，联机分析处理）系统能够帮助决策分析人员从多个角度分析数据。要想理解 OLAP 的概念，必须先了解以下几个重要的概念。

1）维度：是指人们观察事物的角度，如地区维度、时间维度、产品维度等。

2）层次：根据描述维度细节程度的不同，划分数据在逻辑上的等级关系，用来描述维度的各个方面。例如，时间维度包括年、季度、月、日等层次，地区维度包括国家、省、市、县等层次。

维度和层次的关系如图 7-51 所示。

3）维度成员：维度的取值，即维度中的各个数据元素的取值。例如，地区维度中具体的成员有英国、法国、德国、西班牙。

维度和维度成员的关系如图 7-52 所示。

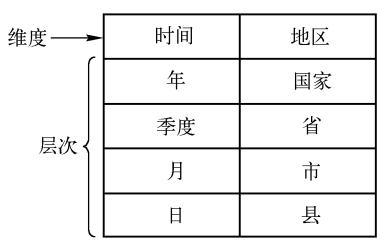


图 7-51　维度和层次的关系

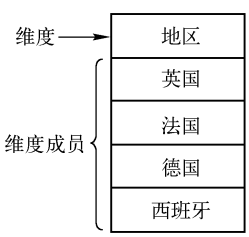


图 7-52　维度和维度成员的关系

4）钻取：通过变换维度的层次，改变粒度的大小。它包括向上钻取（Drill Up）和向下钻取（Drill Down）。向上钻取是将细节数据向上追溯到最高层次的汇总数据。向下钻取是将最高层次的汇总数据深入到最低层次的细节数据中。

5）旋转：通过变换维度的方向，重新安排维的位置，如行列互换。

6）切片和切块：在一个或者多个维度上选取固定的值，分析其他维度上的度量数据。如果其他维度剩余两个，则是切片；如果是 3 个，则是切块。

7）度量：多维数据的取值，如销售额、利润。

8）ROLAP：是基于关系型数据库的 OLAP，即以关系型数据库为基础，对多维数据的存储。

9）MOLAP：是基于多维数据库的 OLAP，其中切片、切块是主要技术。

10）HOLAP：是基于关系型和多维矩阵型等混合型的 OLAP 实现。

总结：OLAP 是针对决策分析人员和企业管理人员从多个角度对数据进行分析，随着市场竞争的日益激烈，OLAP 的应用越来越广泛，它可以从不同的角度去分析各种指标。例如，当分析企业利润指标时，可能综合时间维度、地区维度、产品类别维度、客户类别维度

等多种因素来衡量利润的值是多少，最后通过报表进行展示。OLAP 的最大特点就是通过多维模型，用户可以动态地从多个角度分析数据，增加了分析的灵活性和时效性，大大提高了企业管理的效率，这是 OLAP 发展的根本原因之一。

7.6.1 OLAP 系统与 OLTP 系统的区别

OLTP（在线联机事务处理）系统主要面向细节性的数据，存储的都是当前的数据，用来支持日常业务运作。这些数据都是可以更新的，数据处理量相对较小。OLAP 系统主要是综合的、并且经过提炼的数据，而且主要是历史数据，不可修改，数据处理量相对较大，主要面向决策分析处理。它们的区别如图 7-53 所示。

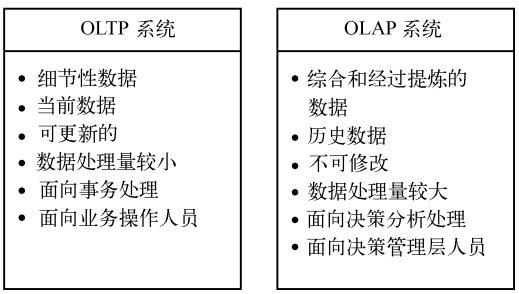


图 7-53　OLTP 系统和 OLAP 系统的区别

7.6.2 OLAP 的实现方法

OLAP 有多种实现方法，根据存储数据方式的不同，可以分为 MOLAP、ROLAP、HOLAP，如图 7-54 所示。

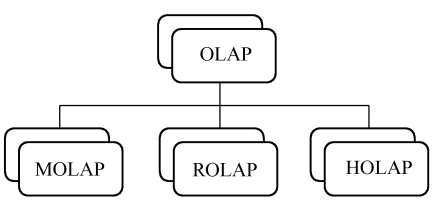


图 7-54　根据存储方式划分的 OLAP 的实现方法

ROLAP（Relational OLAP）表示基于关系型数据库的 OLAP 实现。它的技术依赖于关系型数据，以关系型数据库为核心，以关系型结构对多维数据进行数据存储和展现。通常 ROLAP 将多维数据分成事实表和维表，事实表存储的都是指标数据和维表的关键字段值，维表多数存储维度的层次、维度的成员值等信息。事实表以存储的产品 ID、产品类型 ID、地址 ID 和时间 ID 作为连接维表的关键字段，以销售数量作为指标数据。

维表包括产品维表、时间维表、产品类型维表、地理位置维表。维表和事实表通过主外关键字关联在一起，形成了星形模式，如图 7-55 所示。

对于层次复杂的维，可以使用多个表来描述，这种对星形模式的扩展称为雪花形模式。事实表以存储的产品 ID、产品类型 ID、地址 ID 和时间 ID 作为关联维表的关键字段，以销售数量作为指标数据。维表有产品维表、时间维表、产品类型维表、地址维表，地址维表又包括国家、省级、地市等维表，如图 7-56 所示。通过最大限度地减少数据存储量以及关联

较小的维表来改善数据查询的性能，这是典型的雪花形模式。

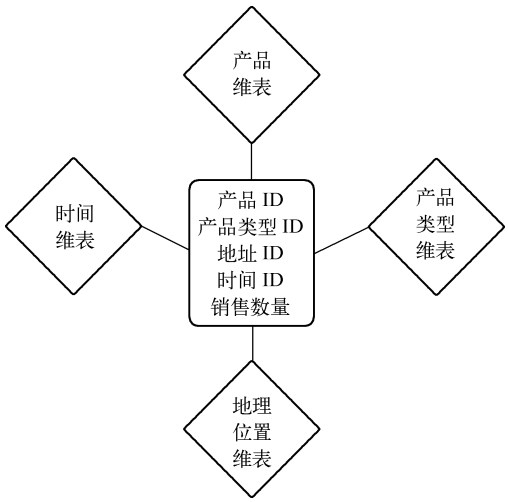


图 7-55　ROLAP 的多维关系图（星形模式）

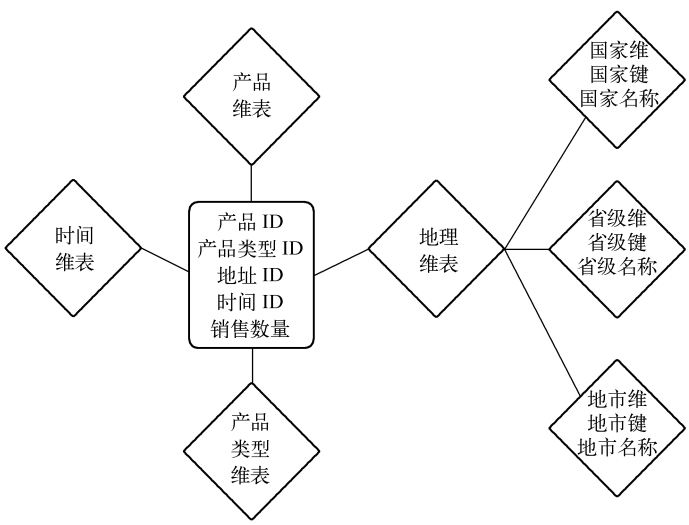


图 7-56　ROLAP 的多维关系图（雪花形模式）

MOLAP（Multidimensional OLAP）表示基于多维数据的 OLAP 实现。它的技术手段主要有“切块”、“切片”，数据检索速度较快，但是生成立方体的时间较长，数据存储在多维立方体中。MOLAP 多维立方体如图 7-57 所示。

HOLAP（Hybrid OLAP）表示基于混合型的 OLAP 实现。它的技术主要结合 MOLAP 和 ROLAP 两种技术的优点。

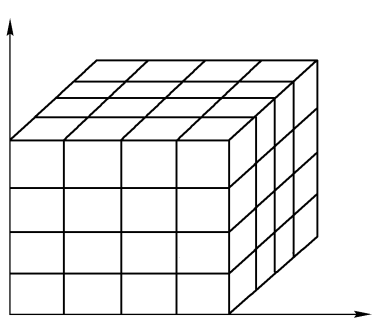


图 7-57　MOLAP 多维立方体

7.6.3 OLAP 的基本目标和特点

OLAP 的基本目标就是支持决策分析和多维数据查询。OLAP 通过对信息的各种形式的存取，满足企业决策人员和管理人员对复杂查询的处理，并且将结果提供给决策分析人员，使他们对企业的运营状况有更深入的了解，能够制定出正确的决策方针。OLAP 是引领企业发展的“灯塔”。

OLAP 系统的特点包含以下几个方面。

1）丰富的报表展示功能：OLAP 系统一般有丰富的报表展示功能，如柱形图、折线图、饼形图。

2）数据访问和多维分析的能力：提供给用户数据访问和多维分析的能力，并以用户希望的方式进行展示。

3）快速的数据分析能力：OLAP 系统有秒级的数据分析能力。

OLAP 的特点如图 7-58 所示。

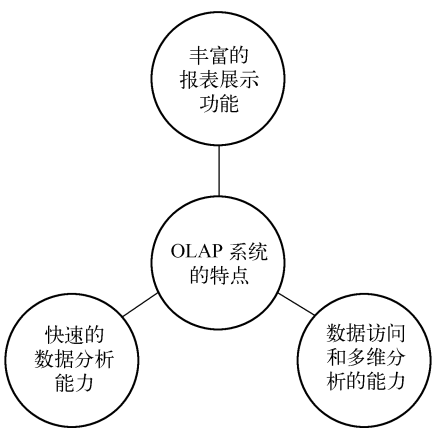


图 7-58　OLAP 的特点

7.6.4 建立 OLAP 的过程

建立 OLAP 的过程如图 7-59 所示。

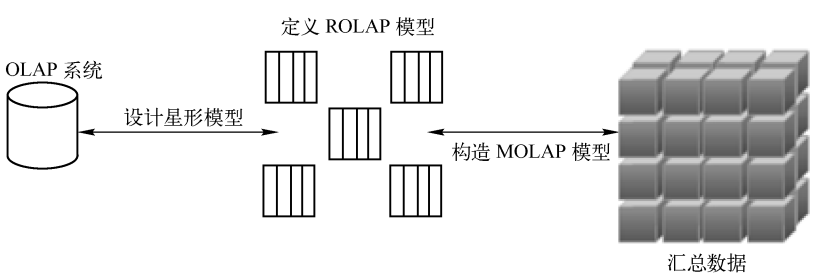


图 7-59　建立 OLAP 的过程

为了提高 OLAP 系统的效率，设计数据仓库时应该考虑如下因素：

1）为事实表和维度表中的关键字创建索引，以提高数据查询的效率。同一类数据尽可

能使用一个事实表，以减少表之间的关联。

2）事实表中尽量不要包含汇总类型的数据。

3）维表的设计应该符合第三范式的约束，维表中不要存储无关的数据。

4）数据仓库设计的好坏直接影响建立 OLAP 系统的难易程度和效率，同时 OLAP 系统又是数据仓库系统的一种多维展现方式。

7.6.5 OLAP 的实施过程

OLAP 系统的实施一般过程（见图 7-60）包括以下几个步骤：

1）源系统经过 ETL 过程装载到 ODS 数据缓冲区中，目的是将所有的业务数据集成起来。

2）从 ODS 数据缓冲区中将数据抽取到 ODS 统一信息视图区，目的是使用户能够通过 ODS 统一信息视图区获得跟某个主题域相关的实时数据。

3）将数据从 ODS 统一信息视图区抽取到数据仓库中。

4）数据集市里的数据在数据仓库中经过转换、汇总计算获取，直接支撑 OLAP 多维分析。

5）最后 OLAP 系统支持多维数据分析。

其中 ODS 数据缓冲区和 ODS 统一信息视图区可以合并成一个，同时具有数据缓冲和集成的功能。

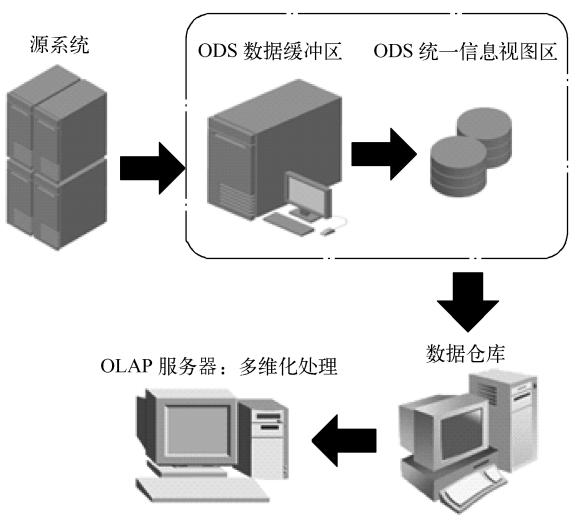


图 7-60　OLAP 系统的实施过程

7.6.6 OLAP 模型的设计与实现

（1）了解用户的需求

作为解决方案的提供者，我们需要理解业务规则，了解当前的业务状况，不仅需要和系统相关人员进行交流，还需要和系统的设计者和开发者进行沟通。

首先，对用户进行分类，理解用户对数据的可用性和访问速度的要求。其次，需要了解

不同用户对系统的访问频率，每类用户的数量和需要分析的数据量是多少。再次，需要大致清楚系统的数据总量应该是多少。当我们已经了解了用户和数据源的基本情况后，可以考虑系统能够满足客户的需求有哪些。用户、开发者、管理者是通过需求文档进行交流沟通的，如图 7-61 所示。而用户最关心的问题就是开发者是否完成了需求文档所要求的功能特性。

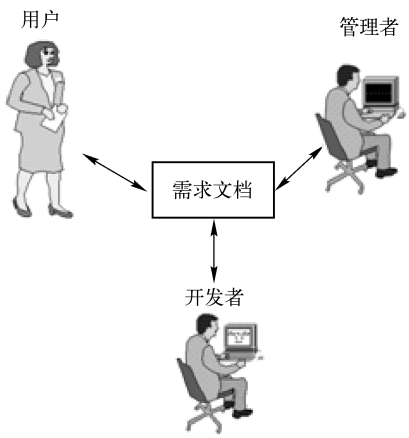


图 7-61　了解用户的需求

（2）逻辑模型的设计

通常所说的数据模型一般有两个层次：逻辑层、物理层。逻辑模型描述现实世界的内在规律和业务规则。物理模型描述数据库内部存储的具体实现。OLAP 模型是一个逻辑概念，主要是对数据进行多角度的分析，以便为企业决策者和管理者提供各种信息和知识。“多维结构”是 OLAP 世界的核心，而多维模型通过维度、层次、度量三者之间的关系分析数据。

举例来说，如果有一个销售系统，度量值可能包括销售额、成本、利润，维度包括时间、产品类型。OLAP 逻辑模型的设计就像是桥梁，一端是用户的需求，另一端是业务数据源。销售额、成本、利润是需要展现、存储的内容，随着时间、产品类型的变化而变化。

（3）OLAP 的分析过程

首先根据逻辑模型定义 OLAP 多维模型，在定义模型的过程中，需要根据业务需求定义“立方体”，分析方法有“切片”和“切块”。例如，在“利润、地区、时间”三维立方体中进行切块和切片，可得到各地区、各产品的销售利润情况，如图 7-62 所示。

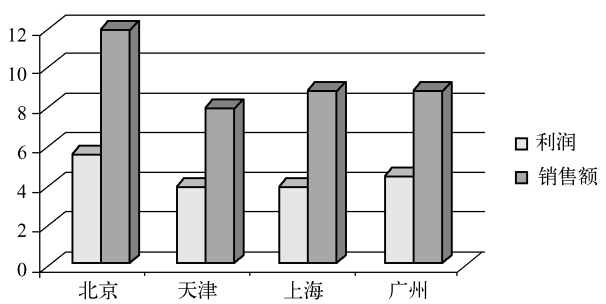


图 7-62　商业智能 OLAP 的例子

7.7　传统商业智能和未来商业智能的关系

对于传统商业智能和未来商业智能的区别，有这样一种观点：如图 7-63 所示，传统的商业智能只是提供类似汽车后视镜的作用，只能看到行驶过的路程，而不能看到远方，即传统商业智能通过查看历史数据，分析以前的情况。事实上，传统的商业智能有查询、报表展示、多维分析、数据挖掘的内容，已经包含了预测分析的能力。因此，上述观点是不正确的。

图 7-63　传统的商业智能和未来商业智能区别的某种观点

而未来商业智能真正要解决的问题是：

1）建立实时动态的数据仓库。因为传统意义上的数据仓库是基于历史数据分析，而动态数据仓库是基于前端应用，增加对非结构化数据的处理，可以大大缩短响应的时间。

2）用户对数据可视化的要求会越来越多，同时增强了对商业智能实时性的要求，甚至在将来，人们可以将商业智能转移到手机移动终端上，实现动态分析和实时分析等。

小结

- 我们对商业智能做一个简单的定义，那就是：帮助用户把一些数据转化成具有商业价值的，而且可以获取的信息和知识，同时在最恰当的时候，通过某种方式把信息传递给需要的人。从专业的角度来说，商业智能就是利用数据仓库、数据分析和挖掘技术，以抽取、转换、查询、分析和预测为主的技术手段，帮助企业完成决策分析的一套解决方案。
- 商业智能的实施方法：

1）项目规划：主要包括项目前期的准备、业务现状的调研、目前系统的现状分析。分析内容包括业务需求的定义和系统实现的目标，系统运行环境的定义，系统的框架结构定义，逻辑模型的设计等。

2）系统设计与实现：主要包括系统体系结构的设计，物理数据库的设计，数据抽取、转换和加载的实现，前端应用的开发，元数据的管理等内容。

3）系统调优：指逻辑、物理模型的调整，系统性能的调优。

4）系统运行及维护：指编写系统运行及维护手册，以及用户操作手册、培训教材等文档。

- 商业智能的实施步骤：

（1）定义需求

需求分析是商业智能项目重要的一步，需要描述项目背景与目的、业务范围、业务目标、业务需求和功能需求等内容，明确企业对商业智能的期望和需要分析哪些主题等方面。

（2）数据仓库模型的建设

在系统设计、开发之前，业务人员和设计人员共同参与概念模型的设计，核心的业务概念在业务人员和设计人员之间达成一致。在系统设计开发时，业务人员和系统设计人员共同参与逻辑模型的设计。最后设计开发人员以逻辑模型为基础进行物理模型的设计。

（3）数据抽取、清洗、转换、加载

抽取主要负责将数据仓库需要的数据从各个业务系统中抽取出来。如果每个业务系统的

数据情况各不相同，可能对每个数据源都需要建立独立的抽取流程，每个流程都需要使用接口将源数据传送给下一环节，即清洗与转换阶段。通过数据抽取程序，可以从业务源系统中不断地将数据抽取出来，抽取周期可以设定为某个固定时间。

（4）建立商业智能分析报表

商业智能分析报表通过对数据仓库的数据分析，使企业的高层领导可以多角度地查看企业的运营情况，并且按照不同的方式去探查企业内部的核心数据，从而更好地帮助企业决策人员对公司未来经营状况进行预测和判断。

- 商业智能项目成功的关键因素：

1）企业高级领导层对商业智能项目的支持和雄厚的资金是项目成功的关键因素之一。

2）拥有实力雄厚的技术团队。技术团队成员不仅精通商业智能相关技术，同时也熟悉相关的业务规则和开发流程。

3）商业智能项目团队的协同合作能力。项目的管理者需要保证团队中每个成员分工明确，沟通及时，并且需要各部门之间有良好的合作能力。总之，商业智能项目的实施是一个长期的不断完善的过程。

- 完整的商业智能系统需要以下几种核心的技术：

（1）数据仓库

（2）数据挖掘和分析

（3）ETL 处理技术

（4）联机分析处理（OLAP）技术

（5）可视化分析

（6）大数据技术

（7）商业智能元数据管理

- 数据仓库是一个面向主题的、集成的、非易失的、反映历史变化的、随着时间的流逝发生变化的数据集合，它主要用来支持企业管理人员的决策分析。
- 数据集市就是满足特定的部门或者用户的需求，按照多维的方式进行存储，包括定义维度、需要计算的指标、维度的层次等，生成面向决策分析需求的数据立方体。数据仓库体系结构中增加了数据集市，数据集市又可以看做部门级的小型数据仓库。
- ODS（Operational Data Store，操作数据存储）是一个面向主题的、集成的、可变的、反映当前细节的数据集合。它主要用于支持企业处理业务应用和存储面向主题的、即时性的集成数据，为企业决策者提供当前细节性的数据，通常作为数据仓库的过渡阶段。
- ODS 的设计原则包括可扩展性、高可用性、可重用性和高性能。
- ETL 是数据抽取（Extract）、转换（Transform）、加载（Load）的英文简写。它的一般过程是指：首先将源数据抽取出来，然后经过数据的清洗、转换，最后加载到目标表中。ETL 过程一般都是批量操作的。
- 维度：是指人们观察事物的角度，如地区维度、时间维度、产品维度等。
- 层次：根据描述维度细节程度的不同，划分数据在逻辑上的等级关系，用来描述维度的各个方面。例如，时间维度包括年、季度、月、日等层次，地区维度包括国家、省、市、县等层次。

- 维度成员：维度的取值，即维度中的各个数据元素的取值。例如，地区维度中具体的成员有英国、法国、德国。
- 钻取：通过变换维度的层次，改变粒度的大小。它包括向上钻取（Drill Up）和向下钻取（Drill Down）。向上钻取是将细节数据向上追溯到最高层次的汇总数据。向下钻取是将最高层次的汇总数据深入到最低层次的细节数据中。
- 旋转：通过变换维度的方向，重新安排维的位置，如行列互换。
- 切片和切块：在一个或者多个维度上选取固定的值，分析其他维度上的度量数据。如果其他维度剩余两个，则是切片；如果是 3 个，则是切块。
- 度量：多维数据的取值，如销售额、利润。
- ROLAP：是基于关系型数据库的 OLAP，即以关系型数据库为基础，对多维数据的存储。
- MOLAP：是基于多维数据库的 OLAP，其中切片、切块是主要技术。
- HOLAP：是基于关系型和多维矩阵型等混合型的 OLAP 实现。

第 8 章　商业智能架构实践

本章目标

通过前几章的学习，我们了解了商业智能的定义、商业智能的功能、商业智能的发展趋势、商业智能的实施方法和步骤、关于商业智能的核心技术、数据仓库理论、数据仓库的特点、数据集市理论、ODS 理论等知识。

学习本章后，读者将掌握：

- 商业智能架构原则
- 商业智能架构典型应用
- 商业智能具有的功能
- 商业智能未来的发展趋势和方向
- 商业智能的传统架构
- 传统商业智能的特点
- 未来商业智能的特点
- 旅游行业 - 分析型客户关系管理的商业智能体系
- 分析型客户关系管理商业智能体系架构
- 实时的商业智能架构
- 电信行业实时商业智能架构体系

8.1　商业智能架构概述

8.1.1　商业智能架构原则和典型应用

商业智能的建设是一个战略性的工程，它直接影响到企业未来的发展方向，对于商业智能的架构应该遵循以下几项原则，如图 8-1 所示。

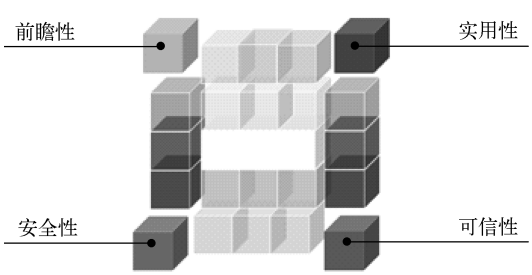


图 8-1　商业智能架构原则

- 前瞻性

商业智能应该建立在可信的数据基础之上，完成商业智能的建设需要投入大量的人力和

财力，具体来说，在开发技术和手段上具有先进性和灵活性。在业务上，需要满足未来竞争的要求。

● 实用性

建立商业智能的目的是服务于决策过程，前期带来的效益可以促进后期的开发，在开发时选择工期较短、重要的和见效快的部门作为突破口，保障系统的实用性和可操作性。

● 安全性

因为商业智能可能会涉及机密数据，所以必须保证其安全性。特别是在查询系统多样化的情况下，商业智能必须符合安全性的要求。

● 可信性

商业智能作为决策支持系统，同时产生大量的报表。商业智能系统应该具有可信性。

下面我们分析一下关于商业智能的典型应用。

商业智能是收集、管理和分析数据，同时将数据转化成有用信息的过程，如图 8-2 所示。

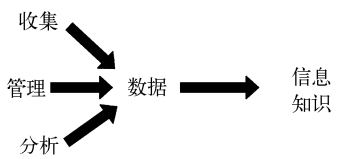


图 8-2　商业智能的典型应用过程

商业智能系统从企业的日常数据中开发基于事实的信息，辅助企业做出更好的商业决策，提高企业运营效率和决策分析的能力。可以帮助企业完成风险分析、欺诈监测、财务分析等。商业智能系统是一个决策支持系统，它是在数据仓库的基础上，利用各种挖掘工具获得信息和知识。目前来说，金融行业、通信行业、制造行业、零售行业、医疗行业、政府机构等已经逐步开始应用商业智能。

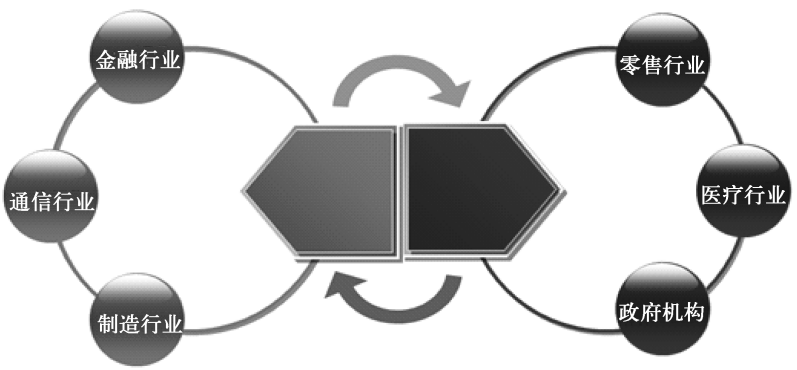


图 8-3　商业智能的行业应用

商业智能的典型应用包括经营分析、绩效管理、战略决策支持、产品管理和创新、客户关系管理和风险管理等，如图 8-4 所示。

（1）经营分析

对于企业的经营分析可以包括指标分析和财务分析等内容。指标分析是针对业务流程相关指标的分析。例如，销售率、利润率和库存量等。财务分析是针对财务数据中的费用支出、利润等指标的分析。

（2）绩效管理

企业管理人员利用商业智能工具衡量员工的工作绩效情况。

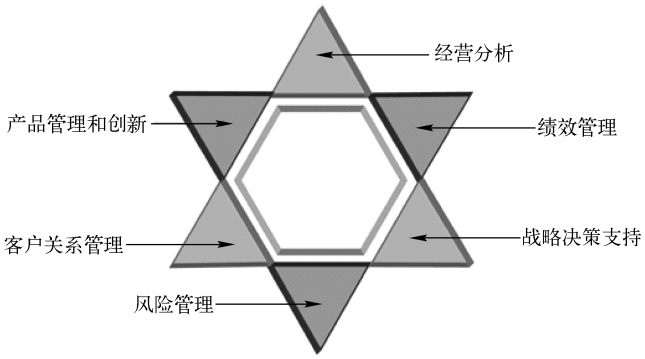


图 8-4　商业智能的典型应用

（3）战略决策支持

通过对各类数据的高度概括和分析，辅助企业高层进行战略决策。

（4）风险管理

利用商业智能技术，降低企业的风险。例如，通过发现客户的异常情况，快速采取措施，提高企业的抗风险能力。

（5）客户关系管理

利用商业智能技术，分析客户的购买习惯和喜好，改进服务和产品的质量，提高客户的忠诚度。

（6）产品管理和创新

利用商业智能技术，通过对历史数据的分析，加强对产品的改进能力和管理能力，同时提高产品的创新能力和推广能力。

8.1.2　商业智能具有的功能

商业智能产品应该建立在稳定的平台上，它可以提供数据关联分析的功能、数据监控的功能、数据展示功能和数据输出功能，如图 8-5 所示。

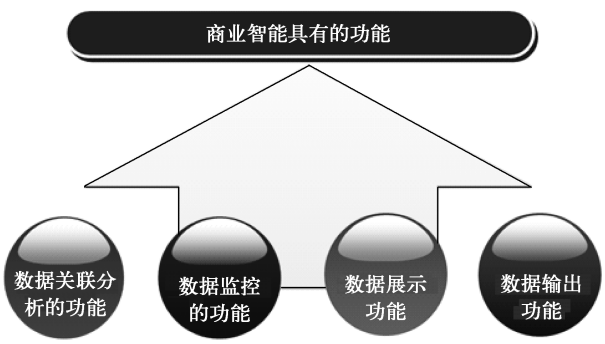


图 8-5　商业智能具有的功能

（1）数据关联分析的功能

关联分析用于发现事物之间的关联性，当一个事件发生时，另一个事件也可能会发生。目的是发现有实用价值的事件。例如，对于商业银行的客户，分析可能进行股票交易和债券

交易的概率，扩展产品范围，吸引更多的客户。

（2）数据监控的功能

可以设置条件，使符合条件的数据显示出来，引起管理人员的注意。

（3）数据展示功能

将结果数据以某种形式展示出来，以支持客户的数据分析和决策。

（4）数据输出功能

将结果数据以某种形式输出，以支持客户的数据分析和决策。

8.1.3 商业智能未来的发展趋势和方向

传统商业智能具有以下几个特点：查询、报表、多维分析和统计分析、数据挖掘，如图 8-6 所示。

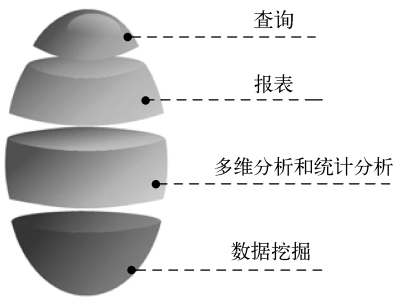


图 8-6　传统商业智能的特点

但是传统的商业智能具有以下几个方面的局限性：

1）传统商业智能的上钻、下钻和比较功能很难满足一些特殊用户的分析需求。

2）传统商业智能的数据准确性、实时性经受着重大的考验。

3）传统商业智能很难处理庞大的数据，只有通过大数据技术才能访问和使用海量的数据，以及各种非结构化数据。

从根本上来说，传统商业智能更侧重历史分析，而未来商业智能更专注于对业务流程的整合，以实现动态分析和实时分析。

举例来说，传统商业智能更加擅长于对历史数据的同期对比、产品分析、企业的绩效管理和统计报表分析等内容，如图 8-7 所示。

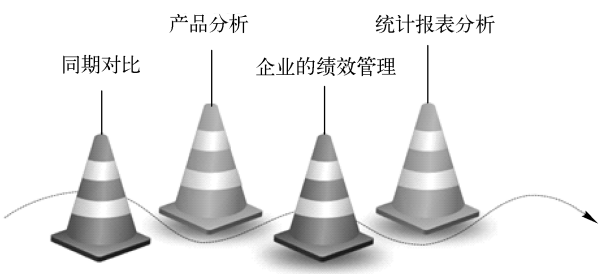


图 8-7　传统商业智能的优势

对于未来商业智能来说，它更专注于对企业的风险管理、提供各种实时报表和实时服务、实现实时或者准实时的精准营销、完成对业务的监控功能等，如图 8-8 所示。

我们总结一下商业智能的发展趋势：

对于传统型的商业智能，主要是基于历史数据做出决策和分析。它面向企业的决策者和分析者，主要以查询为主。

对于未来商业智能，主要是基于实时的数据做出分析和决策。它可以面向一线的客户经理和决策者，通过实时捕获的数据，获取最新的信息和知识。它可以提高商业智能对业务的

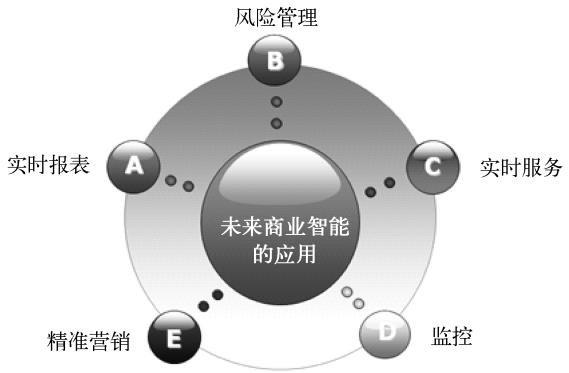


图 8-8　未来商业智能的应用

即时指导作用，同时快速地响应事件，提升企业的竞争力。一般来说，统计报表分析表示已经发生了什么，OLAP 分析和即席查询代表着为什么发生，数据挖掘会预测将来发生什么，而未来商业智能将要解决正在发生什么，如图 8-9 所示。

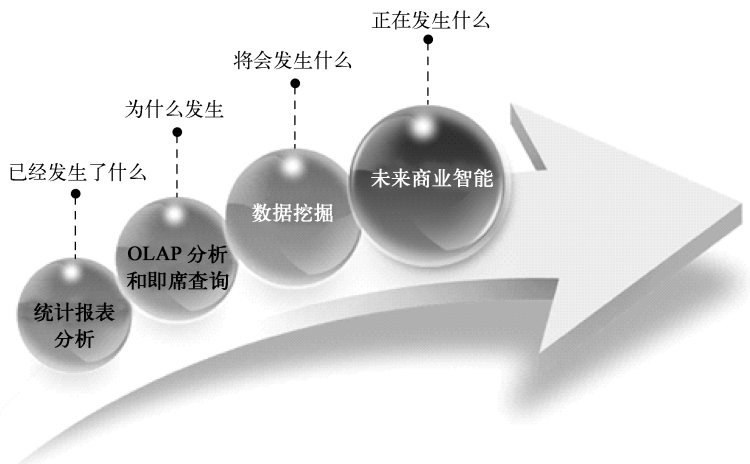


图 8-9　未来商业智能的作用

未来商业智能的方向：

1）建立实时动态数据仓库，一般来说，传统数据仓库是基于历史数据进行分析的，实时动态数据仓库支持前端应用，大大缩短了响应时间。

2）支持大数据技术，增加对非结构化数据的处理。

3）用户已经不再满足于传统的数据展现，要求数据进一步可视化。

4）对于预测分析、假设模拟和数据挖掘技术的应用将会越来越广泛。

5）用户对商业智能的实时性需求越来越多。

8.1.4　商业智能的传统数据架构

商业智能的传统架构类似于传统的物流过程，即各地运来的货物首先存放在暂存库，主要目的是对各类货物进行清洗、筛选、检查、贴标签等工作，然后统一发往货仓，最后在各

个超市中进行集中销售，如图 8-10 所示。

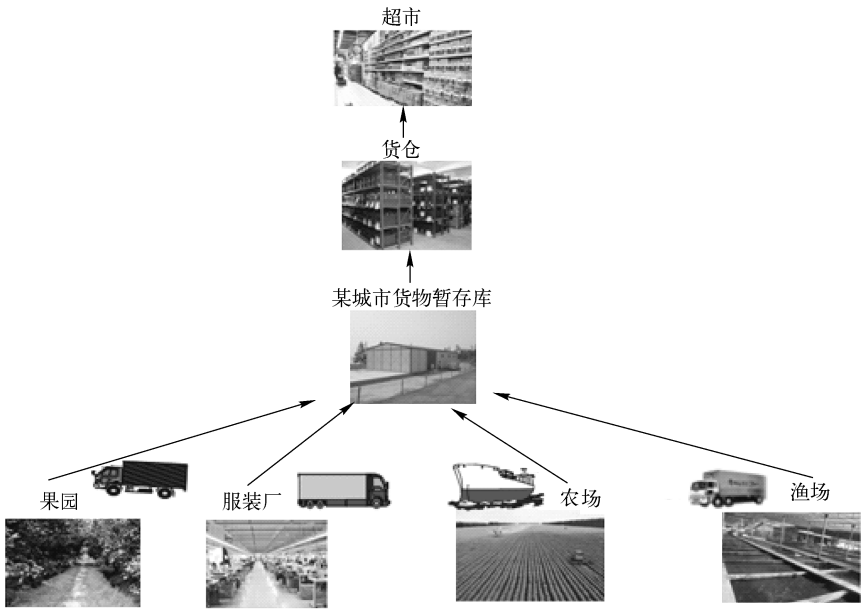


图 8-10　商业智能的传统架构类似于传统的物流过程

其中，果园、服装厂、农场、渔场类似于各个业务系统，货物暂存库的功能与 ODS 系统相似，货仓相当于数据仓库系统，而超市类似于数据集市系统，如图 8-11 所示。

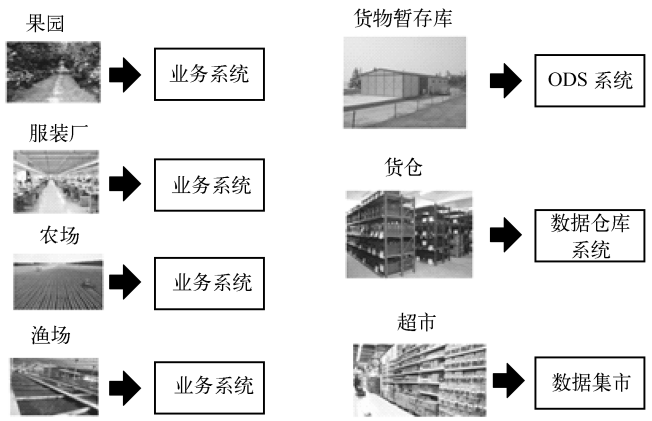


图 8-11　商业智能的传统架构与传统物流过程的映射关系

随着商业智能使用越来越广泛，使用者已经不再局限于业务分析人员或者高层领导，可以通过外部网络延伸到企业的客户、合作伙伴等。为了满足这些用户的需求，商业智能架构需要满足可扩展性和可靠性，同时保证快速的响应能力。

对于传统的商业智能体系，底层是软硬件平台、安全管理和元数据管理等。商业智能的工作流和数据流分别是数据源、数据整合、数据处理、分析和应用。综上所示，我们可以得到商业智能的传统数据架构的分布情况，如图 8-12 所示。

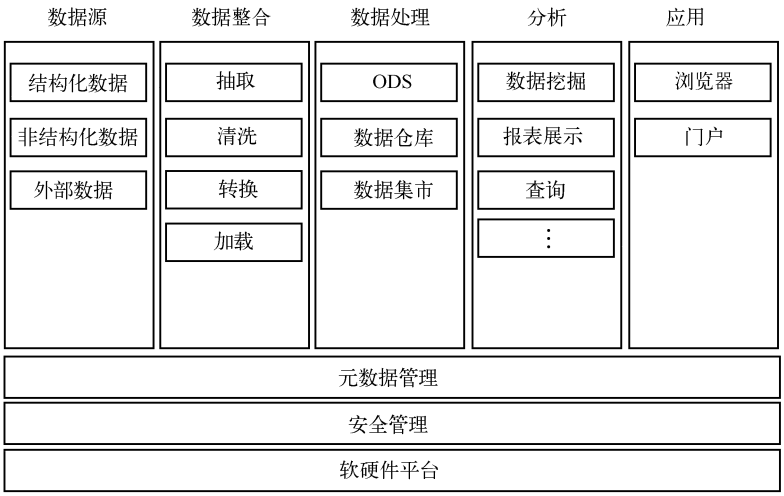


图 8-12　商业智能的传统数据架构的分布

商业智能的处理过程是从各个业务系统或者其他数据源中抽取有用的数据，然后对采集的数据进行清洗、转换和加载，以保证入仓之前的数据是完整的、一致的，经过重构之后，将数据存储到数据仓库或者数据集市中。数据仓库的数据反映的是企业的整体情况，最后利用数据挖掘工具、OLAP 分析工具对数据进行处理，完成数据到信息和知识的转变。

传统商业智能数据架构中的数据流转如图 8-13 所示。

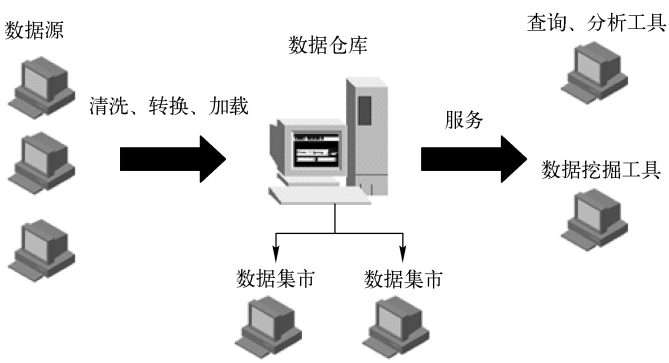


图 8-13　传统商业智能的数据流转

对于数据仓库的建设，是以业务系统和大量的业务数据积累为基础，然后将这些数据进行整理和归纳，提供给决策分析人员。数据仓库建设是一个工程，主要包含企业内部信息和外部信息。内部信息包括各种业务处理数据和各类文档数据，外部信息包括各类市场信息、各种手工收集的信息等。

数据仓库的关键是数据的存储和管理。针对各类业务数据，进行抽取、清理和集成，按照主题进行组织。可以按照多维模型进行组织，分析。

其中前端工具主要包括报表工具、查询工具、数据分析工具、数据挖掘工具等应用开发工具，如图 8-14 所示。

传统商业智能体系和未来实时商业智能体系的区别如图 8-15 所示。

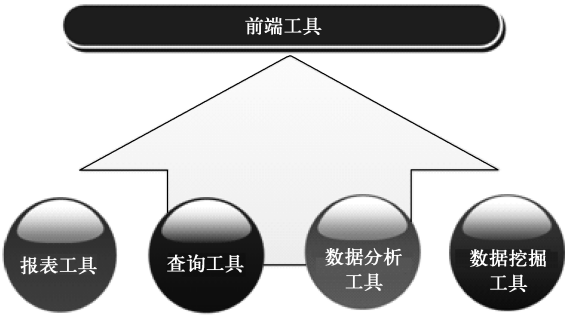


图 8-14　前端工具

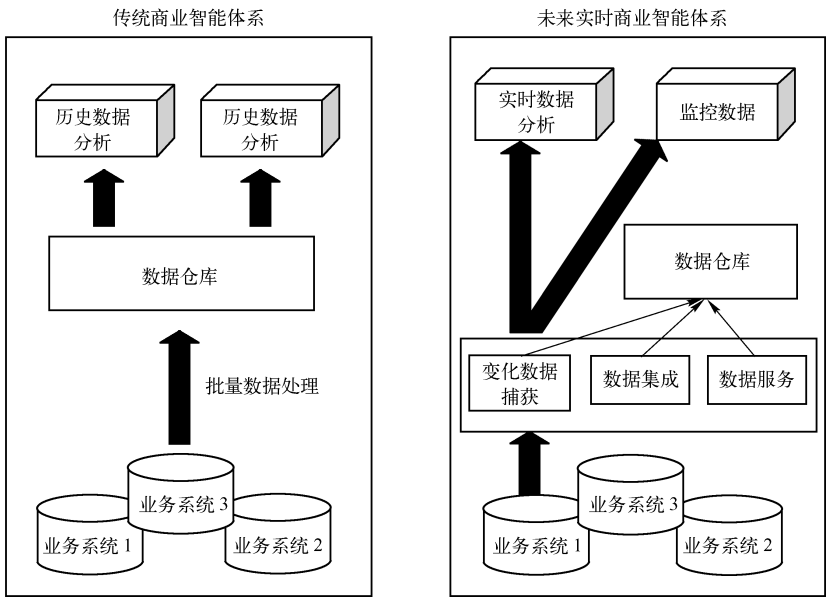


图 8-15　传统商业智能体系和未来实时商业智能体系的区别

传统商业智能体系主要是对历史数据的分析。数据仓库作为前端应用主要的数据源。

未来实时商业智能体系主要是对实时数据的分析和监控数据。它可以快速捕获变化的数据。数据仓库作为前端应用的部分数据源。

8.2　未来商业智能的架构

8.2.1　旅游行业分析型客户关系管理的商业智能体系

目前，旅游行业遇到了很多问题和挑战，如图 8-16 所示。

(1) 市场竞争不断加剧

在许多地方，旅游行业已逐渐发展成国民经济的战略性支柱产业，导致竞争异常激烈。

(2) 部分业务收入开始下降

因为受到多重因素的影响，旅游行业中的部分子行业收入开始下降，传统的粗放型营销策略已经不能适应旅游行业的未来发展。

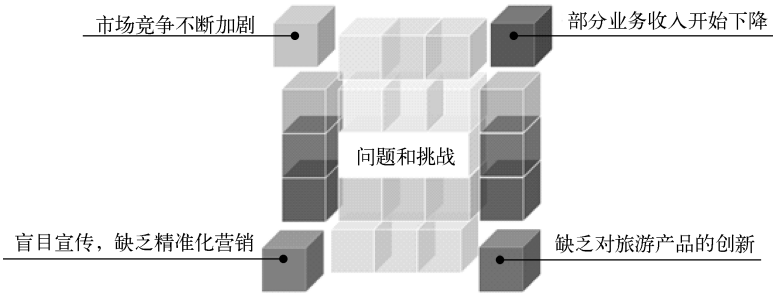


图 8-16　旅游行业目前遇到了很多问题和挑战

（3）盲目宣传，缺乏精准化营销

很多旅游公司没有真正了解消费者的需求，广告宣传缺乏创意，没有达到真正的销售拉动作用，不清楚目标群体的特征、喜好，不能及时地将信息送到目标人群中，缺乏精准化营销。

（4）缺乏对旅游产品的创新

在旅游行业中，普遍存在的问题是缺乏对旅游产品的创新，无法吸引消费者。

新的问题和挑战，对旅游机构提出了以下更高的要求。

1）在市场竞争不断加剧的情况下，旅游机构应该建立一体化的客户营销体系，为客户提供个性化服务，细分目标客户，增强客户满意度，提高营销的精准度。

2）部分业务收入开始下降的情况下，旅游机构应该采取精细化的客户发展策略，满足客户个性化的需求，提高客户的忠诚度，让旅客有不同的体验和感受，从而乐于重复消费，以增加行业的收入。

3）对于旅游机构的营销宣传，需要结合消费者对产品的印象、喜好和市场的实际情况。同时包括对目标群体的需求和竞争对手情况的掌握，提高营销策划的科学化程度。

4）旅游产品项目需要不断拓新，增加与消费者之间的互动，提高产品的精细化程度。另外，可以不断进行新产品设计，提供个性化产品。

面对问题、挑战和更高的要求，基于挖掘技术，建立分析型客户关系管理的商业智能体系，以解决精准营销和产品的创新问题，某旅游机构的建设流程如图 8-17 所示。

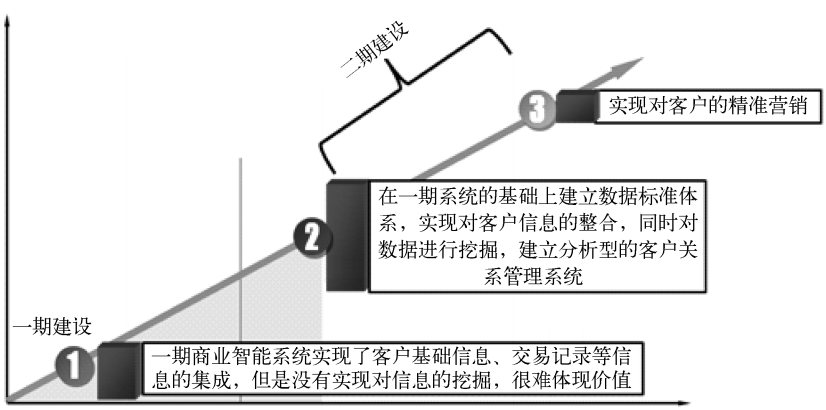


图 8-17　商业智能体系解决精准营销和产品的创新问题

1）一期商业智能系统实现了客户基础信息、交易记录等信息的集成，但是没有实现对信息的挖掘，很难体现价值。

2）在一期系统的基础上建立数据标准体系，实现对客户信息的整合，同时对数据进行挖掘，建立分析型的客户关系管理系统。

3）最后，在此基础上，增加客户体验和推送成功率，由被动营销改为主动营销，对客户信息进行评估与深入挖掘，主动向客户推送一些旅游产品，从而实现对客户的精准营销。

分析型客户关系管理商业智能数据架构如图 8-18 所示。

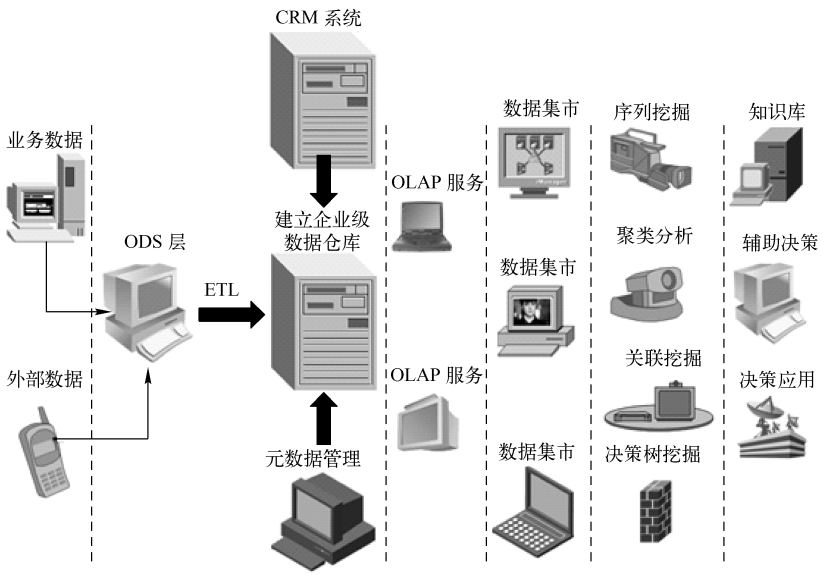


图 8-18　分析型客户关系管理商业智能数据架构

其中分析型客户关系管理系统的数据流转如图 8-19 所示。

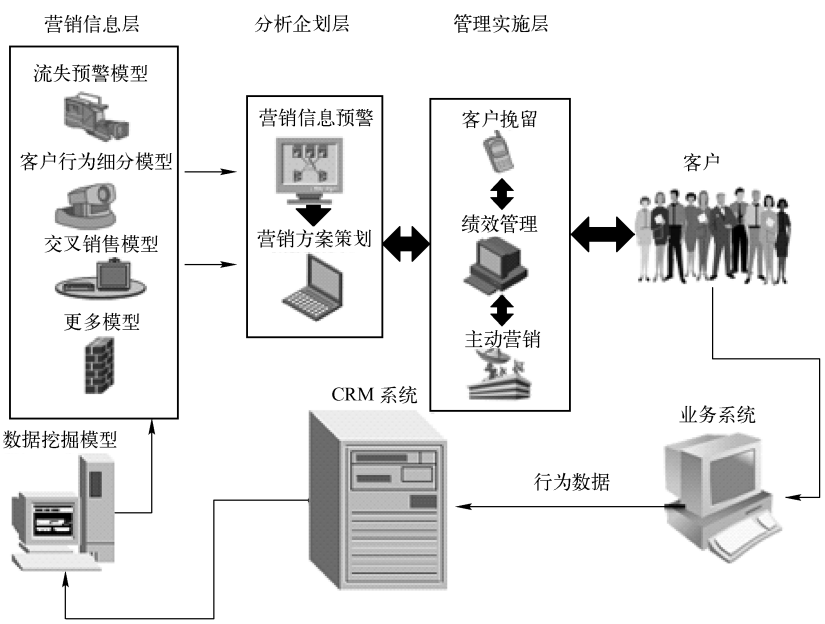


图 8-19　分析型客户关系管理系统的数据流转

8.2.2 电信行业实时商业智能架构体系

电信行业实时商业智能应用架构如图 8-20 所示。

图 8-20　电信行业实时商业智能应用架构

- 监控

监控包括用户欺诈监控、服务监控、高额话费预警等内容。

- 数据查询

数据查询包括话费财务报表、统一客户视图查询、产品套餐销售报表等内容。

- 多维分析

多维分析包括客户分析、竞争分析、产品与套餐分析、增值业务营销分析、人力分析、财务分析等。

电信行业实时商业智能数据架构如图 8-21 所示。

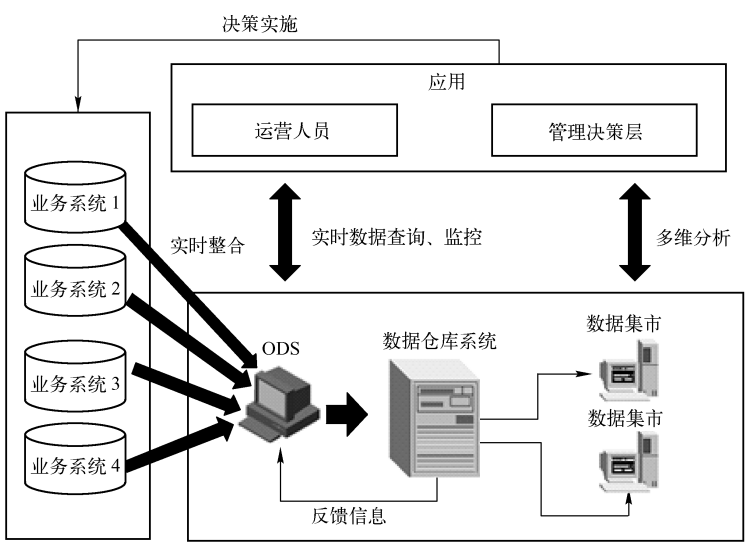


图 8-21　电信行业实时商业智能数据架构

这种实时的商业智能架构的目的是在合适的时机，通过合适的渠道，向客户推送合适的产品和服务。这也要求我们从以“产品为中心”向以“客户为中心”转变。

我们可以分析客户的特征，规划其产品，选择客户喜爱的营销渠道，在适当的时候对客户进行推荐。

实时的商业智能的具体流程如图 8-22 所示。

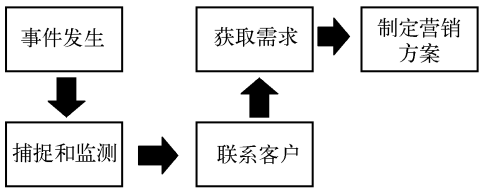


图 8-22　实时的商业智能的具体流程

通过实时获知客户的变化，捕捉客户的需求和购买产品的概率，最后形成完整的营销方案。例如，通过数据仓库中的客户行为信息的自动检测，我们主动联系客户，获取真实的客户需求，最后制定完整的营销方案，形成一个事件式营销流程。

电信行业传统的营销方式和事件式营销方式的区别如图 8-23 所示。

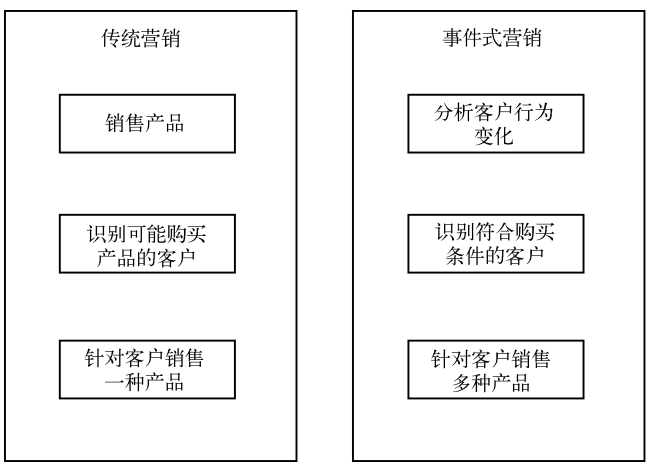


图 8-23　传统的营销方式和事件式营销方式的区别

1）传统的营销有以下几种方式：销售产品、识别可能购买产品的客户和针对客户销售一种产品。

2）事件式营销有以下几种方式：分析客户行为变化、识别符合购买条件的客户和针对客户销售多种产品。

小结

- 商业智能的建设是一个战略性的工程，它直接影响到企业未来的发展方向，对于商业智能的架构应该遵循以下几项原则：前瞻性、实用性、安全性和可信性。
- 商业智能系统从企业的日常数据中开发基于事实的信息，辅助企业做出更好的商业决

策，提高企业运营效率和决策分析的能力。可以帮助企业完成风险分析、欺诈监测、财务分析等。

- 商业智能的典型应用包括：经营分析、绩效管理、战略决策支持、产品管理和创新、客户关系管理和风险管理等。
- 商业智能产品应该建立在稳定的平台上，它可以提供数据关联分析的功能、数据监控的功能、数据展示功能和数据输出功能。
- 传统商业智能具有以下几个特点：查询、报表、多维分析和统计分析、数据挖掘等。
- 传统型的商业智能主要是基于历史数据做出决策和分析。它面向企业的决策者和分析者，主要以查询为主。未来商业智能主要是基于实时的数据做出分析和决策，它可以面向一线的客户经理和决策者，通过实时捕获的数据，获取最新的信息和知识。它可以提高商业智能对业务的即时指导作用，同时快速地响应事件，提升企业的竞争力。
- 商业智能的传统架构类似于传统的物流过程，即各地运来的货物首先存放在暂存库，主要目的是对各类货物进行清洗、筛选、检查、贴标签等工作，然后统一发往货仓，最后在各个超市中进行集中销售。
- 随着商业智能使用越来越广泛，使用者已经不再局限于业务分析人员或者高层领导，可以通过外部网络延伸到企业的客户、合作伙伴等。为了满足这些用户的需求，商业智能架构需要满足可扩展性和可靠性，同时保证快速的响应能力。
- 电信行业实时商业智能架构如下所示：

（1）监控

监控包括用户欺诈监控、服务监控、高额话费预警等内容。

（2）数据查询

数据查询包括话费财务报表、统一客户视图查询、产品套餐销售报表等内容。

（3）多维分析

多维分析包括客户分析、竞争分析、产品与套餐分析、增值业务营销分析、人力分析、财务分析等。

- 这种实时的商业智能架构的目的是在合适的时机，通过合适的渠道，向客户推送合适的产品和服务。这也要求我们从以“产品为中心”向以“客户为中心”转变。
- 传统的营销方式和事件式营销方式的区别：

1）传统的营销有以下几种方式：销售产品、识别可能购买产品的客户和针对客户销售一种产品。

2）事件式营销有以下几种方式：分析客户行为变化、识别符合购买条件的客户和针对客户销售多种产品。

第 9 章　商业智能—数据仓库架构和案例

本章目标

通过前几章的学习，我们了解了商业智能的定义、商业智能的功能、商业智能的发展趋势、商业智能的实施方法和步骤、关于商业智能的核心技术、数据仓库理论、数据仓库的特点、数据集市理论、ODS 理论等知识。同时也掌握了商业智能架构原则和相关典型应用，商业智能具有的功能，商业智能未来的发展趋势和方向，商业智能的传统架构，未来商业智能的架构等内容。

学习本章后，读者将掌握：

- 数据仓库的定义
- 数据仓库产生的背景和原因
- 数据仓库的特征
- 数据仓库和商业智能之间的关系
- 数据仓库的优势
- 数据仓库面临的挑战
- 数据仓库的技术特性
- 数据仓库建设方法
- 数据仓库设计原则
- 数据仓库架构规划
- 数据仓库数据模型
- 数据仓库建设路线图
- 数据仓库系统的灾难备份规划
- 商业银行数据仓库面临概况和瓶颈
- 商业银行数据仓库建设及改进建议
- 商业银行数据仓库建设案例分析
- 商业银行数据仓库建设启示
- 电力行业数据仓库建设难点
- 电力行业数据仓库体系架构
- 电力行业数据仓库能力蓝图
- 数据仓库对电力业务发展的促进作用
- 数据仓库建设策略比较
- 电力行业数据仓库的数据架构设计

9.1　数据仓库概述

9.1.1　数据仓库的定义

数据仓库在比尔·恩门所著的《如何构建数据仓库》一书中的定义：“数据仓库是一个

面向主题的（Subject Oriented）、集成的（Integrated）、相对稳定的（Non－Volatile）、反映历史变化（Time Variant）的数据集合，主要用于支持决策分析”。该定义被广泛接受。换句话说，数据仓库是为企业的决策分析提供支持的所有类型的数据的集合。

1. 如何理解数据仓库

数据仓库是一个过程，而不是一个产品。数据仓库的整个过程包括很多产品和实施服务。例如，数据仓库包含一些平台产品、数据处理工具和前端应用工具。对于平台产品来说，包括数据库、服务器和存储设备。数据处理工具主要是 ETL 工具和一些数据管理工具。对于前端应用工具来说，包括 OLAP 工具、数据挖掘工具、报表展现工具和门户等。

2. 企业级数据仓库的数据架构

企业级数据仓库的数据架构如图 9－1 所示。

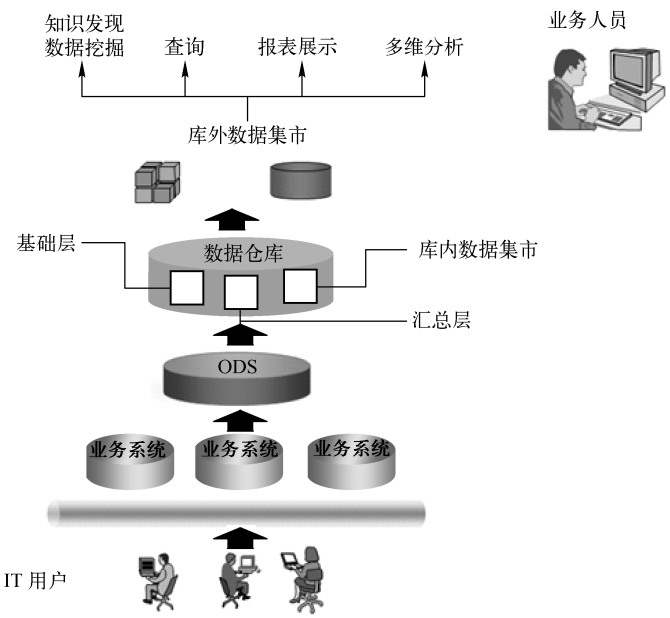


图 9－1　企业级数据仓库的数据框架

3. 数据仓库系统建设应该考虑的问题

1）首先选择数据仓库系统的成功案例作为重要参考。

2）学习行业内的先进经验。

3）具备专业的数据仓库实施队伍和业务领域的专家。

4）考虑数据仓库是否满足海量数据的复杂、并发查询。

5）数据仓库应该满足可扩展的能力。

6）数据仓库应该考虑高可靠性，并且满足高质量的要求。

4. 商业银行数据仓库的应用及需要考虑的主要因素

商业银行数据仓库有很多具体应用，如财务管理、绩效管理、风险管理、资产负债管理和客户管理，如图 9－2 所示。

数据仓库规划时需要考虑的主要因素包括业务需求、技术、投资成本、系统的适用对象等。

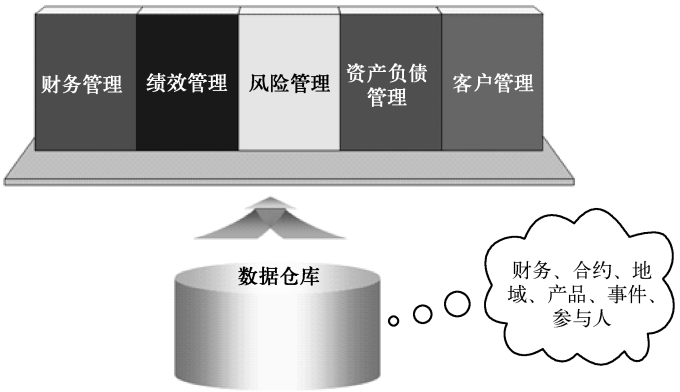


图 9-2　商业银行数据仓库具体应用

● 业务需求

业务需求是指随着市场的开发，客户的需求有哪些转变，业务的发展重点是什么。要理解建设数据仓库不是目的，而是一种手段。

● 技术

技术方面要考虑业务系统的历史数据量和用户数是多少。

● 投资成本

投资成本包括购买数据仓库产品的成本、使用成本、维护成本和管理运行成本。

● 系统的适用对象

系统的适用对象包括管理决策层或者业务部门。

5. 数据仓库的建设目标

数据仓库建设目标是将数据转化成信息、知识，最后辅助企业高层进行决策分析。其中数据是原始业务数据的记录。信息表示整合的数据提供特定的信息。信息间的逻辑关系成为知识。决策是基于对知识的掌握采取相应的行动。

6. 数据仓库项目失败的标志

数据仓库项目失败的标志有以下几种，如图 9-3 所示。

1）数据仓库项目周期延长，费用严重超支。

2）日常工作不依赖于数据仓库。

3）业务人员对数据仓库中的数据质量不信任。

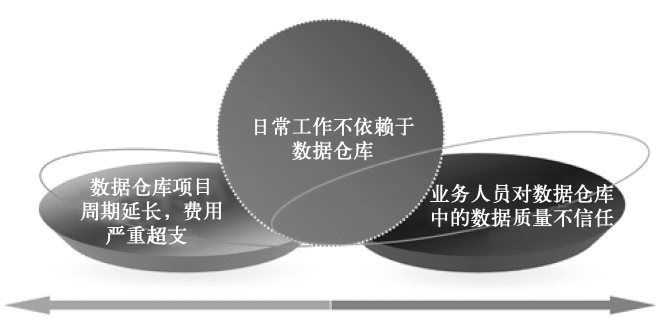


图 9-3　数据仓库项目失败的标志

7. 数据仓库普遍存在的问题

数据仓库普遍存在的问题包括数据仓库平台问题、数据仓库质量问题和数据仓库应用问题，如图 9-4 所示。

图 9-4　数据仓库普遍存在的问题

8. 数据仓库项目需要考虑的因素

数据仓库项目需要考虑很多因素，例如：系统应该实现的目标、项目实施的条件、系统现状和技术平台应该拥有的能力。

对于应用规划的目标来说，数据仓库应该满足企业管理层的决策分析需求，提高客户的满意度。项目实施的条件包括：数据仓库项目实施的成功经验，具有数据仓库实施的一般方法论，同时具备团队建设和管理的能力。

对于系统现状的调研来说，应该调研数据源的质量问题、业务系统的运行状况和各个部门对于数据仓库系统的理解程度。对于数据仓库技术平台的要求包括：具有海量数据处理能力，数据分区的能力，同时具备一定的技术先进性。

9. 对数据仓库有效的使用方式

对于数据仓库的建设过程来说，首先应该实现对业务的分析，帮助高层领导加深对业务运营状况的了解，提高企业的市场竞争能力，然后将分析结果反馈到业务系统中，实现分析应用和业务应用的交互闭环过程，加强对业务运营的指导，为企业带来可持续的价值，如图 9-5 所示。

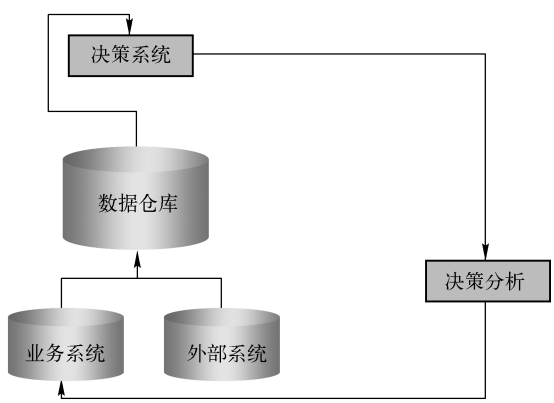


图 9-5　对数据仓库有效的使用方式

9.1.2　数据仓库产生的背景和原因

1. 数据仓库产生的背景

随着信息量的不断增大，企业需要把这些数据当作一种资产，通过多个角度去分析这些海量数据，并从中获取有用的信息和知识。因为事务处理操作型的数据库很难满足这种需求，所以数据仓库技术应运而生。

数据仓库是面向主题的、集成的、稳定的，并且反映历史变化的，数据仓库在保证数据存储的基础上，挖掘信息，使数据变得更有价值。

2. 数据仓库产生的原因

数据仓库的出现和发展是计算机应用到一定阶段的产物，很多企业经过多年的数据积累，保存了大量的原始数据和各种业务数据，这些数据真实地反映了企业的经济情况。但是因为缺乏对数据的有效管理，所以无法体现这些数据对企业的价值。

在 20 世纪 70 年代，出现了关系型的数据库技术，为这一类问题提供了解决方案。

在 20 世纪 80 年代中期，很多用户已经不能满足数据库技术处理事务数据的需求，而是更希望满足决策分析的需要。

随着决策分析的需求慢慢深入人心，在 20 世纪 80 年代末和 20 世纪 90 年代初，终于出现了数据仓库的概念，它为决策支持打下了基础。数据仓库经历了一段时间的发展，加之经过多年的市场和运营积累，企业也已经坐拥了大量的业务数据，这些数据为数据仓库技术的后续发展打下了重要的基础。

20 世纪 90 年代初期，比尔·恩门在《如何构建数据仓库》中提出了“数据仓库”的概念，几年后，数据仓库的研究和应用得到了广泛关注。

9.1.3 数据仓库的特征

数据仓库有以下几个特征：面向主题的、集成的、相对稳定的和反映历史变化，如图 9-6 所示。

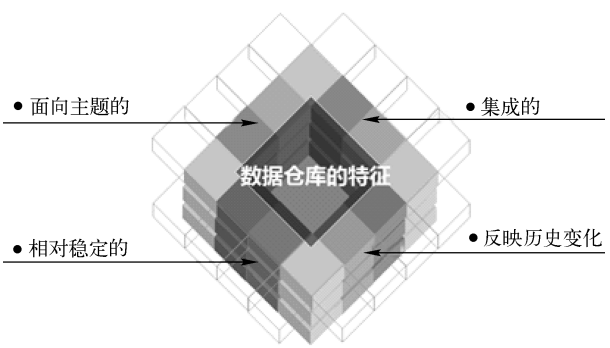


图 9-6　数据仓库的特征

1. 数据仓库是面向主题的

数据仓库是面向主题进行组织的。什么是主题？主题是对业务数据的一种抽象，是从较高层次上对信息系统中的数据进行归纳和整理。面向主题的数据组织方式，就是在较高层次上对分析对象的描述，所谓面向主题的特性是根据业务的不同而进行的内容划分。

2. 数据仓库是集成的

数据仓库中的数据是按照主题存储的，与业务系统中的数据可能会存在较大差别，数据仓库中的数据来源于不同的业务系统，因此，在进入到数据仓库之前，需要经历一个整合、清洗的过程，保证数据的一致性，同时进行数据的集成、计算和汇总。

集成的特性表现在：数据是独立分散的，如核心业务系统、电子渠道系统、信贷系统、票据系统，每个系统只保留单独的数据，如果进行公共的汇总，那么必须纳入到一个统一的

平台进行分析、挖掘。这是数据仓库产生的根本动因，数据仓库也可以做一个公共标准，例如有的系统用 0 和 1，分别代表男和女，有的用 m 和 f 代表，需要有一个统一的标准。

3. 数据仓库是相对稳定的

数据仓库通常保存数据不同历史时期的各种状态，并不对数据进行任何更新操作，一般来说，数据仓库的数据主要是做查询，以供企业决策分析之用。数据仓库中的数据反映的是很长时间内的历史数据，可以看作不同时点的数据库快照的集合。并且在这些快照的基础上进行统计分析。当操作型数据库经过联机处理后，将数据集成并且输入到数据仓库中。而数据仓库将这些历史数据保存起来，如果超出存储期限，这些数据可能会进行归档处理，或者进行删除操作。

因为数据仓库只进行数据查询的操作，并且查询量相对很大，对数据查询的效率提出了更好的要求。例如，可以利用索引、分区等技术对数据仓库进行优化。数据仓库的数据一般不进行删除，但是超过 10 年的数据都放入到归档库中。有些银行单独建设 ODS，不仅给数据仓库供数，而且也为其他应用供数。有些银行是将 ODS 放在数据仓库中建设。

数据可以分成两类，即交易类的和状态类的。交易类的，每天都在增加，如还款记录。另一部分，是状态类的，如合同余额、借据余额。可以基于拉链或者快照的方式放入到数据仓库中。数据仓库尽量不做频繁修改。

4. 数据仓库是反映历史变化的

数据仓库的历史特性是指数据保留时间戳字段，记录每个数据在不同时间点内的各种状态。数据仓库反映历史变化的特性表现在以下几个方面：

1）数据仓库不断地捕捉业务系统中已经变化的数据，然后将这些数据追加到数据仓库中，将不断生成的业务快照经过统一集成后进入到数据仓库中，对于捕捉到的新的变化数据只进行新增操作，而不进行更新操作。

2）一般来说，数据仓库的数据会有存储期限，一旦超出了期限，过期数据就会被归档，或者直接删除。

9.1.4 数据仓库和商业智能之间的关系

从图 9-7 中可以看出，数据仓库是实现商业智能的基础平台，没有数据仓库的搭建，真正的商业智能是无法实现的。

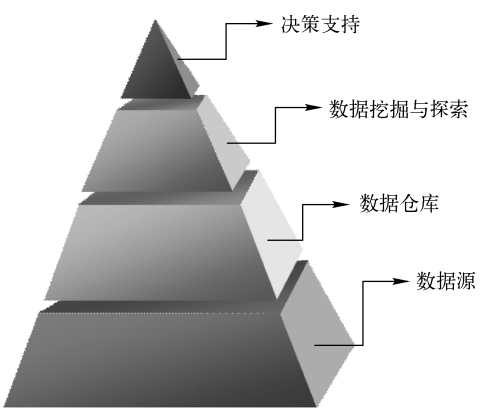


图 9-7　数据仓库和商业智能之间的关系

9.1.5 数据仓库的优势及面临的挑战

1. 数据仓库的优势

数据仓库系统相比其他系统有哪些优势呢？有以下几种：

1）数据仓库系统可以获取生产系统综合的信息，作为科学决策分析的重要依据。

2）数据仓库可以从宏观的角度理解信息，也可以从微观的角度探查信息。

3）通过数据仓库系统，可以建立企业内部各个部门之间的联系。

2. 数据仓库面临的挑战

全球经济的起伏，行业竞争的日益激烈，数据信息的迅速增长，都要求今天的企业具备访问、整合各种数据的能力，并通过数据分析帮助企业管理层做出更快、更好的商业决策。

一方面很多企业拥有多个系统，这可能导致各个系统之间数据互相冲突，从而使管理人员无法及时、有效地获得准确的信息。此外，企业大多数的分析解决方案都是与数据仓库分离的，增加了系统的维护成本和运营负载。另一方面，信息的快速增长使数据仓库规模扩展到一个新的层次，同时还产生了更加复杂的数据关系，对海量数据的查询、挖掘与分析变得更加复杂，从而导致系统性能降低，这对决策分析的及时性和灵活性产生重大影响。

传统数据仓库所带来的挑战，使企业管理层无法获得及时、准确、有效的业务信息，这会对企业的运营和竞争力带来影响，原因如下所示：

1）缺乏有效的目标市场定位，难以推出有针对性的产品。

2）不能够根据个性化的服务需求，制定出对应的营销策略。

3）不能及时了解客户的真实需求和特征，无法提高客户的忠诚度。

因此，企业需要一种全面、多功能的数据仓库平台，它不仅提供唯一事实的版本，更需要实时洞察的功能。

9.1.6 数据仓库的技术特性

数据仓库的技术特性主要包括海量数据处理能力、高可用性、线性的扩展能力和数据压缩能力，如图 9-8 所示。

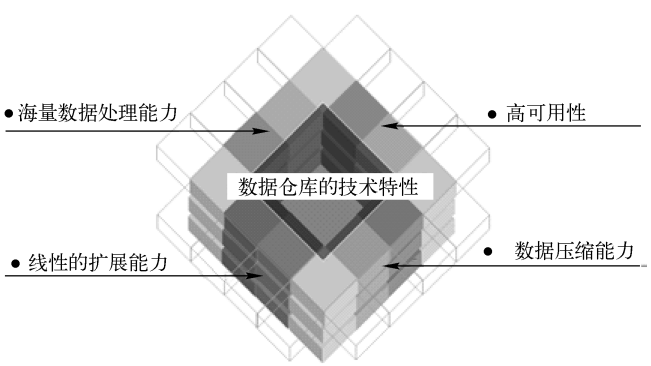


图 9-8 数据仓库的技术特性

（1）海量数据处理能力

数据仓库汇集了系统的全部数据，数据量不断增长，这就需要数据仓库平台能够处理高

并发和大数据的能力。

（2）高可用性

数据仓库平台需要提供高可用方案，满足系统的高可用性需求。

（3）线性的扩展能力

随着用户需求的多样化，数据仓库平台不仅能够满足现有的处理需求，而且可以提供良好的扩展能力，以满足不断增长的数据量和复杂的查询需求。

（4）数据压缩能力

数据仓库平台应该提供良好的数据压缩能力，降低成本，满足系统恢复的时间要求。

9.2 数据仓库设计

9.2.1 数据仓库建设方法

1. 数据仓库建设的方法论

“制定数据标准，建立数据管控机制，以数据、应用驱动为主”是数据仓库基本的建设方法论。如图 9-9 所示，对于数据仓库的建设应该首先建立分析类数据标准和基础类数据标准，同时成立数据管控机制，最后以数据、应用驱动为主，建立数据仓库系统。其中数据仓库可以分成基础数据层、汇总数据层和库内集市层。基础数据层的数据是以主题域的方式进行划分，汇总数据层在基础数据层的基础上按照时间或者机构等维度进行汇总。库内集市层一般是在汇总数据层或者基础数据层的基础上建立起来的。应用可以建立在库内集市层或者库外集市层中。

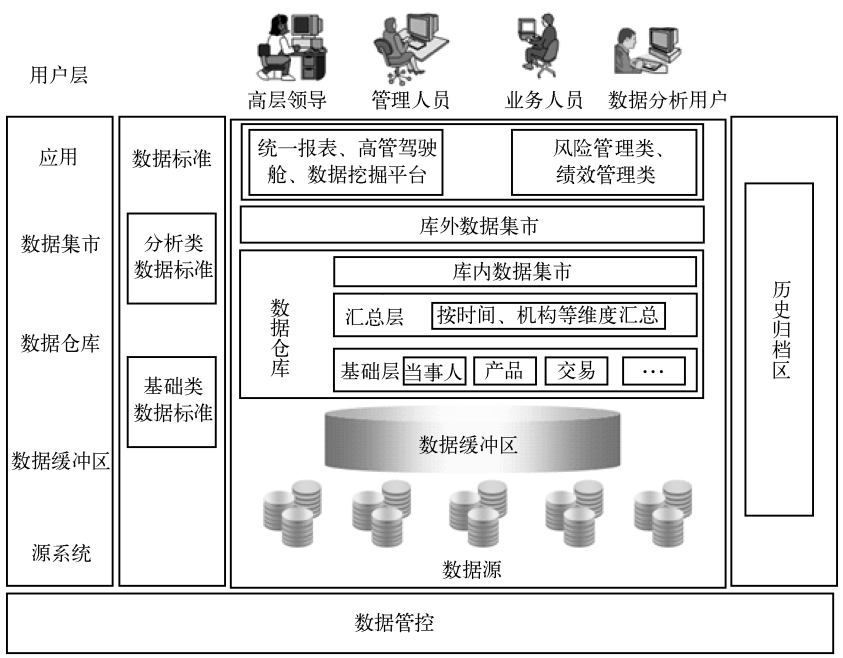


图 9-9　数据仓库建设的方法论

（1）制定数据标准

制定数据标准时，可参考金融数据模型，同步设计标准参考模型，直接构成数据仓库逻辑数据模型的“骨架”和核心内容。同时推进数据标准化的工作，保证基础信息项的名称、定义、口径一致。在此基础上，逐步建立数据仓库，将各个系统的数据分类汇集到数据仓库中，实现数据管理的规范化和标准化。例如，将客户、产品、机构等基础信息建立统一的数据标准，以确保名称、定义、口径和来源的一致性，然后在数据仓库的建设过程中遵循这些统一的标准。

（2）建立数据管控机制

以元数据管理为基础，以管控流程为手段，使数据仓库成为可信、可控的数据源。

（3）以数据、应用驱动为主

提升经营管理、决策分析和监管报送水平。

2. 数据仓库规划的原则

数据仓库满足高效、灵活的多层次的数据应用需求，以更高的效率和质量来支持复杂的分析应用。数据仓库能够整合各类数据源，提高数据架构的灵活性、数据处理高效性和数据加工的自动化水平，使系统设计更具前瞻性和易扩展性，保证系统安全稳定性的提高。明确各个部门管理职责，指定信息管理岗位职责，完善数据管理技术岗位，制定管理岗位绩效考核等指标。

3. 数据仓库的实现方式

一般来说，数据仓库的实现方式可以分成两类：数据驱动的实现方式和业务驱动的实现方式，如图 9-10 所示。

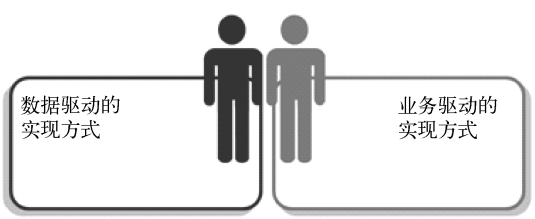


图 9-10　数据仓库的实现方式

（1）数据驱动的实现方式

当业务需求不明确的时候，按照这种方式，首先进行全企业的数据建模，并且按照 E/R 模型建立数据仓库，然后根据业务部门的需求建立相应的数据集市，数据仓库作为数据集市的唯一来源。从整体的角度进行总体规划，例如 6 ~ 8 个月搭建基础数据平台，形成初步成果，采用循环式的开发方式，向业务部门提供切实的成果。可以边开发、边投产、边推广、边收益。

（2）业务驱动的实现方式

当业务需求非常明确的时候，按照业务需求迭代地建设数据仓库，在建设数据仓库的过程中，有什么样的业务需求就抓取什么数据。

两种方式的不同点如图 9-11 所示。

数据驱动的实现方式需要将数据仓库的模型在前期做扎实，使得大部分的人力集中在数据仓库的基础上开发应用，数据仓库的模型只需要微调就可以满足应用。特点是前期建立数

图 9-11　两种方式的不同点

据仓库模型需要的时间周期长，见效慢，但是一旦数据仓库模型建立扎实后，后期的基于数据仓库的应用开发时间就会大大缩短，数据仓库模型只需进行微调就可以满足应用需求。

业务驱动的实现方式是在业务需求很明确的情况下，按照业务需求迭代地建立数据仓库模型，即有什么样的业务需求就为数据仓库抓取什么样的数据。特点是前期建立数据仓库需要的时间周期较短，对项目来说，具有“短、平、快”的特点。但是对于后期如果需要增加新的应用，那么数据仓库模型需要有较大的调整。简单来说，业务驱动就是有什么业务需求就抓取什么数据，而对于数据驱动来说，是当很多业务需求讲不清楚的时候，先把所有有用的数据全部都放进数据仓库中。

总结来说，数据仓库的整体建设思路主要是：

首先是整体规划和分步实施，也就是先设立分阶段的目标，再逐步实施。

然后是完全将业务需求作为数据仓库系统建设的驱动，最终让数据仓库的分析系统和业务系统能够互相交互和影响，形成一个闭环的结构。

最后还可以采用齐头并进的方式建设数据仓库。例如，以数据为驱动的系统分析和以业务需求为驱动的系统分析同时进行。

9.2.2　数据仓库设计原则

数据仓库架构设计遵循以下原则：可重用性、高性能、可扩展性、可管理性和高可用性，如图 9-12 所示。

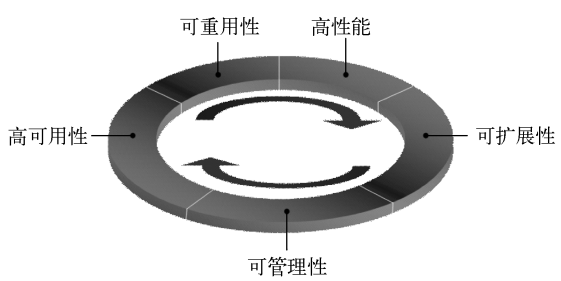


图 9-12　数据仓库架构设计遵循的原则

（1）可重用性

数据仓库的可重用性是指数据仓库系统的组件可以被多次利用。例如，使用 ETL 工具或者数据服务组件，提高数据和组件的可重用性，从而减少重复的开发。

（2）高性能

数据仓库应该满足高性能的需求。对数据仓库来说，可以采用诸如负载均衡、多机并行的技术提高数据仓库系统的响应处理能力，这样可以多方面、多层次地提升数据仓库的性能。

（3）可扩展性

数据仓库系统应该尽量支持以第三范式为主的逻辑数据模型的设计方法。同时需要考虑架构灵活的原则，将业务需求封装到数据仓库模型中，减少数据不必要的重复。保证在业务需求发生变化的时候，改动量最小化。这样，可以满足未来数据仓库系统的可扩展性。

（4）可管理性

数据仓库的可管理性是指当局部发生变化的时候，应该从全局的角度估计出这个变更可能产生的影响。

（5）高可用性

数据仓库的高可用性是指在规定的服务时间范围外，数据仓库系统可以安排计划内的停机。但是如果在服务时间范围内，出现因为硬件或者其他原因导致的系统服务或者数据不可用时，那么应该保证数据仓库系统尽快恢复，尽量避免因停机带来的损失。

对于简单加工、以查询为主的数据服务，不需要使用数据仓库技术。数据仓库的应用需要建立在海量历史数据和复杂多维的计算上。

9.2.3 数据仓库架构规划

1. 数据仓库的架构和定位

数据仓库可以作为数据架构规划中的重要内容之一。一般来说，在系统中的定位如图 9-13 所示。

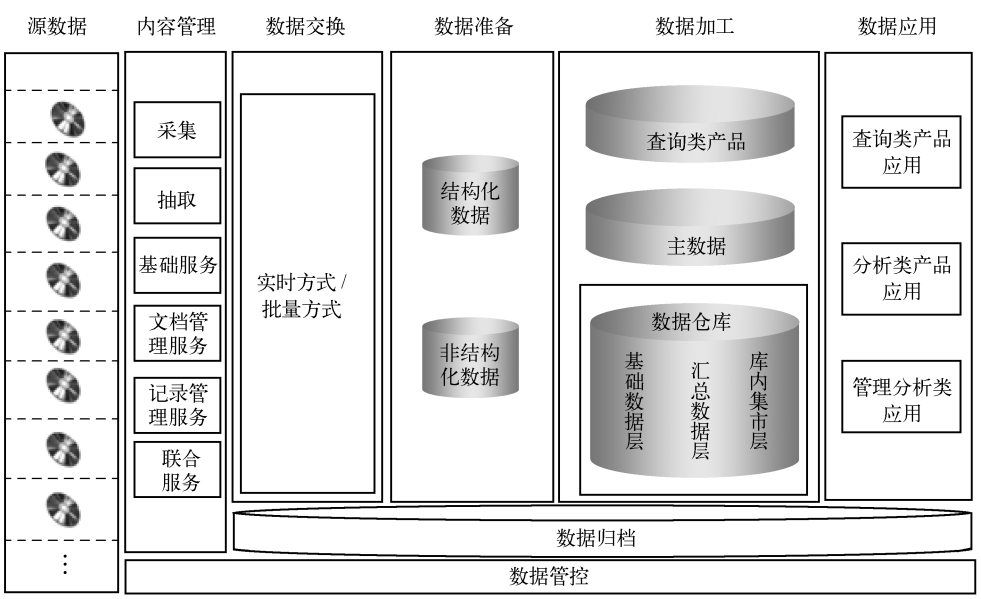


图 9-13 数据仓库的架构和定位

数据仓库包括基础数据层、汇总数据层和库内集市层，如图 9–14 所示。

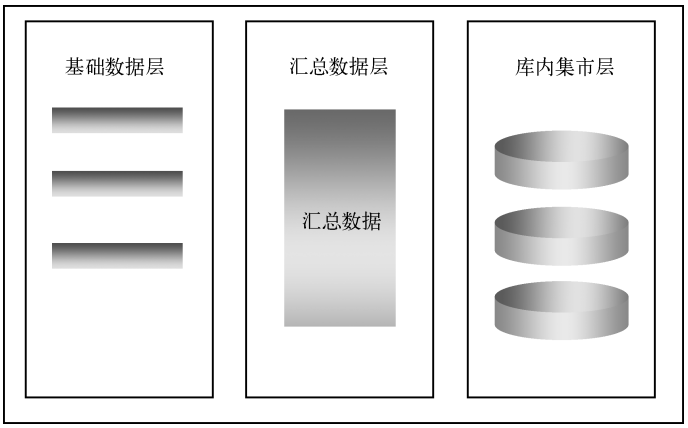


图 9–14　数据仓库包括基础数据层、汇总数据层和库内集市层

数据仓库有以下几个特征：

1）数据仓库整合系统的全局信息，包括基础数据层、汇总数据层和库内集市层。

2）数据仓库中的数据通常包含历史信息，记录了从过去某一时间点到目前各个阶段的信息，通过这些信息，可以对企业的发展历程和未来趋势做出分析和预测。

3）数据仓库的数据来源可以是结构化的基础数据、非结构化数据结构化的信息，也可以是产品数据或者主数据。

4）数据仓库中的库内集市层是根据应用需求形成的数据集合，它支撑了各种专业化的应用。

下面分别对基础数据层、汇总数据层和库内集市层进行描述：

（1）基础数据层

对于数据仓库中的基础数据层来说，它存储了数据仓库最细节层次的数据。它的数据源来自于数据准备区中的数据，是最贴近数据源的一层。

基础数据层的特性：

1）基础数据层一般是按照数据仓库的第三范式进行数据组织的。

2）基础数据层作为汇总数据层的数据源。

3）基础数据层一般不做删除操作。

数据仓库基础数据层同数据准备区中的基础层相似，但是两者在组织形式、用途、内容、访问频率等方面存在差异，见表 9–1。

表 9–1　数据仓库基础数据层与数据准备区中的基础层的差异

差　异　点	数据仓库基础数据层	数据准备区中的基础层
组织形式不同	按第三范式存储，强调完整性、一致性，时效性相对较低	存储贴源，按第一范式或第二范式存储，时效性高
用途不同	支持仓库汇总加工	主要支持基础产品加工，并对数据仓库供数
数据内容不同	除包括基础数据存储外，还包括主数据等内容	包括所有采集数据
访问频率不同	数据仓库数据访问频率较低	采取准实时批量方式加载及供数

（2）汇总数据层

对于数据仓库汇总数据层来说，它是对基础数据层的数据进行轻度汇总，同时为分析型的应用提供数据服务。

汇总数据层的特性：

1）随着应用需求的增加，汇总数据层的建设需要不断扩展。

2）汇总数据层是对明细数据的必要整合，目的是对一些共性需求进行加工整合，提高数据的利用率。

3）汇总数据层的来源应该是数据仓库中的基础层，汇总的问题可以直接反映一些业务需求。

通过创建中间汇总表，预关联和汇总常用的数据，使其多个数据集市可以共享该数据，以提高数据仓库的性能，同时也降低了 ETL 工作的复杂性。

（3）库内集市层

数据仓库规划库内集市，首先数据在基础层整合后，做一些汇总设计。可以把基础数据层的数据和汇总数据抽取出来做成接口数据，提供到库外建设。对于一些应用较为复杂的、独立的情况，把数据下放到库外应用。而相对应用简单的，直接在仓库内实施。

数据仓库建设一期先把所有源数据纳入数据仓库的基础数据层，然后加工到汇总数据层，随着业务需求的增加，慢慢扩展数据仓库的内容。

例如，个人基本信息、地址等信息直接在基础数据层抽取；余额类的、每天都可能发生变化的数据先在汇总数据层加工，再抽取到集市。

汇总数据层是公共加工层。汇总加工一次，可以支持多个应用。例如，按客户、产品进行日均、月均加工汇总。可以把基础数据层的明细数据汇总到汇总数据层，将汇总数据层作为公共数据提供给应用。

对于数据仓库中的库内集市来说，有以下几个特点，如图 9-15 所示。

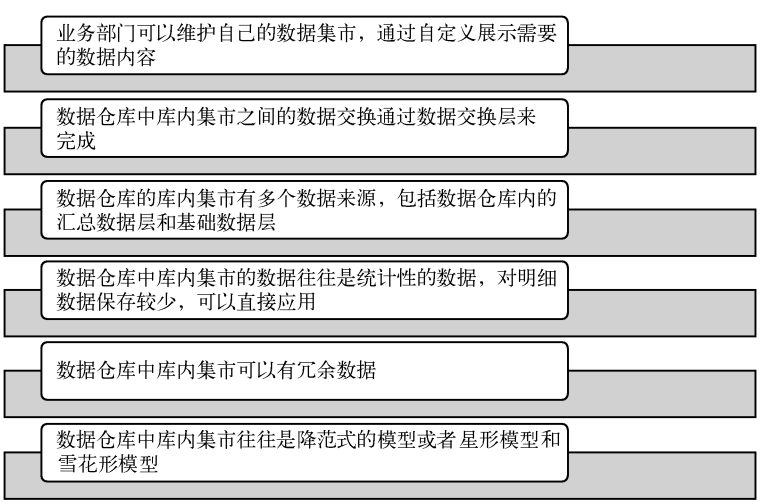


图 9-15　数据仓库中的库内集市的特点

1）业务部门可以维护自己的数据集市，通过自定义展示需要的数据内容。

2）数据仓库中库内集市之间的数据交换通过数据交换层来完成。

3）数据仓库的库内集市有多个数据来源，包括数据仓库内的汇总数据层和基础数据层。

4）数据仓库中库内集市的数据往往是统计性的数据，对明细数据保存较少，可以直接应用。

5）数据仓库中库内集市可以有冗余数据。

6）数据仓库中库内集市往往是降范式的模型或者星形模型和雪花形模型。

下面分析数据仓库建设的一个重要方法，如图 9-16 所示：

1）在数据仓库的基础数据层，建立稳定的数据模型，同时建立数据标准，实现数据的标准化和数据集中。

2）在数据仓库的汇总数据层，建立分析类的数据标准。对常用的、重要的业务指标进行统一加工计算。实现业务的汇总，创建高效的数据共享平台。

3）数据仓库的库内集市层包括基础数据层的视图、汇总数据层的视图和各种加工视图，如图 9-16 所示。

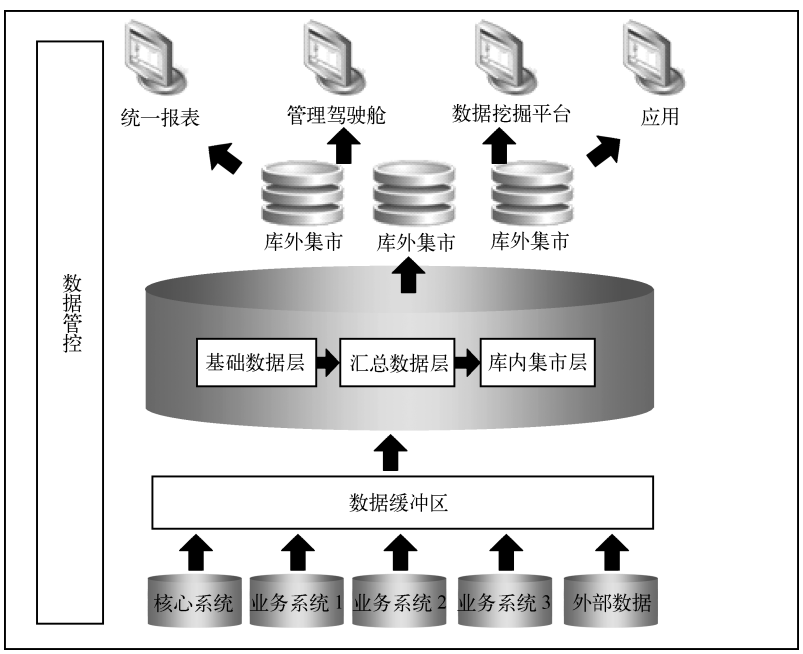


图 9-16　数据仓库建设的另外一个重要方法

数据管控的范围包括元数据管理、数据质量管理和数据维护管理。

数据管控的手段包括数据管理系统、调度与监控系统和数据管理考核系统等。其中，数据管理考核系统帮助系统有效提升数据质量，一般采用系统检查和人工核对的方式进行数据管控考核工作，这种方式可以有效地推动数据管控制度的执行，提升数据质量，促进业务人员使用数据管理系统加强数据质量的管理。

建设数据仓库的方法之一就是首先制定数据标准，形成数据仓库逻辑模型的核心骨架，然后以元数据管理为基础，保证数据仓库成为可信和可控的数据源，最后提高管理水平。

总的来说，关于数据仓库的定位有如下几个方面：

1）数据仓库中的来源数据为结构化的，或者是已经结构化的基础数据。

2）数据仓库中的数据都是有用的数据，是经过清洗后的数据。

3）数据仓库加工后的数据可以同步到数据应用层，由应用层对外提供服务。

4）数据仓库中的基础数据层、汇总数据层和库内集市层都有各自的定位和用途。

数据仓库从生产系统采集数据，经过 ETL 过程将数据加载到数据仓库中，然后进行汇总和加工，最后在数据仓库的基础上提供各种应用和分析。

2. 数据仓库关键设计点

基于业务及整体架构规划，我们讲解数据仓库的关键设计点。

（1）数据仓库的构建方法

数据仓库的构建方法主要包括自上而下的实现方式和自下而上的实现方式，如图 9-17 所示。

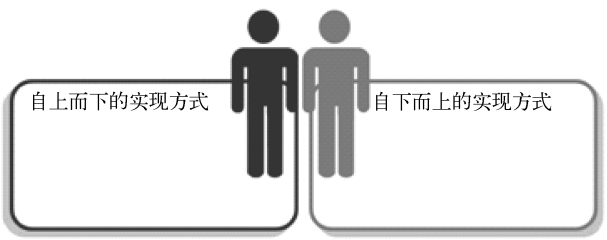


图 9-17　数据仓库的构建方法

- 自上而下的实现方式

这种方式是企业级的数据建模和数据整合，一般按照第三范式模型建立数据仓库，然后根据业务部门的需求，基于已有的数据仓库建立相应的数据集市。数据集市的数据来源是数据仓库。建设的条件是业务需求较少，而数据准备区中的基础数据又比较完整，这样可以采用自上而下的构建方法。将基础数据在数据仓库中进行全面存储，而后续的应用可以随着需求的完善而不断扩展。优点是可以进行宏观的全局规划，有较好的数据一致性和较低冗余。缺点是建设周期长，前期很难见到实际效益。

- 自下而上的实现方式

按照业务需求通过渐进的方式建设数据仓库。首先根据业务需求建立数据集市，然后把一系列维度相同的数据集市纳入到数据仓库中，这种分阶段的建设方式就是自下而上的实现方式。其中每一阶段的数据集市必须兼容到数据仓库中，可以先建设部门级、面向主题的数据集市，然后扩建为数据仓库。它的优点是从业务需求出发，项目周期短。缺点是数据仓库的一致性难以保证，数据的冗余度较高。

如果分析类的业务需求比较多，同时为了快速满足应用的开发，可以采用自下而上的构建方法，先将有业务需求的数据存储到数据仓库中，继而开发应用，然后慢慢地补充数据仓库中的数据。

总之，数据仓库的构建方法是以业务需求为导向的，并且不断完善的闭环流程。

（2）数据仓库 ODS 建设方法

ODS 的概念也是由比尔·恩门在《建立运营数据仓储》一书中提出来的。他认为分析决策需要基于实时的和细节性的运营数据，同时也需要这些数据是集成的和面向主题的，因

此提出了 ODS 的概念。

ODS 的数据来自于各个分散的业务系统，这些数据是面向主题的、集成的、变化的和反映当前情况的数据。一般来说，ODS 和数据仓库作为独立的系统分别进行建设。但是随着硬件水平的提高，有时候 ODS 也被纳入到数据仓库中进行建设。

（3）数据集市设计

数据集市基于业务需求的复杂度，考虑设置库内集市还是库外集市。例如，当业务需求比较单一，复杂度较高的时候，为了性能上的考虑，可以建立库外集市。如果复杂度较低，那么可以在数据仓库内建立集市。

（4）非结构化数据在数据仓库的应用

对于非结构化数据，可以通过 Hadoop 平台建立非结构化数据的标签、摘要、索引、日志等信息，然后提取非结构化数据的元数据信息，如类别、索引、摘要等，实现与结构化数据的整合和关联分析。在统计分析应用中，可能涉及结构化数据和非结构化数据的联合应用，也可能是对非结构化数据的单独应用，如图 9-18 所示。

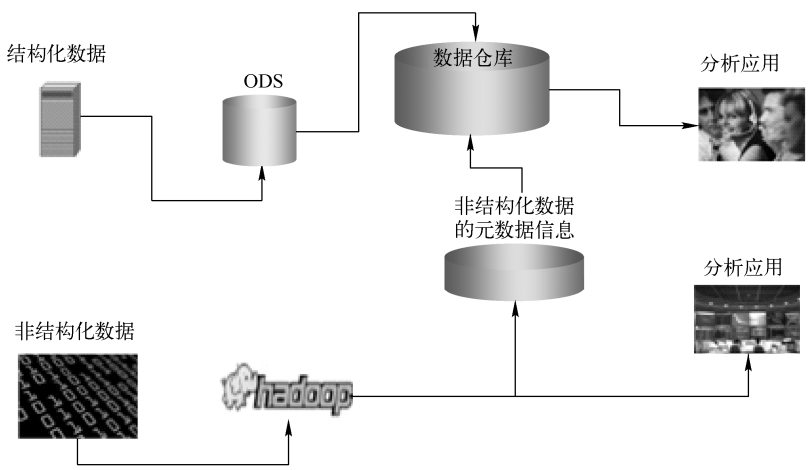


图 9-18　非结构化数据在数据仓库的应用

（5）数据标准在数据仓库中的落地

在建设数据仓库之前，如果可以先完成数据标准的建设，则有利于数据仓库在数据一致性方面的控制。

（6）关于数据仓库的灾难备份设计

由于地震、水灾等因素可能会对数据安全造成危害，因此国家出台了一系列法规、政策，要求各重要机构进行灾难备份体系的建设。后面会详细介绍数据仓库系统的灾难备份建设。

3. 数据仓库质量的控制是数据治理的重要内容

对数据仓库的质量控制是数据治理的重要内容之一。数据仓库系统服务于经营决策，数据应该是全面的、真实的和有意义的。如果数据质量得不到保证，就会使决策分析者做出错误的判断，可能会引起不可挽回的商业损失。因此，提高数据质量是数据仓库系统建设的重要环节。

总体来说，数据仓库对数据质量的要求可以归纳为以下几点：数据的正确性、数据的完

整性、数据的一致性、数据的有效性、数据的时效性、数据的可获取性和数据的冗余性，如图 9–19 所示。

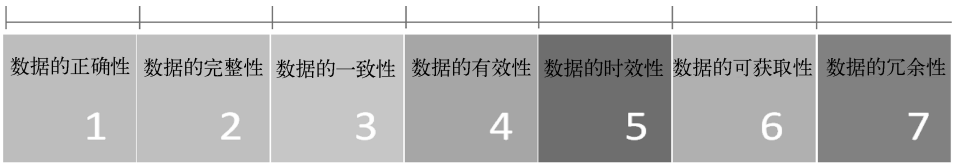


图 9–19　数据仓库对数据质量的要求

- 数据的正确性

数据在数据仓库中是否会正确体现。

- 数据的完整性

数据仓库中的数据是否是完整的。

- 数据的一致性

数据仓库中的数据是否是一致的。

- 数据的有效性

数据是否在企业定义的可接受范围之内。

- 数据的时效性

数据在给定的时间内是否有效。

- 数据的可获取性

数据是否易于获取、理解和使用。

- 数据的冗余性

数据仓库中是否存在不必要的数据冗余。

技术类数据质量指标见表 9–2。

表 9–2　技术类数据质量指标

指 标 类 型	说　　明
完整性	实体的每个属性都有明确的值，不存在“空”或“未知”的属性
相关性	对于数据库中的某些实体，它们的存在可能要依赖于其他的实体
唯一性	一个表中的一组属性的值是唯一的
有效性	实体属性的值要在用户定义的有效范围之内
及时性	是否满足业务应用对数据的时间要求
非重复记录	是否存在多个记录表现同一个实体的现象

业务类数据质量指标见表 9–3。

表 9–3　业务类数据质量指标

指 标 类 型	说　　明
真实性	数据库中实体必须与现实世界中的对象是一致的
精确性	指数据精度是否符合业务需要
一致性	数据是否和其他系统的业务含义是一致的
可理解性	数据本身的含义是否简单、明确
可获得性	数据是否可获得，并满足业务使用要求

数据仓库的数据质量面临的挑战见表 9-4。

表 9-4　数据仓库的数据质量面临的挑战

质 量 分 类	关 键 问 题
一致性	同一条记录被多个应用程序访问时，信息含义是否保持一致性
时效性	数据从被知道到使用，需要多长时间，这种延迟是否可以被用户接受
可访问性	数据是否可以被需要的人访问
可理解性	数据是否容易理解
完整	数据是否有足够的完整信息，并且能够用于决策分析
正确反映现实	数据是否在任何时期内都符合实际情况
汇总数据的准确性	数据汇总是否准确和可信
无冗余数据	是否有多条记录表示同一个实体

数据质量存在问题的根本原因：

我们可以把数据质量存在问题的原因归为以下几类，如图 9-20 所示。

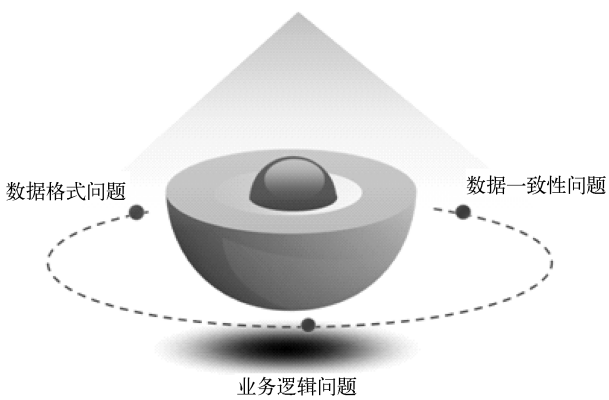


图 9-20　数据质量存在问题的原因

1）数据格式问题。例如，数据的缺失、超出了数据范围、无效数据格式等。

2）数据一致性问题。出于性能考虑，可能会去掉一些外键或者检查约束，这样可能会出现数据一致性的问题。

3）业务逻辑问题。通常是由于数据库设计出现问题所致。

原因分析和解决思路：

1）在构造数据仓库的时候，如果数据质量得不到保证，那么在后续的构建过程中，数据质量所引发的问题会逐渐被放大。

2）数据质量问题会贯穿于项目的整个生命周期，必须面对并且给出解决办法，尽量把影响降低到最小。通常情况下，当遇到错误数据时，通过记录，同时打上错误的标记，先保证这些数据顺利通过，然后根据这些错误标志，通过报表反映出来。这样可以确保数据的完整性，并且真实反映数据源的质量，保证数据仓库的顺利实施和任务的正常调度。

3）技术检测数据仓库质量的方法有多种。例如，第一种方法，对于记录级的，可以先分离出主表，再验证目标表和源表中主表的记录数是否一致。第二种方法，对于字段级别的，如有两个团队，一个是开发组，另一个是数据质量组，当开发团队抽取出数据后，再由数据质量组通过业务规则编写验证脚本，验证两边的结果是否保持一致。第三种方法，寻找

不同目标表中相同口径的值，验证数据是否一致。

4）在大多数情况下，解决数据仓库质量问题最根本的方法就是从源头解决质量的问题，但是这种方式需要投入大量人力成本和时间成本。

5）可以通过手工方式对数据仓库质量问题进行处理。

数据质量的检查应该尽量在靠前的位置进行，这样确保错误的数据在前面就被消除掉，因为每一点的错误都会导致在后续的处理过程中被无限放大。数据的完整性和正确性问题都可能因为 ETL 的错误导致，可以通过源和目标的汇总对比，找出差异，从而确定数据的完整性和正确性是否有问题。

数据仓库质量问题解决办法可参考案例如图 9-21 所示。

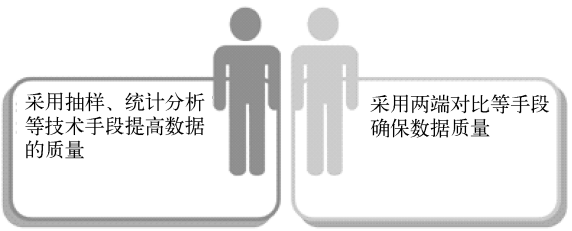


图 9-21　数据仓库质量问题解决办法参考案例

（1）采用抽样、统计分析等技术手段提高数据的质量

对于数据仓库来说，它主要存储的是大量的历史数据，无形当中加大了数据质量检查的难度，如何保证进入数据仓库中的数据是高质量呢？可以采用抽样和统计分析等技术手段提高数据的质量，并且提高数据仓库的高效性。

它的具体做法是通过抽数定理抽取少部分的样本数据，然后进行系统级别的数据校验。如果出现系统级别的错误，则马上返回。如果没有出现系统级别的错误，则对数据抽样取得的数据进行质量打分。如果数据质量的分数较高，那么它的入库校验相对简单，即校验规则相对简单，入库的效率就很高。如果数据质量的分数较低，那么它的入库校验相对复杂，也就是校验规则相对复杂。因此，可以将入库的校验可配置化。

（2）采用两端对比等手段确保数据质量

对于数据仓库的数据质量来说，它可以进行入库时的格式校验和逻辑校验，当入库后，再通过两端对比等手段确保数据质量。所谓两端对比是指在源系统中抽取出一部分数据，再和数据仓库中的一部分数据进行核对。

综上所述，第一种方式是采用抽样、统计分析的方法发现数据的系统错误，以及提高数据校验的效率，将数据质量校验都集中在入库前完成，入库后的数据质量问题主要通过异议处理等手段来实现。

这种采用抽样、统计分析的校验数据方法，对于质量好的数据采取相对宽松的校验规则，对于质量差的数据采取相对严格的校验规则，这样会大大提高数据的加载效率。然后对于通过检验的数据，再逐条进行检查，同时对于校验规则的有效性，不断进行调整，尽量保证入库数据的质量。因此，第一种方法是较为先进的方法。

4. 在大数据环境下的数据仓库的建设

大数据是指无法在一定时间内，用传统型的数据库软件对其内容进行抓取、管理和处理

的数据集合。大数据用于在成本可承受的条件下，通过非常快速采集、发现和分析，从大量的、多类别的数据中提取价值。大数据是一系列技术的集合，汇集了如 Hadoop/Mapreduce、一体机、NoSQL、数据分析与挖掘、商业智能、数据仓库等。

通过对大数据的处理和分析，可以发掘出巨大的价值，包括商业价值和社会价值。

关于大数据环境下的数据仓库架构，如图 9-22 所示。

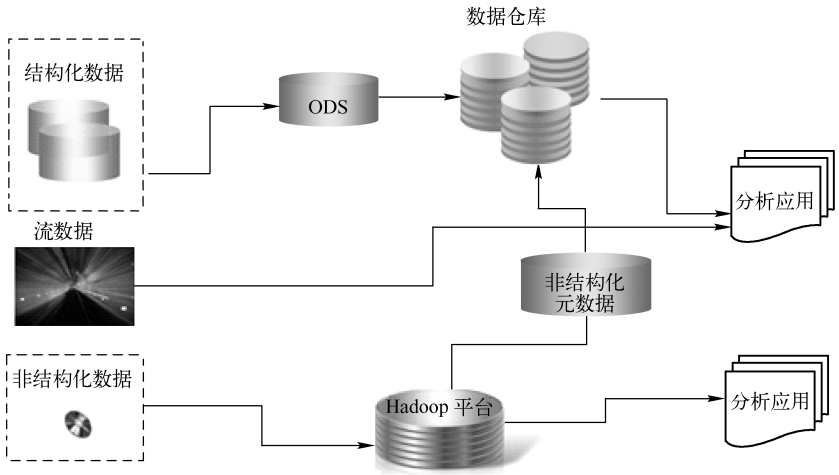


图 9-22　大数据环境下的数据仓库架构

流数据强调的是实时处理与分析，而不是数据存储，因此一般只在内存中进行处理，随着数据的流动、处理和分析，只针对一定时间内的流数据进行处理和分析。

对于数据仓库来说，存储的都是“金子”，全部都是有用的信息。而 Hadoop 平台存储的都是“金矿石”，Hadoop 平台的目的是为了把“金矿石”里的“金子”筛选出来。

所有的非结构化数据都是通过 Hadoop 平台进行分析，例如通过网络收集信息，分析人们对银行的情感分析，包括正面、负面的信息。同时包括针对银行的预警分析等。

非结构化的数据可以经过结构化处理，再与数据仓库中的数据结合起来分析，或者单独对非结构化数据进行分析。

9.2.4　数据仓库数据模型

1. 数据仓库模型设计原则

数据仓库模型的设计原则包括一致性、可扩展性、不倾向性、高效性和可回溯性，如图 9-23 所示。

图 9-23　数据仓库模型的设计原则

数据仓库模型的设计原则的相关内容见表 9–5。

表 9–5　数据仓库模型的设计原则的相关内容

设 计 原 则	相 关 内 容
模型的一致性	数据仓库的数据模型必须在设计过程中保持一个统一的业务定义。统一业务的定义和概念，方便不同系统的设计、开发人员在进行功能设计和数据展现时的沟通和交流
模型的可扩展性	业务需求是随时变化的，因此模型设计需要遵循“以不变应万变，以小变应大变”的设计思想，当业务部门后续有新的需求时，模型不需要做更改，或者只需做轻微的更改即可满足业务需求
基础数据层模型的不倾向性	模型不倾向性的含义是：模型中的数据结构不倾向于源系统，也不倾向于上层应用，不应该和它们发生耦合，即模型底层存储的是基础明细数据，不应倾向于某数据源，也不应该为某个业务部门的应用需求做任何特殊加工
数据加工高效性	数据仓库处理的数据量巨大，而且随着业务量增加，数据的处理效率必然受到影响，因此，在模型设计时，需要能够在给定时间窗口内处理海量数据
数据加工可回溯性	例如，数据仓库上线后，业务人员查看报表时，发现前几天数据有误，需要重新加载数据，此时模型需要支持重新加载之前的数据

设计数据仓库模型的方法原则包括：可维护性、规范性、粒度、历史性和可用性，如图 9–24 所示。

图 9–24　数据仓库模型的方法原则

数据仓库模型的方法原则的相关内容见表 9–6。

表 9–6　数据仓库模型的方法原则的相关内容

方 法 原 则	相 关 内 容
模型的可维护性	数据流向清晰，依赖关系简单，当有需求变更或者出现问题时，将影响降至最低，能够快速维护
模型设计规范性	模型设计时，必须遵循一定的设计规范，如命名规范、业务规则规范等
模型的粒度	为了满足将来不同的应用需求，数据仓库模型能够提供最小粒度的详细数据，以支持各种可能的分析查询
模型的历史性	数据仓库要存储历史记录，比如保留账户、客户信息每次变化的痕迹，账户的转账交易数据等
模型的可用性	数据仓库模型需要很方便地支持业务需求，数据仓库模型设计完成后，基于之上的报表开发、查询开发都很方便、快捷

2. 数据仓库模型设计策略

按照数据仓库模型的设计原则，建议在数据仓库模型设计中采取如图 9–25 所示的设计策略，包括：数据仓库模型设计分层；失效日期填写为默认值，不采用空值；利用时间戳，

保留历史数据；对大表进行分区；将设计流程规范化；采用主流的设计工具；数据和索引分别存储在不同的表空间中；对于特定的缓慢变化维，使用代理键；公用数据处理前置。

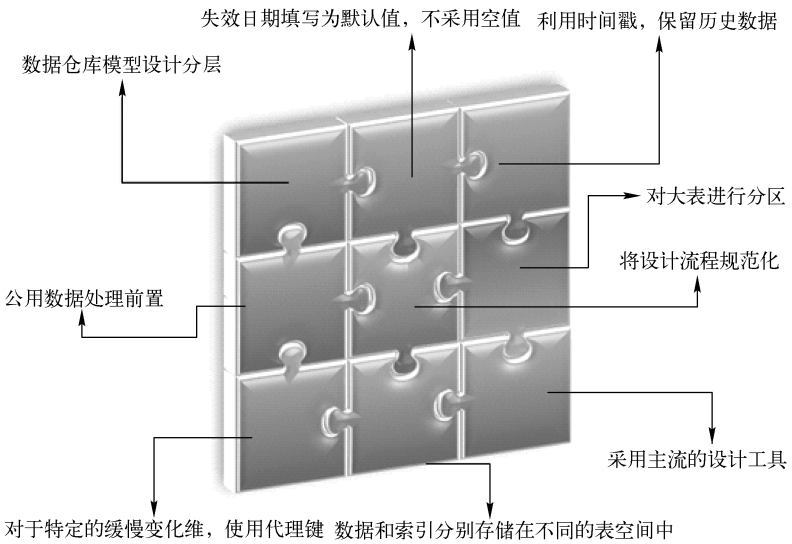


图 9-25　数据仓库模型设计策略

9.2.5　数据仓库建设路线图

数据仓库的建设一般可以分成三个阶段进行。

第一阶段：

完成数据仓库模型的建设。因为数据仓库是面向主题的、集成的、历史的、相对稳定的数据的集合。对于面向主题、集成的特点，数据仓库的数据应该按照仓库的模型进行存储和摆放。对于模型的建设，在整个数据仓库建设中占了相当大的比重。建设的内容主要包括完成数据仓库企业级的概念模型和应用级的逻辑模型的建设，最后完成基于数据仓库物理模型的实现。

第二阶段：

按照数据仓库的模型，将基础数据、产品数据或者日志数据在数据仓库中进行存放，并且完成历史数据的迁移。具体的建设内容可以包括：建设数据仓库的基础数据层，开发校验规则，对入库的数据进行检查，最后完成历史数据的迁移。因为数据仓库需要对历史数据进行统计和分析，所以包含了历史数据迁移的工作。它的数据流转如图 9-26 所示，源数据通过数据交换层将数据放入到数据仓库中。

第三阶段：

完成数据仓库汇总数据层的设计，包括数据集市的设计，最后将数据仓库数据加工后导入到数据集市中。数据仓库建设的内容包括完成数据仓库汇总数据层的设计和数据仓库库内集市层的加工。数据仓库的数据流转如图 9-27 所示，汇总数据层的数据来自于基础数据层的数据，库内集市层的数据来自于汇总数据层数据或者基础数据层数据，库内集市层数据加工完成后同步到各个应用中，并且对外提供相应的服务。

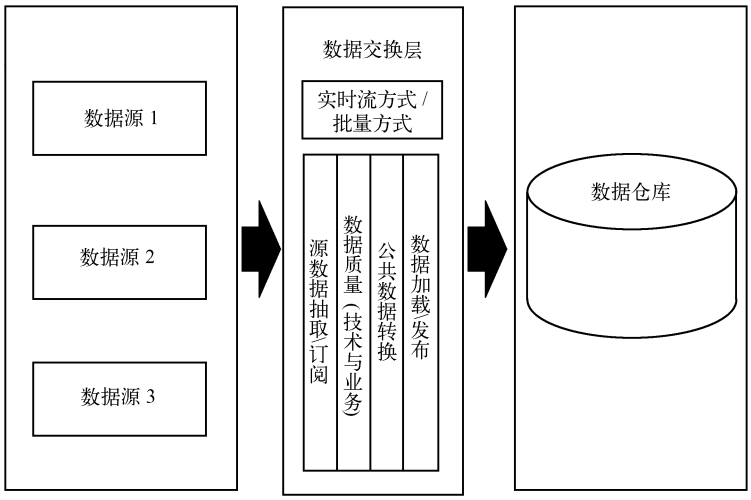


图 9-26　数据流转图

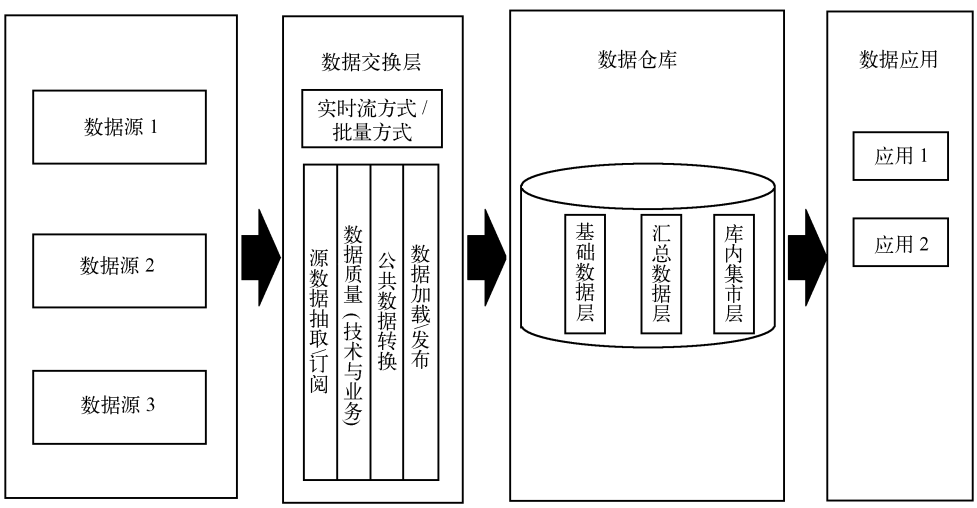


图 9-27　建设的内容

9.2.6　关于数据仓库系统的灾难备份规划

任何灾难造成的数据仓库系统停运，都会对一些重要机构产生重大的影响，特别是金融机构。根据国务院信息办《重要信息系统灾难恢复指南》《信息安全风险评估指南》和中国人民银行《银行业信息系统灾难恢复管理规范》，对灾难做如下定义：

灾难是由于人为或自然的原因，造成信息系统运行严重故障或瘫痪，使信息系统支持的业务功能停顿或服务水平不可接受、达到特定的时间的突发性事件，通常导致信息系统需要切换到备用场地运行。

很多金融机构为了预防灾难，都会对重要的系统建设同城和异地的数据备份中心，对于同城的数据备份中心来说，它可以接管所有核心的业务系统，而异地数据备份中心应该具备

恢复所需环境的能力，并且时刻处于运行或者就绪状态。下面详细介绍关于数据仓库的灾难备份（灾难备份）架构规划。

1. 灾难备份建设的方法论

关于灾难备份建设的方法论，主要分成以下几个阶段：分析阶段、架构设计阶段、技术方案选择阶段、实施阶段、维护阶段，如图 9-28 所示。

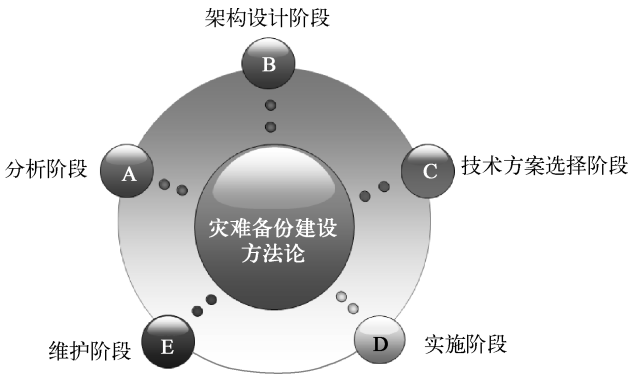


图 9-28　灾难备份建设的方法论

（1）分析阶段

分析因为中断和灾难对该机构造成的影响，确定系统恢复的优先顺序和相关性，包括恢复的时间目标和恢复点目标，明确关键功能的业务连续性需求等。

（2）架构设计阶段

确定灾难备份业务恢复策略，进行灾难备份架构的设计，以便在规定时间内恢复业务系统。

（3）技术方案选择阶段

在选择方案方面，首先了解 IT 系统建设现状以及发展趋势，其次是对灾难备份技术进行评估，提出方案建议，最后结合成本收益，选择最佳方案并实施。

（4）实施阶段

制定实施业务连续性的计划，便于在规定时间内完成业务的恢复。包括建立紧急事件处理中心。对于金融行业来说，一般都需要进行“两地三中心”的建设，例如，在第一期完成同城灾难备份中心的建设，第二期完成异地灾难备份中心的建设。

同城灾难备份中心是指生产中心和灾难备份中心在同一个城市或者相近区域内，主要防范火灾、建筑物破坏等灾难风险，保证在生产中心遭到灾难打击后，在极短的时间内可以快速恢复运营。但是同城灾难备份对大规模灾难的防范能力较弱。

异地灾难备份中心是指生产中心和灾难备份中心距离比较远，可能是跨省或者跨区域。利用先进的远程数据备份技术和可靠的网络通信可以实现异地灾难备份。

（5）维护阶段

开展对全部工作人员的灾难备份意识培养和技能培训工作。制定合适的规章制度和策略，以保证各个部门之间的协调响应。

2. 需求分析与灾难备份策略

针对灾难备份建设的需求分析与灾难备份策略设计，主要包含以下几个步骤：现状分析、风险分析、业务影响分析和灾难备份策略选择，如图 9-29 所示。

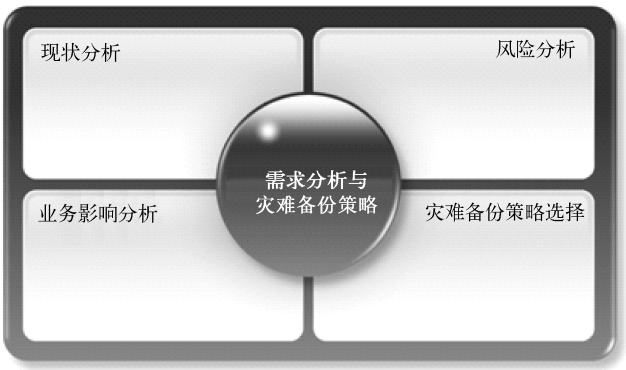


图 9-29　需求分析与灾难备份策略

（1）现状分析

现状分析主要包括对应用系统、网络情况和数据系统的分析。

应用系统的分析主要包括服务器部署环境分析、操作系统分析、数据库系统分析和应用关联关系分析等内容。其中服务器部署环境分析包括生产中心服务器系统现状分析（小型机服务器系统、PC 服务器系统等）、生产中心存储系统现状分析（存储系统情况、生产数据情况）、数据备份情况等。数据系统分析是对各个业务系统的数据存储情况进行分析。

（2）风险分析

风险分析需要对数据中心的物理环境、运行状况进行梳理。风险分析的结果可以作为业务连续性规划的工作数据。

从系统可靠性和性能的角度识别服务器、操作系统、数据库、存储和网络的风险。识别可能造成系统中断的各种风险。

根据识别出来的风险，判断是否在用户能够接受的范围之内。对于不能接受的风险，判断是否可以通过技术或者管理手段去防范和控制风险。同时提供降低风险和控制风险的合理建议。

风险分析工作的流程主要包括：前期调研、问卷整理、现场访谈，以及撰写及提交报告，如图 9-30 所示。

图 9-30　风险分析工作的流程

1）前期调研。

主要针对业务系统进行调研，了解 IT 系统的架构、业务运行情况和应用系统运行情况等内容，确定风险评估的应用范围。通过前期调研，了解相关部门的组织架构、人员职责等，为后面的问卷调研做好准备。

2）问卷整理。

通过对用户管理现状的调研，编写调查问卷，可以把调查问卷内容分成以下几个部分：IT 系统基础架构，开发和运维管理，基础设施建设，机房管理等。针对相应的管理人员和技术人员进行访谈，整理问卷的访谈结果，识别管理过程中存在的各种问题，制定对各种风险的分类和定义。双方达成一致。

问卷涉及的内容如图 9-31 所示。

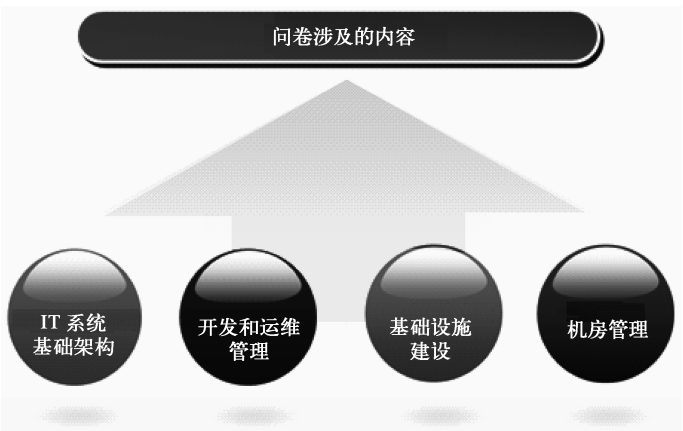


图 9-31　问卷涉及的内容

- IT 系统基础架构

IT 系统基础架构是针对 IT 基础架构管理设计的，问卷主要偏重于主机、数据库、网络和各种存储设备等。该问卷的目的是了解 IT 系统基础架构和运行维护方面的情况。

- 开发和运维管理

问卷主要偏重于软件架构的灵活性、安全性、可用性和可靠性的调研，目的是从架构的角度了解软件开发的部署、运维管理方面的情况。同时问卷也倾向于服务水平、故障处理、故障分类等领域。目的是收集在过去运维过程中发生的各类安全事件等信息。

- 基础设施建设

基础设施建设主要是针对基础设施现状和运维能力而设计的，问卷主要偏重于基础设施的建设标准、运行现状、管理水平和运行监控等能力的调研。

- 机房管理

机房管理主要针对机房基础设施管理进行调研，包括机房的运行能力，目的是收集机房运行的潜在风险和曾经发生的各类安全事故。

3）现场访谈。现场访谈是在问卷调研的基础上进行的，首先对问卷调研结果进行初步整理，确定访谈的策略，然后总结访谈的结果，得出相关系统脆弱性的列表。可以在 IT 部门范围内选择技术骨干进行访谈和交流。将调研结果和行业标准、最佳实践进行比较，把握

企业管理水平的现状，为降低和控制信息管理风险提供可行的意见。

4）撰写及提交报告。根据前期讨论的结果，结合信息管理风险的评估方法，进行风险识别、等级分析等工作。同时，撰写风险评估报告，正式提交文档。

（3）业务影响分析

业务影响分析（Business Impact Analysis），简称 BIA。英国标准协会制定的关于业务连续性管理对其定义为“一种分析机构的业务功能以及一旦业务中断所带来的影响的过程”。业务影响分析是通过调研，分析信息系统事故或者灾难造成业务中断时所产生的影响和业务恢复所依赖的资源，评估各业务功能的灾难恢复需求，为制定灾难恢复策略提供依据。

业务影响分析的流程如图 9-32 所示。

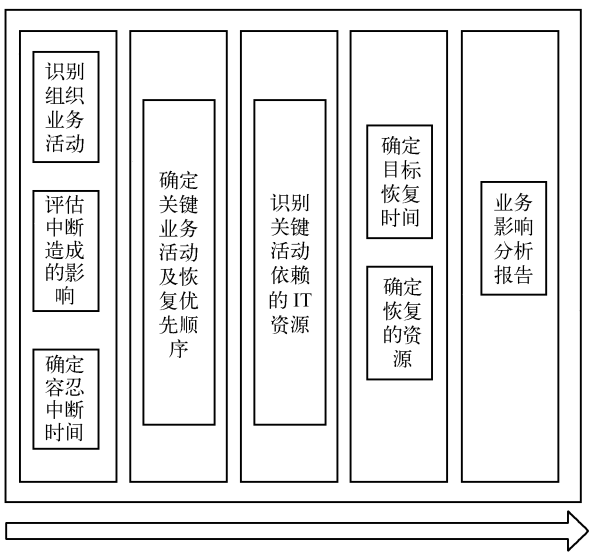


图 9-32　业务影响分析的流程

业务影响分析的流程是首先识别组织业务活动、评估中断造成的影响、确定容忍中断时间，然后确定关键业务活动及恢复优先顺序，识别关键活动依赖的 IT 资源，确定目标恢复时间、恢复的资源，最后形成业务影响分析报告。

具体的业务影响分析实施步骤如图 9-33 所示，主要包括前期沟通、调研问卷、培训、访谈和撰写报告。

1）前期沟通。主要针对业务影响分析的工作内容和方法进行沟通，根据实际情况，确定业务影响分析的工作范围和实施方式。

2）调研问卷。根据前期 IT 现状梳理及应用关联分析的结果，对调研问卷进行客户化修订，以方便业务人员能够准确、客观地进行填写。

3）培训。对业务人员进行业务影响分析问卷填写的培训，使参与实施的业务部门了解

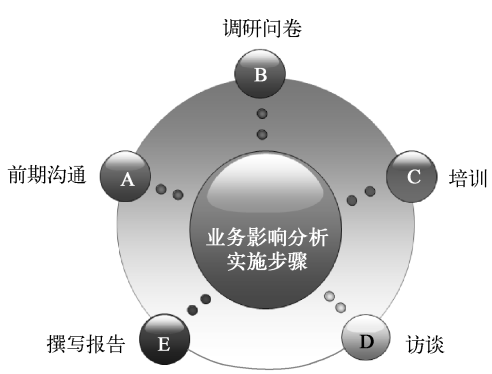


图 9-33　业务影响分析实施步骤

实施业务影响分析的意义。

4）访谈。各业务部门对业务影响分析调研问卷进行填写。确认问卷填写内容，了解补充信息。

5）撰写报告。根据调研收集的资料，完成报告初稿。最后整理出业务影响分析报告。

（4）灾难备份策略选择

通过业务影响分析，确定业务之间的关键功能和其中的关键点，决定业务连续性策略和所需成本。利用这一信息，管理层可以制订出合适的灾难备份策略。一般来说，典型的灾难备份中心策略包括：系统容灾的等级和灾难备份中心的运行模式。

关于灾难备份策略的选择，只有在充分调研现状的基础上，制定符合机构现状的策略，实现灾难备份系统建设的真正落地，才能发挥出应有的价值。灾难备份策略选择主要包含六个级别：

- 第一个级别

每周至少进行一次数据备份，在灾难应对方面，是经过测试和演练的灾难恢复预案。

- 第二个级别

在满足第一个级别的基础上，对备用数据处理系统和网络系统进行定义。

- 第三个级别

每天进行一次完整的数据备份，利用网络进行定时的数据备份传输。

- 第四个级别

在第三个级别的基础上，配置灾难恢复所需的全部数据处理设备和网络设备，并且处于就绪状态。

- 第五个级别

要求数据备份系统达到实时数据传输的能力，灾难备份中心可以提供 7 ×24 小时的技术支持能力。

- 第六个级别

要求达到对远程数据的实时备份，达到零数据丢失。

3. 灾难备份方案设计

为了提高风险管理能力，需要建立符合国际标准的业务连续性保障体系，主要包括需求分析、灾难识别、灾难备份启动、灾难备份恢复和灾难备份切换演练。通过“两地三中心”的规划布局，保障核心数据的安全和业务的连续性。

完成生产中心灾难备份系统的建设，使生产中心具备较强的防灾、抗灾能力，以避免因为意外灾难引起的不良后果，大大减少损失。

下面讲解主要从几个方面进行灾难备份体系的建设，如图 9-34 所示，包括灾难接管和恢复、应用处理能力、数据备份与数据复制、网络备份系统和“两地三中心”建设总体方案等。

其中“两地三中心”建设总体方案的目的是保证数据仓库系统的抗灾能力，系统可以快速恢复，如图 9-35 所示。

灾难备份系统建设的流程如图 9-36 所示，包括规划设计阶段、实施阶段和运营管理阶段。

- 规划设计阶段

规划设计阶段主要包括灾难备份需求分析、灾难备份建设规划、技术方案设计等内容。

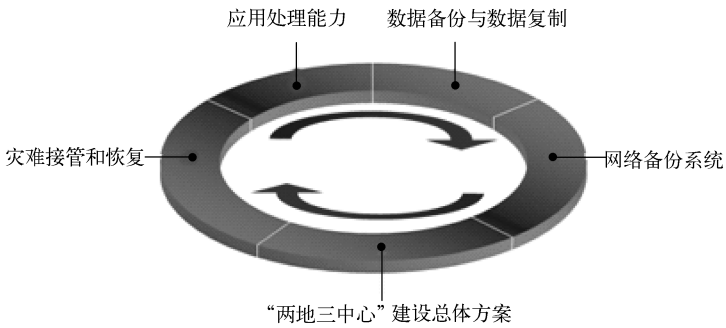


图 9-34　生产中心灾难备份系统的建设

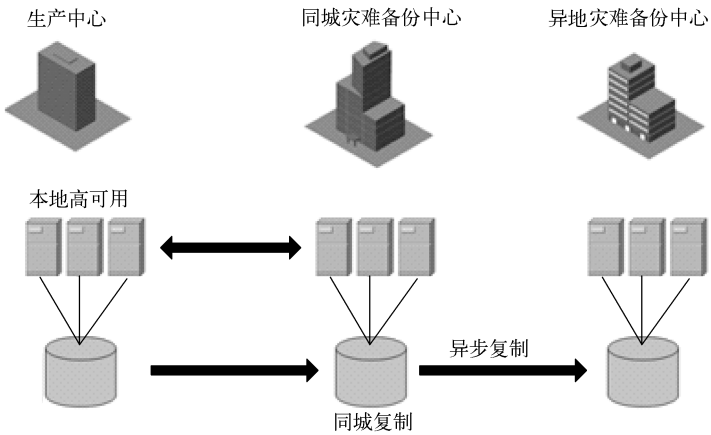


图 9-35　“两地三中心”总体建设方案

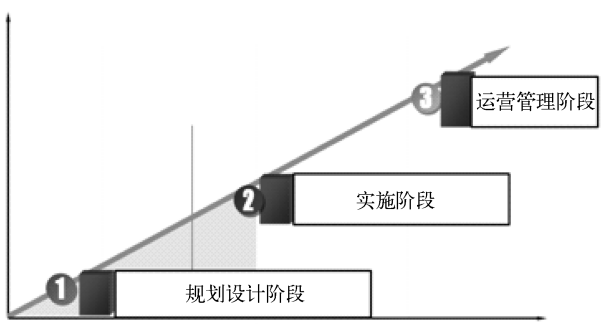


图 9-36　灾难备份系统建设的流程

● 实施阶段

实施阶段主要包括灾难备份中心的建设、灾难备份中心运营管理体系的建设、灾难备份系统的验证等内容。

● 运营管理阶段

运营管理阶段主要包括异地灾难备份系统日常运营管理、灾难备份系统切换、生产运行管理等内容。

4. 灾难备份应急预案与灾难备份演练

（1）灾难备份应急预案

灾难备份应急预案是在数据仓库系统灾难发生之前，建立相应的灾难恢复组织并制定相

关人员职责。这样可以确保灾难备份运行规范。

例如，当应用系统故障，存储系统故障，人为错误，网络故障，水灾、火灾、地震等灾难（见图 9-37）发生时，知道如何进行应急处理。

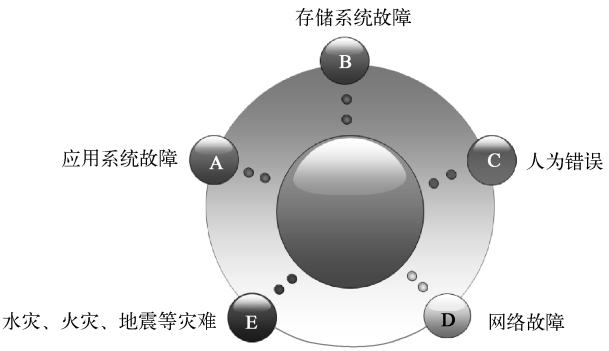


图 9-37　不同故障、错误或灾难发生时均有灾难备份应急预案

1）应用系统故障。当应用系统发生故障时，可以采用双机热备的方式进行切换。

2）存储系统故障。当存储系统发生故障时，可暂时采用本地存储替代生产系统。

3）人为错误。可以通过提取本地数据库快照，将数据恢复到灾难时间点前。

4）网络故障。当生产中心的网络发生故障时，通过设备冗余解决该问题。

5）水灾、火灾、地震等灾难。当水灾、火灾、地震等灾难发生时，通过重新部署硬件设施，利用灾难备份中心的业务数据，在短时间内恢复生产。

（2）灾难备份演练

根据数据仓库系统灾难备份技术方案，对灾难备份演练涉及的部门、人员，系统范围，演练步骤，进度安排，防范措施等内容提出建议。灾难备份演练需要制定灾难备份演练计划，实施容灾技术切换演练，对演练工作进行总结和评估。最后，针对演练过程中的问题提出改进建议。

当演练结束后，需要对相关预案及操作手册进行完善。

灾难备份演练可以有以下两种场景：

1）当生产中心发生火灾、数据丢失等事件时，会造成系统中断。这时可以直接启用同城灾难备份中心。例如，当灾难发生后，生产中心数据遭到损坏，造成系统不可用，业务中断，直接启用同城灾难备份中心接管生产。

2）当生产中心系统恢复后，回切生产中心，继续业务运行。

5. 灾难备份中心建设

数据仓库系统灾难备份中心的建设主要包括基础设施建设、人员组织机构建设、运维管理体系建设，如图 9-38 所示。

图 9-38　数据仓库系统灾难备份中心的建设

（1）基础设施建设

生产中心和灾难备份中心应该保持一定的距离，同时应该保证电力供给的可靠性及交通的便捷性，远离火灾隐患和地质、地震灾害的

高发区域。关于灾难备份中心的选址，应该考虑以下几种因素：地理位置、配套的设施、人力资源条件、地区政策、周边环境、建设和运营的成本，如图 9–39 所示。

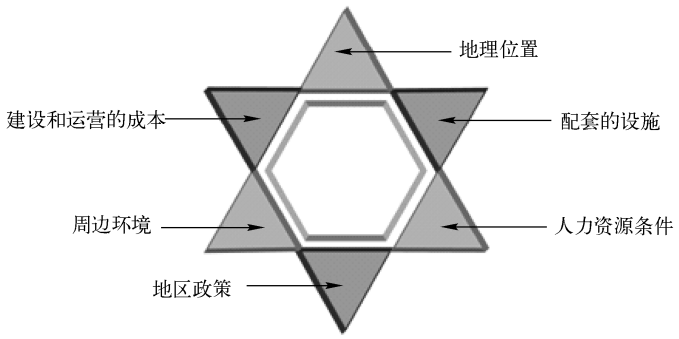


图 9–39　基础设施建设

1）地理位置。应该选择远离地震、台风、洪水等自然灾害频发地区，气候条件要舒适稳定，环境要清洁，交通也要方便。

2）配套的设施。配套的设施主要是指当地的经济发展水平、交通便利条件、人力资源供应等方面。灾难备份中心对各种社会资源的要求都非常高。

3）人力资源条件。人力资源条件主要是指当地的科技文化水平、人力资源是否充足等方面。

4）地区政策。地区政策主要是当地政府提供的政策。

5）周边环境。所在地的周边环境条件。选址时应避开生产或存储易爆物产品的工厂、仓库等，远离高速路、铁路等，避免震动对于主机的影响。

6）建设和运营的成本。成本是一个需要反复权衡的因素。成本一般涉及当地的土地价格、房屋建筑价格、通信费用、用电价格和人力成本等多种因素。

（2）人员组织机构建设

人员组织机构建设主要是指建立或设立项目领导小组、项目技术委员会、项目经理、项目管理组、项目实施组和项目支持组。

● 项目领导小组。

人员构成：由项目负责人和客户项目负责人组成。

具体职责：协调项目参与方与客户相关部门的关系，协调解决各方的重大争议，协调项目与厂商的合作关系。审核项目的总体方案和实施计划等。对项目的进度、质量状况和风险等进行宏观调控，对项目的各个方面进行管理，协调用户内部、各厂商及合作伙伴之间的关系。制定计划，明确分工责任等。

● 项目技术委员会。

人员构成：由技术专家组成。

具体职责：技术专家主要负责项目总体技术的把关，以及解决重大技术问题。

● 项目经理。

人员构成：项目管理人员。

具体职责：负责项目的组织、管理和协调；制定项目实施方案和计划；协调项目成员与

用户人员之间的工作关系；负责监督项目的具体实施，安排各阶段工作任务；负责向项目领导小组汇报项目进展情况。

● 项目管理组。

人员构成：由项目质量管理人员组成。

具体职责：作为项目的质量保障机构，负责制定质量标准和计划等，参与项目的实施，负责监督项目的实施过程，并在发现问题后进行处理和改进。

● 项目实施组。

人员构成：由项目实施人员组成。

具体职责：负责软硬件设备的安装、调试。汇报项目各阶段的进展情况和存在的问题等。负责对用户运营维护人员的技术培训。

● 项目支持组。

人员构成：由技术专家组成。

具体职责：负责系统规划和项目实施的审核工作。为项目实施组提供技术支持。负责解答用户的专业技术问题。

（3）运维管理体系建设

数据仓库系统运维服务管理对象包括基础设施、应用系统、用户、运维部门及供应商。

具体内容如下：

● 基础设施

主要包括网络、主机系统、存储和备份系统、安全系统等。

● 应用系统

主要包括办公系统、门户网站等应用系统。

● 用户

主要包括使用产品或服务的一方和产品或服务的购买者。

● 供应商

主要包括基础设施、应用系统和 IT 运维的供应商。

● 运维部门

主要包括参与运维活动的相关部门和人员。

9.3 商业银行数据仓库的建设规划

9.3.1 商业银行数据仓库建设概况和瓶颈

2000 年以后，多数商业银行都在建设数据仓库，经过前期的数据积累，数据质量的提升，数据仓库建设成功率较高。

商业银行数据仓库的建设一般都采用分阶段建设的策略。

第一阶段，基本实现对数据的集中处理，特别是对内部重要报表系统提供数据支持。

第二阶段，进一步实施诸如资产负债管理、客户关系管理或者某些灵活报表查询等较为复杂的管理分析类应用。

第三阶段是在第二阶段的基础上，实施数据挖掘分析、商业智能等应用。

对于商业银行来说，建设数据仓库是基本功，缺点是实施的周期较长，统一标准困难，见效慢，是一个典型的高投入和慢回报的建设项目。

但是随着时间的流逝，建设数据仓库或者不建设数据仓库给商业银行的发展带来了不同的影响。例如，如果某些商业银行在 5 年前或者 10 年前就开始重视数据仓库的建设，比那些不重视数据仓库建设的商业银行发展态势要好很多。也就是说，后期投入的成本就越高。

一些商业银行在建设数据仓库时面临很多的困难，例如很多银行为每一个应用系统建设数据库，当多个应用系统建设完成之后，增加了数据管理的难度。因为数据标准不统一，所以整合难以实现。

举例来说，客户使用银行服务的渠道很多，除了传统的营业网点，还包括网上银行、手机银行等渠道，这会导致同一个客户可能会拥有多个账户信息，那么识别唯一客户需要大量的数据整合和集成工作。如果不能对客户信息进行唯一识别，就很难进行商业智能分析。

目前商业银行数据仓库面临很多瓶颈，包括业务价值、系统性能、数据质量和后续运维等，如图 9-40 所示。

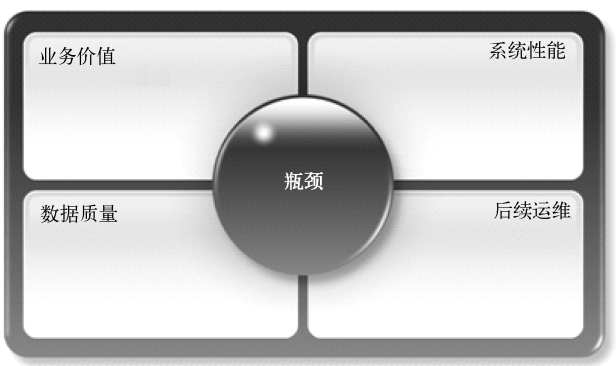


图 9-40　目前商业银行数据仓库面临的瓶颈

（1）业务价值

对于大部分的商业银行来说，数据仓库基本上停留在业务报表和供数层面上，业务价值体现得不够充分。商业银行有大量的数据，有的银行也做了很多的数据分析，但是没有把业务分析结果转换为业务行动。国外很多银行都会把分析结果转变成业务规则或者业务事件，然后和银行的业务系统糅合到一起，最终形成一个闭环结构。国内商业银行的高端分析应用不多，也就是没有把对业务的分析结果转换为业务行动。

（2）系统性能

对于商业银行来说，如果数据仓库系统的数据链路过长，例如数据从核心业务系统加载下来，然后再通过交换系统、缓冲区，经过 ETL 加工，最后到应用系统。这种方式必须考虑数据的混合负载，也就是数据加载、数据加工和前端访问同时进行，可以进行批量加载和实时加载。这种工作负载是混合的，需要重点考虑资源的分配问题。

（3）数据质量

对于大部分商业银行来说，基本上都是先有数据仓库，然后才进行数据标准的建设，这样会导致数据标准很难在数据仓库中落地。大部分商业银行实行了数据质量检查程序，对数据仓库的上游、中游和下游进行全生命周期的质量管理，但是对于前台业务系统，也可能有数据质量问题，所以需要统一起来。

（4）后续运维

当数据仓库建好之后，每天都在加载数据，模型也在不断扩充，如果有新的数据源加进来，模型就需要变化，ETL 程序也需要修改，这样维护的工作量非常大。同时还需要考虑数据自助服务，开放数据接口，也就是业务人员通过接口自助服务，临时取数。但是一般来说，这种灵活查询不能全部开放，因为数据仓库的数据量非常巨大，有可能一个查询会影响整个数据仓库系统，对于开放的查询只能开放一些汇总数据层的数据。而关于明细的基础数据层、交易层的数据是不能开放的。

9.3.2 商业银行数据仓库建设面临的问题和改进建议

我们从 4 个维度（架构、模型、管理、应用）说明商业银行的数据仓库建设存在哪些问题，如图 9-41 所示。

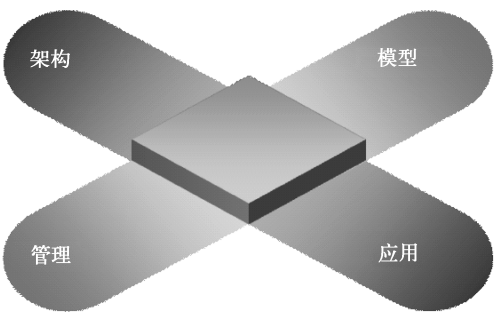


图 9-41　4 个维度（架构、模型、管理、应用）

一般来说，很多商业银行的数据仓库架构面临的较大问题是数据链路过长，架构的灵活度不够，系统在高可用性上还处于较低的水平，模型的稳定性不够，同时语义层不丰富。在管理上，多数商业银行对管理一个越来越庞大的数据仓库系统经验不足，在元数据管理和数据质量管理上都有改进的空间，同时在数据仓库的基础上开发的部门越来越多，如何管理跨部门之间的使用已经成为了一大难题。在应用上，商业银行对数据仓库的建设多采用传统的 OLTP 应用的开发、测试方法，效率较低。

下面针对这 4 个方面的问题，分别提出改进的建议。

首先，对于架构上的问题，可以缩短信息链路，或部署沙盒，执行一些具有高可用性特点的方案。

其次，对于模型上的问题，可以进行相应的模型优化，同时要求数据仓库的上游系统提高稳定性，完善数据仓库的语义层。

再次，对于管理上的问题，可以借鉴同行业的先进经验或者海内外先进经验，同时也可以升级元数据管理系统和数据质量管理系统。为了保证在数据仓库的基础上，各个部门之间的管理和协作，应该制定数据仓库开发规范，并且严格执行，同时制定部门接入数据仓库的准入制度。

最后，对于数据仓库应用上的问题，应该对现有的开发、测试方法进行创新，增强对灵活查询的支持，同时需要敏捷开发。

9.3.3 商业银行数据仓库建设思路及系统情况

1. 商业银行建设数据仓库时遇到的挑战

商业银行建设数据仓库时遇到的挑战主要包括高可用性、组织架构、数据质量和性能/数据延迟性，如图 9-42 所示。

（1）高可用性

在单一物理环境中集中了数据缓存、ODS、数据仓库和数据集市，这样会严重影响系统

的高可用性，同时会引发一系列关于性能、可扩展性和可维护性等问题。

因为缺乏对负载的管理或者是相关政策实施监管不到位，所以造成了资源的相互争夺，使得系统不能提供很好的服务。

（2）数据质量

由于如果数据仓库中存在大量的不一致的数据和冗余的数据，则对于数据质量的维护来说是非常被动的，所以应该保证数据仓库中的数据都是有用的。

（3）组织架构

很多商业银行缺少与数据治理相关的人员角色和岗位，不能保证业务部门和 IT 部门的目标是一致的，导致数据仓库的建设缺乏长远的、与商业银行的业务战略一致的规划。

图 9-42　商业银行建设数据仓库时遇到的挑战

（4）性能/数据延迟性

对于很多商业银行的数据仓库来说，查询的并发度是一个很大的挑战，多用户使用数据仓库运行的报表或者是即席查询的时候，系统很难进行扩展和对负载进行优先级的处理。

2. 商业银行数据仓库架构问题及案例分析

（1）第一个案例

商业银行在建设数据仓库的时候，可能会存在其他某商业银行的数据仓库架构问题，下面分析一下对这类数据仓库有哪些可以改进的地方，如图 9-43 所示。

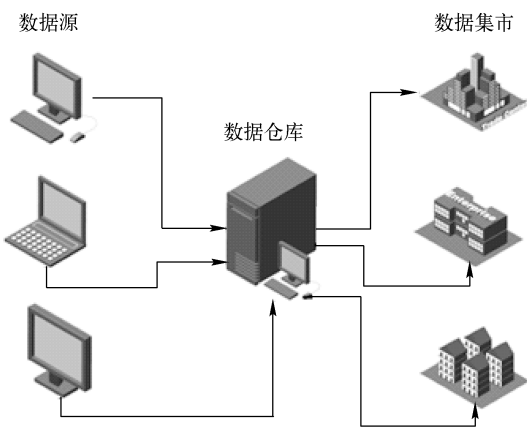


图 9-43　某商业银行的数据仓库架构

现状：

该商业银行的业务系统每天将文件放入到数据仓库中，如今的数据仓库在压缩前存放着大约 80TB 的数据，压缩后有 45TB，日增量大概有 300 ~ 400GB，在峰值时可能会有800 ~ 900GB 的数据。

需要优化的地方：

整体的数据架构需要优化，面临着数据如何迁移，缺少统一的数据管控体系，缺乏大数据处理机制，数据模型没有统一规划等很多问题。

在核心银行业务系统向数据仓库传送文件的过程中缺少文件交换平台，文件被直接送入到数据仓库中，缺少数据缓冲区。因为业务系统与数据仓库之间缺少缓冲区，这意味着数据仓库缺少了一道屏障。

首先，因为数据仓库存储着大量的历史数据，同时为多个应用提供服务，所以系统的效率可能是个瓶颈，如果再与多个业务系统建立连接，会大大降低数据仓库系统的高效性。

其次，缓冲区相当于数据进入到数据仓库系统的一道闸门，很多事情可以在缓冲区完成。例如，对数据质量的校验，对“垃圾”数据的“清洗”，目的是保证数据的一致性和正确性。然后从缓冲区中将数据迁移至数据仓库，保证流到数据仓库的数据都是高质量的数据。

最后，数据仓库面对的是数据缓冲区这唯一的数据源，把该缓冲区当作唯一可信的数据源，只需要建立一个连接即可，会大大提高数据仓库系统的性能。

同时该系统缺乏库内集市和库外集市的合理规划，根据性能的要求，应用可以分成库外数据集市和库内数据集市。划分的原则是需要考虑性能问题，如果数据访问量很大，计算复杂，则需要用库外数据集市；如果访问量小，计算简单，则考虑库内数据集市。

（2）第二个案例

下面看一下某商业银行的数据仓库数据架构，如图 9-44 所示。

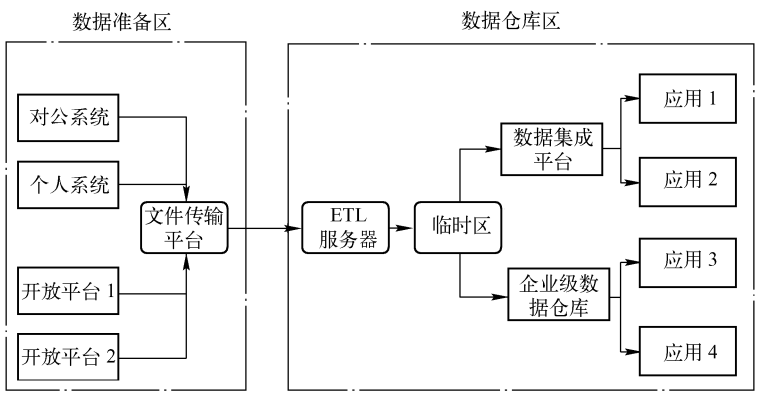


图 9-44　某商业银行的数据仓库数据架构

现状

从主机对公系统、主机个人系统和开放平台，每天通过文件传输平台，到 ETL 服务器，数据通过解压、压缩，每天传输的数据量是 450GB，先放入临时区（该临时区一般只存储一周的数据）。该临时区的数据是为了做数据加工准备的，是贴数据源的。从临时区出来分了两条路径，所谓的数据集成平台相当于 ODS 系统。如果应用是不跨系统的，同时要求数据的时效性高，则该应用从数据集成平台中取数据；如果该应用要求跨系统取数，但是要求的时效性不高，则该应用从企业级的数据仓库中取数据。

企业级的数据仓库分成基础数据层、汇总数据层。针对数据仓库的应用也可以分成库外的数据集市和库内的数据集市，原则是考虑性能的问题。如果数据访问量很大，要求的时效性高，则需要考虑库外的数据集市。如果数据访问量小，则可以考虑使用库内的数据集市，

也就是在数据仓库内做视图。

需要优化的地方

该商业银行的数据仓库逻辑架构存在问题，例如时间窗口过长，也就是数据的链路太长。解决的办法是通过主机直接连到数据集成平台，可以通过产品实现。在时间调度上，如果某个业务的数据很快加载完了，就可以先提供访问，不需要等所有的业务全部加载完之后再提供数据访问。可以通过 ETL 将业务之间的相互关系拆开，在没有相互依赖的情况下，某个业务的数据加载完之后就可以提供访问了。

3. 对商业银行数据仓库目标数据架构的建议

对于数据仓库的目标数据架构，可以提供以下建议，如图 9-45 所示。

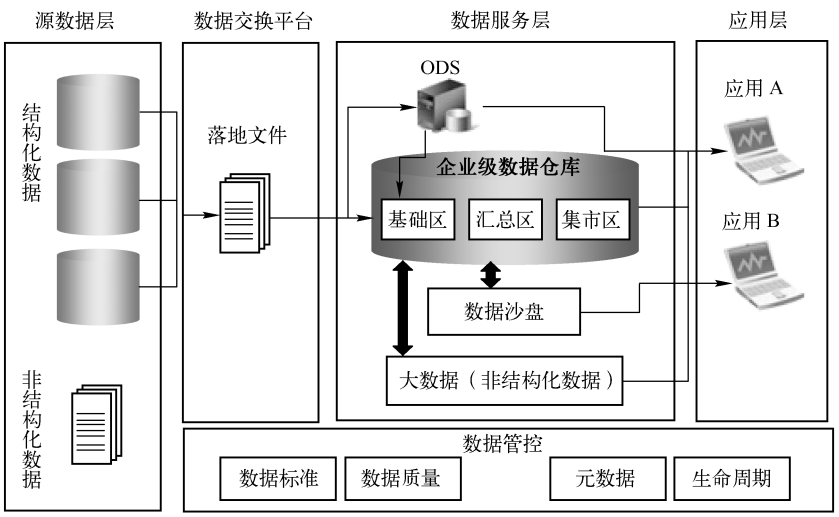


图 9-45　数据仓库的目标数据架构

1）在数据源层和数据服务层之间建立一个数据交换平台。数据服务层内部的数据流动和数据交换都通过数据交换平台。ODS 相当于数据的集成平台，存储的都是实时性的数据，而数据仓库存储的都是历史数据。

2）数据仓库可以分成数据基础区、数据汇总区和集市区。

3）数据沙盘的使用。如果某个应用从数据源层通过数据交换平台到 ODS，到数据仓库层，再到数据集市层，可能数据的链路过长，从而影响应用的时效性，这样就可以建立一个数据沙盘，可以直接从 ODS 取数，或者从数据仓库、数据集市中取出数据，当稳定和固化后，再把应用挪到 ODS 或者数据仓库、数据集市中。

4）所有的数据流动都有统一的调度工具进行调度。

5）同时建立对数据的分布和流转的管控，包括元数据管理、数据质量管理、数据标准管理和数据生命周期管理等内容。

关于商业银行数据仓库的目标数据架构，主要包括源数据层、数据交换平台、数据服务层、应用层。源数据层对于各个 OLTP 生产系统，如一些核心业务系统等，时效性要求较高，一般只存储生产数据，不存储历史数据。它一般作为数据仓库的主要数据来源。源数据层还可能包括文件系统、Web 等非结构化数据源。

数据服务层为数据仓库所在层，通过对历史细节数据的存储和汇总数据的加工，支持后续的应用。数据服务层结合业务的需要可以设计成库内集市或者库外集市。

应用层将数据服务层加工出的数据，通过静态报表、动态 OLAP 等处理方式提供给用户。

9.3.4 商业银行数据仓库建设启示

对于大多数商业银行来说，数据仓库的建设不是一蹴而就的，一般是分阶段、分期实施的，然后逐步建设数据仓库的模型，最后对应用形成支持。在数据仓库的建设过程中，需要业务部门主导及深入参与，深入发掘和分析业务管理方面的需求，并且建立相应的数据管控体系。同时需要团队培养和知识积累等工作。

如图 9-46 所示，首先启动数据仓库的建设，由业务部门主导及深入参与，然后将数据仓库的建设和数据管控的工作结合起来，最后分期、分阶段地进行建设，同时注重团队培养。

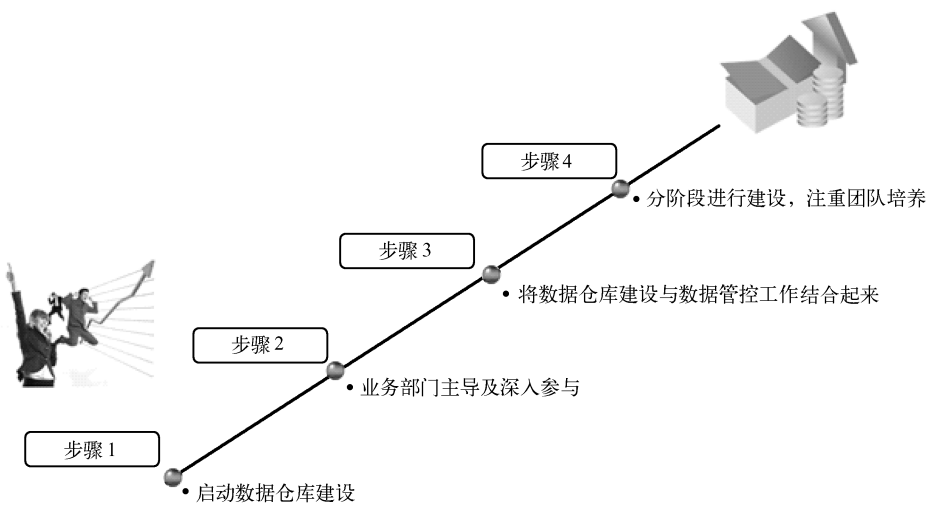


图 9-46　商业银行数据仓库建设启示

1）根据银行业务运营、客户关系管理、资产负债管理和决策分析等需求，在合适的阶段启动数据仓库项目的建设。

很多商业银行为了满足多方面的需求，例如风险管理、绩效管理或者监管合规等多方面的要求，在核心业务系统数据集中和建立统一数据源之后，启动企业级数据仓库的建设。

在业务数据量相对较小的时候启动数据仓库项目的建设，可以降低数据仓库系统建设的难度和风险，能够尽快体现出数据仓库的价值。

2）数据仓库的建设需要业务部门主导及深入参与。

对于商业银行来说，数据仓库的工作需要业务部门人员的广泛参与，并且由业务部门牵头发起数据仓库的建设，深入挖掘和分析业务管理方面的需求，从而指导数据仓库模型的设计等核心工作。

数据的集中过程也需要业务部门的参与，包括完成数据的清洗和整合工作，在此基础

上，深入挖掘信息，有效发挥数据仓库的价值。

3）将数据仓库建设与数据管控工作结合起来。

对于商业银行来说，数据标准为数据仓库提供统一的定义，它是数据仓库的重要基础，如果先进行数据仓库的建设，后期再进行数据标准的建设，会对数据仓库的建设带来一定的负面影响。数据仓库的建设需要与数据管控结合起来，这样会有效提升数据仓库的数据质量，从而保障数据的可信度。

4）分阶段进行建设，注重团队培养

商业银行的数据仓库项目一般是分期、分批迭代进行的，不能一蹴而就。而且数据仓库项目复杂度相对较高，需要有丰富专业知识的技术人员和业务人员才能将数据仓库项目建设好，因此，需要商业银行重视对数据仓库方面人才的培养，包括技术开发人员、设计人员和运维人员的培养。

9.4 电力行业数据仓库的建设规划

9.4.1 电力行业数据仓库建设难点

电力行业数据仓库的建设存在以下难点（见图 9-47）：

缺乏统一的数据规划
数据质量问题
缺乏数据仓库建设在试点单位的验证
缺乏对数据仓库建模的具体指导意见

数据不统一、不完整、不正确
成功的经验无法得到有效的推广

图 9-47 电力行业数据仓库的建设存在以下难点

1）在电力行业里，有些部门缺乏统一的数据规划。

2）因为缺少标准化的数据模型和统一的编码管理，所以经常导致出现数据质量问题。

3）电力行业有时会缺乏数据仓库建设在试点单位的验证。

4）电力行业有时同样缺乏对数据仓库建模的具体指导意见。

5）因为以上的原因，电力行业的系统建设很容易形成信息孤岛，导致数据不统一、不完整、不正确。

6）同样也会导致电力行业系统建设的成功经验无法得到有效推广。

解决数据仓库建设难点问题的方法

解决电力行业数据仓库建设的难点问题有以下方法：

可以通过试点建设积累经验。形成统一的数据模型标准、管控方法和建设流程，再大面积推广，如图 9-48 所示。

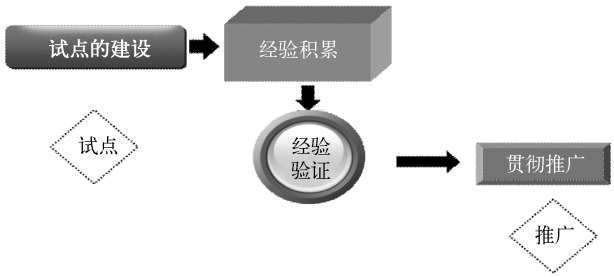


图 9-48　解决电力行业数据仓库建设的难点问题

9.4.2　电力行业数据仓库体系架构

对数据的应用分析通过数据仓库和数据集市提供数据支持，并通过前端展示层，将分析的结果展现给最终用户。电力行业关于数据仓库的体系架构如图 9-49 所示。

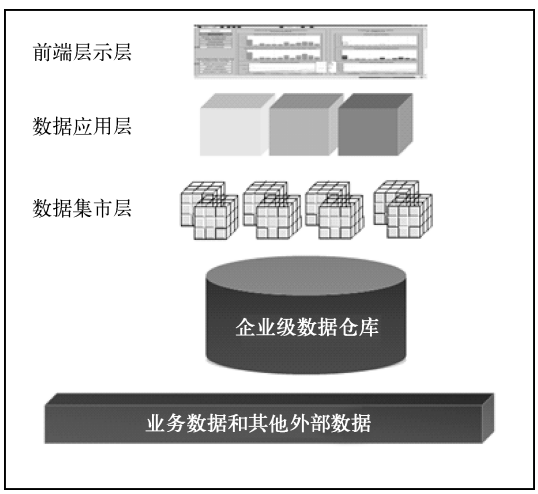


图 9-49　电力行业数据仓库体系架构

9.4.3　电力行业数据仓库能力蓝图

电力行业的数据仓库应该具备以下几种能力：集中整合能力、分析展现能力、高级应用能力、数据移动能力、质量保障能力和信息描述能力，如图 9-50 所示。

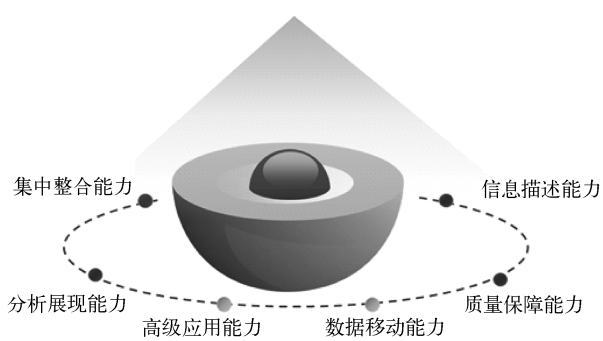


图 9-50　电力行业数据仓库应该具备的几种能力

（1）集中整合能力

数据按照业务主题的方式进行存储，完成对零散数据的整合工作，形成“唯一数据”。

（2）分析展现能力

通过标准化的报表和图表帮助管理人员做出正确分析和决策，针对具体应用建立相应的集市，并且提供高效的数据查询和服务。

（3）高级应用能力

为分析人员和管理人员提供多维分析的能力，帮助用户从多个维度深入分析需要的指标；同时还需要具备数据挖掘的能力，能够对企业的状况和未来发展作出完整、合理和准确的分析预测。

（4）数据移动能力

提供数据抽取、转换与加载的能力。可以高效地将业务分析需要的各类数据移到数据仓库中。

（5）质量保障能力

数据仓库应该具备完善的数据质量管理机制，保障企业内部数据的一致性与准确性，提升数据分析的可信度。

（6）信息描述能力

应该具备强大的元数据管理功能，以实现各类技术术语与业务术语在公司内部的统一定义。

9.4.4 数据仓库对电力业务发展的促进作用

数据仓库可以促进电力业务的发展，如图 9–51 所示。

图 9–51　数据仓库可以促进电力业务的发展

数据仓库的建设可以提高电力安全运营的能力、绩效分析的能力、电力营销管理的能力和决策分析的能力。具体表现是通过对电力设备的运行状况、检修情况和事故的及时掌握，提高电力安全运营的能力，通过对电量、电费、电价的分析，提高对电量的需求预测能力和价格制定能力，这样可以提高电力营销管理的能力。通过完善报表管理的能力，为分析人员提供全面的关键业务信息，同时对运行状态进行分析和监控，可以提高绩效分析和决策分析的能力。

9.4.5 数据仓库建设策略比较

（1）第一种数据仓库建设策略

由业务部门建立各自的数据集市，这种方式会造成重复的 ETL 开发，导致缺少企业层面的统一规划和协调，造成资源的浪费，同时因为缺少跨业务系统数据的支撑，所以无法提供全面的分析能力，也容易出现不一致的情况。但是由于建设方式简单，一般来说，设计、开发的周期都较短。

（2）第二种数据仓库建设策略

业务部门根据自身需求，在统一的数据仓库平台上建设更深层次的数据分析应用，这种建设方式可以有效地形成企业范围的统一信息视图，可重用 ETL 流程，减少资源的浪费，通过更丰富的企业数据支撑，提供全面的企业级的数据分析能力。可以制定统一的数据管理机制，提升数据的质量，但是因为设计、开发的难度较大，所以时间周期也相对较长。

9.4.6 电力行业数据仓库的数据架构设计

电力行业数据仓库的数据架构设计如图 9-52 所示。

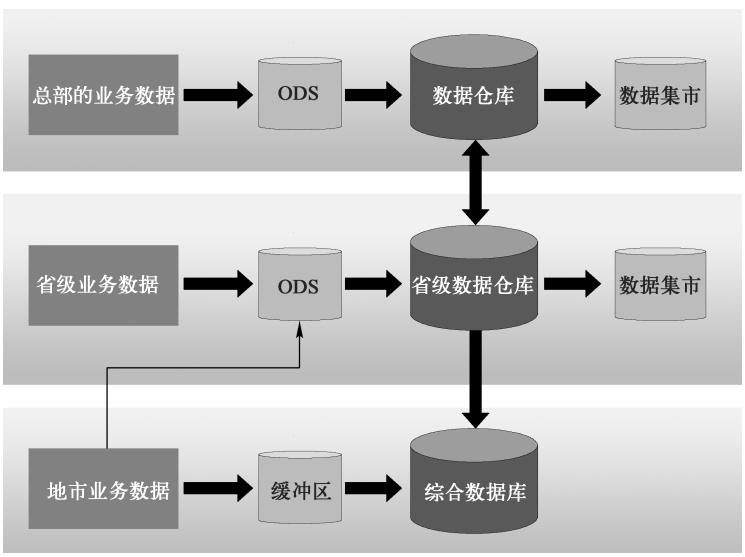


图 9-52　电力行业数据仓库的数据架构设计

1. 具体说明

1）电力公司总部 ODS 的组成与省级单位的 ODS 相同，主要区别在于数据源的不同，电力公司总部 ODS 主要的数据源来自于总部的业务系统，而省级单位 ODS 主要数据源来自于省级单位的业务系统。

2）电力公司总部数据仓库的数据来源包括业务明细和汇总的数据，省级电力公司数据仓库的数据是以中度或者高度汇总的数据进行存放。

3）电力公司总部的数据集市主要针对公司整体发展分析，跨系统和跨省地对数据进行全面挖掘。

4）省级电力公司的数据仓库主要覆盖多个主题域的企业信息，这些信息主要是低级别的、细粒度的数据，同时根据分析需求建立一定粒度的汇总数据。它们主要为数据集市提供整合后的、高质量的数据。省级数据仓库和总部的数据仓库存在数据交换的功能，同时将一部分数据下发到地市级中。

5）省级电力公司的数据集市是一组特定的、针对某个主题域的、部门的数据集合。这些数据需要针对用户需求进行快速访问，数据集市可以保障数据仓库的高可用性、可扩展性和高性能。

2. 数据移动说明

（1）ODS 缓冲区数据抽取到数据仓库区

数据仓库区是核心的数据存储区域，它支持大部分的数据应用。

数据仓库内的数据一般按照面向主题的方式进行组织和存放。数据模型满足第三范式，这些数据在线存储的周期一般较长，而 ODS 缓冲区中的数据结构和业务系统相似，它起到缓冲的作用。从 ODS 缓冲区，数据经过转换、映射、清洗，最后加载到数据仓库区中。中间的过程包含了合并、匹配、数据的追加（覆盖、更新）等操作，如图 9-53 所示。

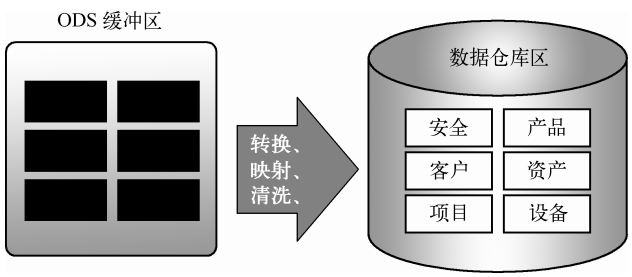


图 9-53　ODS 缓冲区数据抽取到数据仓库区

（2）数据仓库区数据抽取到数据集市区

数据集市是针对某个主题域、部门的数据集合。这些数据需要被快速访问。数据集市的数据模型可以是星形结构和雪花形结构。而数据仓库的数据模型满足第三范式。从数据仓库到数据集市的数据迁移，应该重点考虑从规范化建模到多维建模的映射关系，包括实体表和事实表、维表之间的映射关系以及转化过程。主要的过程包含了汇总、缓慢变化维等操作，如图 9-54 所示。

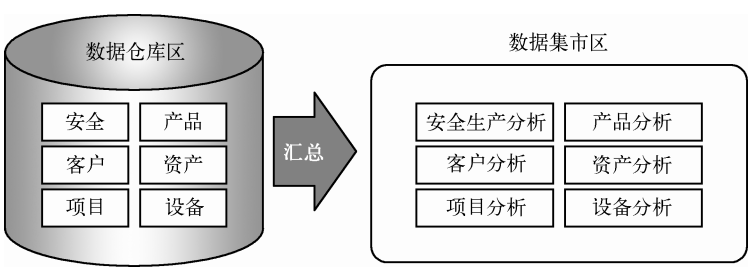


图 9-54　数据仓库区数据抽取到数据集市区

（3）总部数据仓库和省级数据仓库之间的数据交换

总部数据仓库的数据源主要包含两部分的内容：一部分是总部的业务系统数据，另外一

部分是省级电力公司数据仓库的数据。省级电力公司定时向总部数据仓库上传数据以供分析使用，同时总部数据仓库也会定期将汇总的数据下发到省级电力公司，如图 9-55 所示。

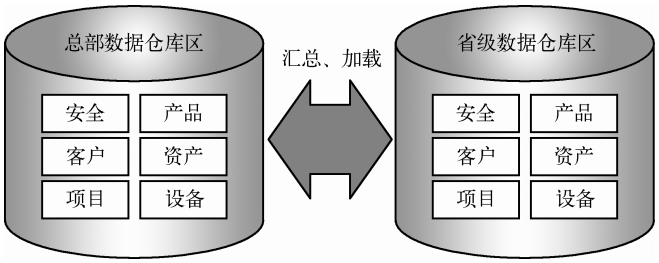


图 9-55　总部数据仓库和省级数据仓库之间的数据交换

一般来说，总部的数据仓库只抽取部分省级电力公司的数据，同时存储跨系统、高度汇总和集成的数据。

（4）省级数据仓库和地市级综合数据库的数据交换

省级数据仓库会定期将相关数据加载到地市级综合数据库，如图 9-56 所示。对于地市综合数据库来说，它的数据主要来源于省级数据仓库下发的数据和部署在地市级别的业务系统的数据。地市级综合数据库也可以看作地市级的数据仓库。

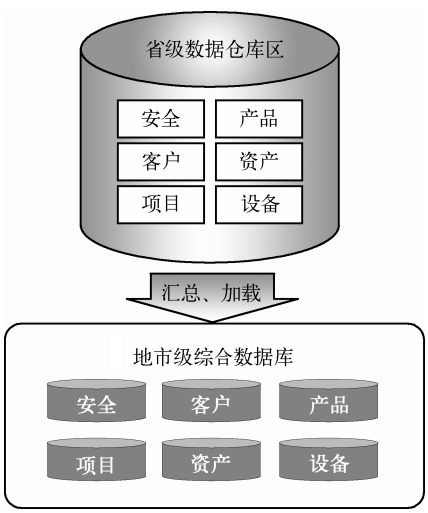


图 9-56　省级数据仓库和地市级综合数据库的数据交换

小结

- 数据仓库在比尔·恩门所著的《如何构建数据仓库》一书中的定义："数据仓库是一个面向主题的（Subject oriented）、集成的（Integratel）、相对稳定的（Non - Volatile）、反映历史变化的（Time - variant）数据集合，主要用于支持决策分析"。
- 数据仓库是一个过程，而不是一个产品。数据仓库的整个过程包括很多产品和实施服务。

- 数据仓库是实现商业智能的基础平台，没有数据仓库的搭建，商业智能是无法实现的。
- 数据仓库系统建设应该考虑以下问题：

首先选择数据仓库系统的成功案例作为重要参考。

学习行业内的先进经验。

具备专业的数据仓库实施队伍和业务领域的专家。

考虑数据仓库是否满足海量数据的复杂、并发查询。

数据仓库应该满足可扩展的能力。

数据仓库应该考虑高可靠性，并且满足高质量的要求。

- 数据仓库系统相比其他系统有下面几种优势：

数据仓库系统可以获取生产系统综合的信息，作为科学决策分析的重要依据。

数据仓库可以从宏观的角度理解信息，也可以从微观的角度探查信息。

通过数据仓库系统，可以建立企业各个部门之间的联系。

- 传统数据仓库所带来的困难，使企业管理层无法获得及时、准确、有效的业务信息，这会对企业的运营和竞争力带来影响，原因如下所示：

缺乏有效的目标市场定位，难以推出有针对性的产品。

不能够根据个性化的服务需求，制定出对应的营销策略。

不能及时了解客户的需求和特征，无法提高客户的忠诚度。

- 数据仓库的技术特性：

海量数据处理能力。

高可用性。

线性的扩展能力。

数据压缩能力。

- “制定数据标准，建立数据管控机制，以数据、应用驱动为主”是数据仓库基本的建设方法论。
- 数据仓库架构设计遵循原则：

可重用性。

高性能。

可扩展性。

可管理性。

高可用性。

- 数据仓库有以下几个特征：

数据仓库整合系统的全局信息，包括基础数据层、汇总数据层和库内集市层。

数据仓库中的数据通常包含历史信息，记录了从过去某一时间点到目前各个阶段的信息，通过这些信息，可以对企业的发展历程和未来趋势做出分析和预测。

数据仓库的数据来源可以为结构化的基础数据，非结构化数据结构化的数据，也可以是产品加工后的产品数据，或者是主数据等。

数据仓库中的库内集市是根据应用需求形成的数据集合，它支撑了各种专业化的应用。

- 关于灾难备份建设的方法论，主要分成以下几个阶段：分析阶段、架构设计阶段、技

术方案选择阶段、实施阶段、维护阶段。

- 大数据是指无法在一定时间内，用传统型的数据库软件对其内容进行抓取、管理和处理的数据集合。大数据用于在成本可承受的条件下，通过快速采集、发现和分析，从大量的、多类别的数据中提取价值。大数据是一系列技术的集合，汇集了如Hadoop/Mapreduce、一体机、NoSQL，数据分析与挖掘、商业智能、数据仓库等。
- 商业银行数据仓库的建设一般都采用分阶段建设的策略：第一阶段，基本实现对数据的集中处理，特别是对内部重要报表系统提供数据支持。第二阶段，进一步实施诸如资产负债管理、客户关系管理或者某些灵活报表查询等较为复杂的管理分析类应用。第三阶段就是在第二阶段的基础上，实施数据挖掘分析、商业智能等应用。
- 对于大多数商业银行来说，数据仓库的建设不是一蹴而就的，一般是通过分阶段、分期实施的，然后逐步建设数据仓库的模型，最后对应用形成支持。在数据仓库的建设过程中，需要业务部门主导及深入参与，深入发掘和分析业务管理分析方面的需求，并且建立相应的数据管控。同时需要团队的培养和知识的积累等工作。
- 对于数据仓库的目标数据架构，可以提供以下建议：

在源数据层和数据服务层之间建立一个数据交换平台，包括数据服务层内部的数据流动都通过数据交换平台，ODS 相当于数据的集成平台，存储的都是实时性的数据，数据仓库存储的都是历史数据。

数据仓库可以分成基础区、汇总区和集市。

对于数据沙盘的使用，如果某个应用从源数据层通过数据交换平台到 ODS，到数据仓库层，再到数据集市层，可能数据的链路过长，影响应用的时效性，这样可以建一个数据沙盘，数据可以直接从 ODS 取数，或者从数据仓库、数据集市中取出数据，当稳定和固化后，再把应用挪到 ODS 或者数据仓库、数据集市中。数据沙盘也可以称为数据试验区。

对于大数据来说，可能本身会有应用，或者和结构化数据结合起来一起应用。

所有的数据流动都有统一的调度工具进行调度，同时建立对数据的分布和流转的数据管控，包括元数据管理、数据质量管理、数据标准管理、数据生命周期管理等内容。

- 解决电力行业数据仓库建设的难点问题有以下方法：可以通过试点建设积累经验，形成统一的数据模型标准、管控方法和流程，再大面积推广。
- 数据仓库的建设可以提高电力安全运营的能力、绩效分析的能力、电力营销管理的能力和决策分析的能力。具体表现是通过对电力设备的运行状况、检修情况和事故的及时掌握，提高电力安全运营的能力，通过对电量、电费、电价的分析，提高对电量的需求预测能力和价格制定能力，这样可以提高电力营销管理的能力。
- 电力行业数据仓库模型的建立过程：首先建设企业级的概念数据模型，然后在此基础上建设企业级逻辑数据模型，最后建设电力物理数据模型。

第 10 章　商业智能—ODS 数据架构和案例

本章目标

通过前几章的学习，我们了解了数据仓库的定义、数据仓库产生的背景、数据仓库的主要特征、数据仓库面临的挑战和技术特性。同时我们也了解了数据仓库的建设方法、数据仓库的架构规划，包括大数据环境下的数据仓库建设、数据仓库模型的设计、关于数据仓库系统的灾难备份规划，最后我们学习了关于商业银行的数据仓库建设和电力行业数据仓库的建设等相关内容。

学习本章后，读者将掌握：

- ODS 的定义
- ODS 的系统目标
- ODS 的业务目标
- 某商业银行 ODS 系统的数据架构规划
- 某商业银行 ODS 系统案例
- ODS 逻辑模型设计
- ODS 物理模型设计

10.1　ODS 概述

10.1.1　ODS 的定义

关于 ODS 的概念，在前几章已经进行了介绍，即 ODS 是面向主题的、集成的、可变的、并且反映当前细节性的数据集合，用于支持即时性的、操作性的全局信息的需求，它是数据仓库的过渡阶段。关于 ODS 有很多的解释和定义，最根本的就是 ODS 需要集成多个系统的数据，同时又要给一个或者多个系统使用。通常数据有较频繁的更新以及保存即时性的信息。

对于企业来说，ODS 系统可以解决很多的问题。例如，ODS 拥有最少的历史数据，而尽可能接近实时地监控企业目前的运转情况，提供企业内部或者外部的信息以支持决策分析，提供实时的全局信息以便于制定未来的发展战略。

ODS 的建设流程一般包括 4 个步骤：

1）对数据进行统一整合，构建全企业的数据标准化体系。

2）实现对应用系统的统一供数和数据分发。

3）实现数据架构和技术架构的统一，不断完善 ODS 系统的建设。

4）将 ODS 系统的数据转入到数据仓库中，以便对历史数据进行分析和挖掘。

通过 ODS 系统的建设，可以有效地缩短应用系统的实施路径，降低重复开发率，同时可以提高对数据需求的快速响应，为更深层次的挖掘分析奠定基础。

10.1.2 ODS 的系统目标和业务目标

（1）ODS 的系统目标

ODS 系统作为企业运营数据共享的平台，应该集成各个业务系统的数据，支持跨系统的数据应用，有效地提升数据的质量。因此，ODS 的系统目标包括以下几个：

- 数据共享

通过 ODS 系统为各个业务系统提供共享数据，降低接口的复杂度，提高系统接口的效率。

- 数据质量的校验和管控

通过 ODS 系统提高数据质量的校验能力和管控能力，包括提升数据的完整性、唯一性、一致性和及时性。对于校验能力，主要包括唯一性校验、一致性校验和主外键校验等内容。例如，在某 ODS 系统中，客户主题中的客户信息不允许重复，客户身份证号码字段可以作为客户唯一识别的标识。为了保证客户信息的正确性，需要在 ODS 系统中增加对客户基本信息表的唯一性校验。

- 数据整合的能力

通过 ODS 系统的建设，提升数据整合的能力，包括统一的数据模型、数据标准和数据视图等。

- 实时或者准实时地提供数据应用

通过 ODS 系统的建设，可以为用户提供固定报表应用、查询类应用、动态决策分析应用、风险监控类的应用等内容。

（2）ODS 的业务目标

ODS 系统是商业智能架构的重要组成部分，它可以实现跨系统的数据整合。ODS 系统的业务目标主要包括：

- 为客户提供统一的视图和展示。
- 为客户提供生产经营类的报表展示。
- 为客户提供关键绩效类的报表展示。
- 为客户提供经营风险类的报表展示。
- 为客户提供决策分析类的报表展示。

10.2 关于 ODS 系统的数据架构

10.2.1 某商业银行 ODS 系统的数据架构规划

关于某商业银行 ODS 系统数据架构规划的设计思路，主要包括以数据源作为驱动、统一管理和规范、完善共性加工层等几个方面的内容。

- 以数据源作为驱动

对数据源系统进行分析，按照模型贴近源系统的原则，确定源系统的增量层和标准增量层。

- 统一管理和规范

我们可以基于银行的数据统一标准，在源系统分析基础上，对标准增量层的数据进行整合，然后按照业务主题重新组织，形成基础数据层。

● 完善共性加工层

根据业务需求，整理共性加工层，以满足公共加工的要求。

某商业银行 ODS 系统的数据架构规划设计如下：

ODS 系统的架构设计可以分成几个层次：源数据增量层、标准增量层、基础数据层、共性加工层。每个应用系统都独立设计各自的数据集市。

其中源数据增量层和标准增量层与源系统结构类似，对数据进行标准化处理，以避免源系统的变化对基础数据层的影响。基础数据层按照业务主题进行整合，在设计过程中，考虑业务发展的需求，为分析类应用提供标准化的基础数据。共性加工层根据业务特点，结合实际应用，对一些指标进行统计分析，为集市提供统计数据。一般来说，共性加工层只进行简单的汇总计算，随着应用系统的不断扩充，可以整理出相关的共性指标。

最后在基础数据层和共性加工层的基础上为分行和总行的应用系统提供数据，或者为每个应用系统建设独立的数据集市。

基于以上思路，关于某商业银行 ODS 系统的数据架构规划如图 10-1 所示。

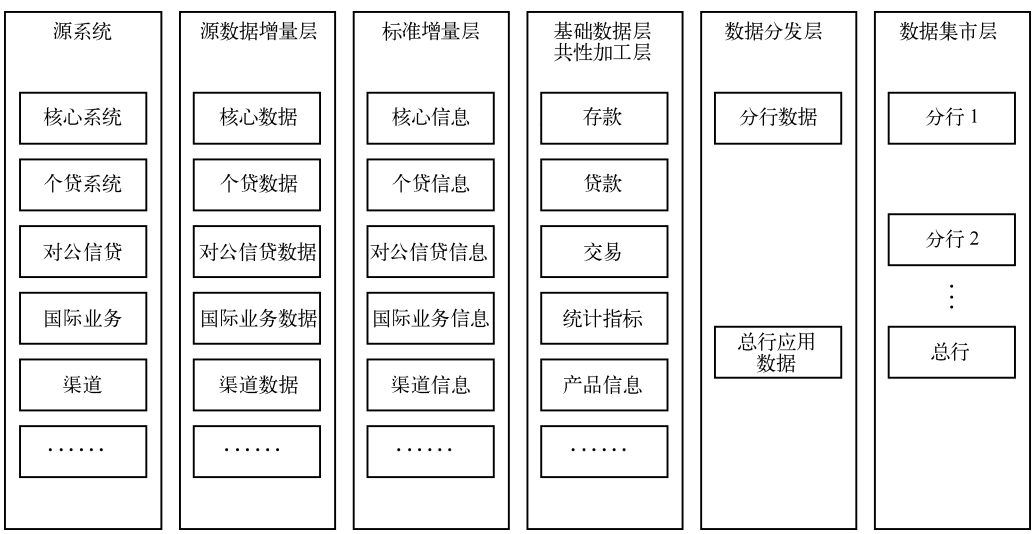


图 10-1　关于某商业银行 ODS 系统的数据架构规划

对各个层次的描述主要包括源数据增量层、标准增量层、基础数据层和共性加工层。

● 源数据增量层

源数据增量层存放各个业务系统的增量文件，可以由 ETL 工具进行增量抽取。源数据增量层可以将数据保存一周左右。

● 标准增量层

标准增量层是介于源数据增量层和基础数据层之间的模型，它的数据结构是贴源的，是经过清洗和标准化后的数据。

● 基础数据层

基础数据层是 ODS 系统的核心，对业务数据进行轻度的整合，该模型贴近源系统，同

时保证数据的标准化。该层需要保留必要的历史数据，可能是几个月，也可能是若干年。

- 共性加工层

共性加工层是 ODS 系统的重要组成部分之一，目的是提高数据查询的效率，对查询请求频率较高的数据做进一步的整合。方便对共性基础指标进行统计分析，该层只包含基本的汇总数据。共性加工层将共性指标提炼出来，减少系统的重复处理。

10.2.2 某商业银行 ODS 系统案例

下面介绍某商业银行 ODS 系统建设的案例。在 ODS 系统未建之前，如图 10-2 所示，这种复杂的网状结构会带来一系列的问题，可能会造成信息孤岛，数据的可共享性降低，缺乏完整的数据解决方案。

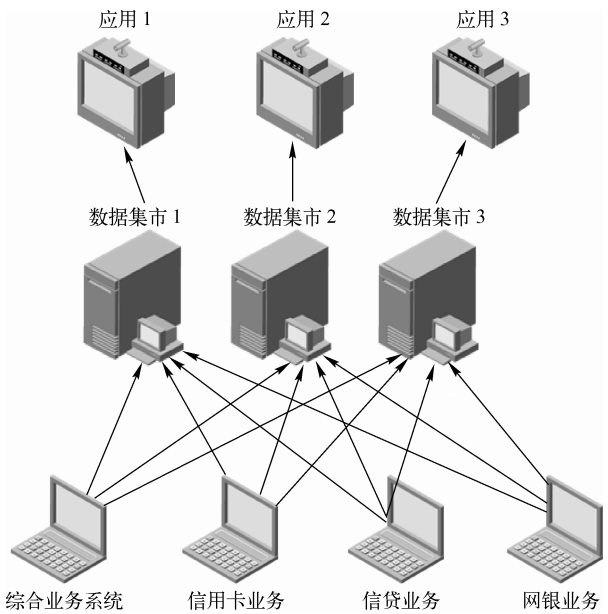


图 10-2 ODS 系统未建之前

按照此种思路，ODS 系统未建之前，系统的复杂度是 M × N，如图 10-3 所示。

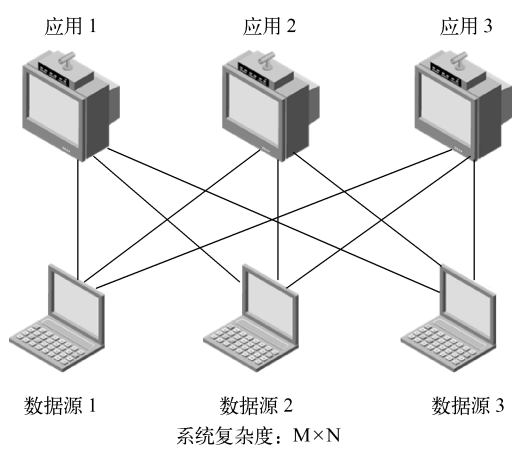


图 10-3 ODS 系统未建之前的系统复杂度

ODS 建成之后，作为一个中间的层次，它包含全局一致的、细节的、当前的数据。经过 ODS 系统的初步集成和标准化加工，对具有共性的数据加工需求进行抽象，以供后续加工使用。数据仓库的数据来自于 ODS 系统，ODS 系统的数据经过转换后，根据需要可以移入数据仓库中，如图 10-4 所示。

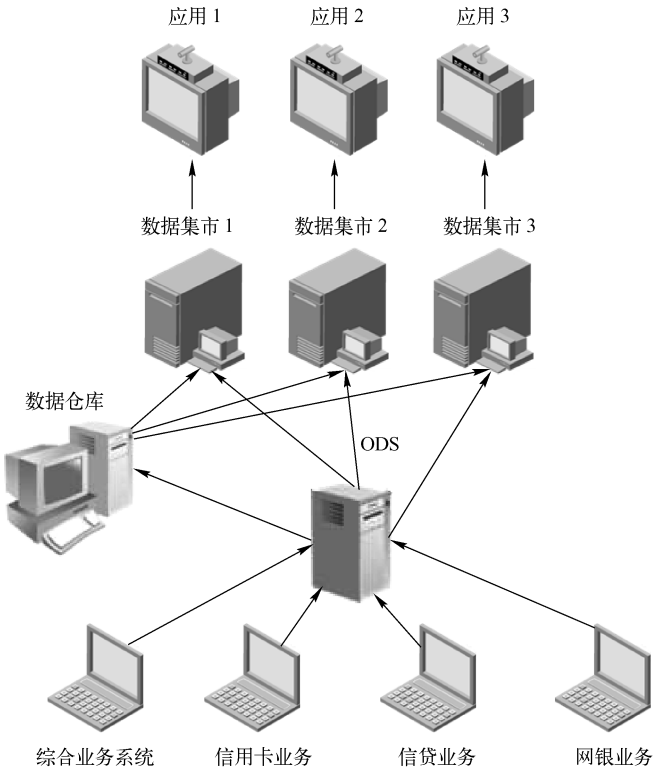


图 10-4　ODS 系统建成之后

ODS 建成之后，系统的复杂度是 M + N，如图 10-5 所示。

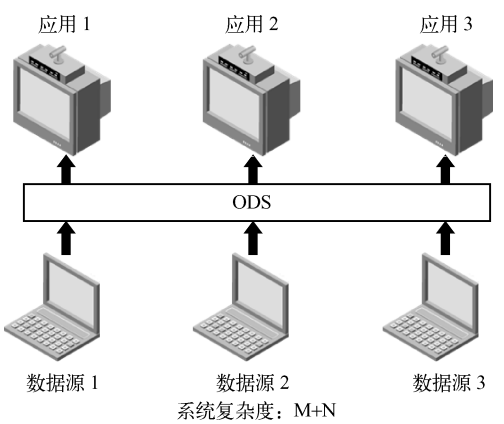


图 10-5　ODS 建成之后的系统复杂度

综上所述，ODS 系统在业务系统数据源和各个应用之间形成一个缓冲带，它可以对各

个业务数据进行标准化、规范化，进行数据质量管理，最后实现全企业的统一数据视图。同时支撑跨系统的数据应用，提供数据共享，满足银行在业务经营和精细化管理方面对高质量和高时效的需求。

10.3 ODS 模型设计

关于 ODS 的模型设计，主要包括数据调研、确定数据范围和主题定义。数据调研是对现有业务系统的逻辑模型和物理模型进行了解。确定数据范围是在业务系统调研的基础上进行的，目的是确保应用所需的数据都已经从业务系统中抽取出来了。主题的定义是以业务系统为基础，参考业务系统的企业模型来定义数据主题，主要以 ER 模型为主。

10.3.1 ODS 逻辑模型设计

关于 ODS 逻辑模型的设计过程，主要包括：逻辑结构定义、存储周期定义和存储粒度定义。

逻辑结构定义主要包括：定义各个实体的概念特性、实体和实体之间的关系等。

存储周期定义主要是指数据在 ODS 中的存储期限。例如，有些数据在 ODS 中保存一段时间后再加载到数据仓库即删除，有一些数据可能会在 ODS 中长期保存。

存储粒度定义是指数据在 ODS 中存储的细节程度。粒度层次的划分决定了 ODS 中的数据量和查询的灵活度。

关于 ODS 逻辑模型的设计步骤，如图 10-6 所示，主要包括：定义数据范围、主题定义、形成逻辑模型说明书。

图 10-6　ODS 逻辑模型的设计步骤

（1）定义数据范围

确定数据范围是在对业务系统调研的基础上进行的，确保应用所需的数据都已经从业务系统中抽取出来了。一般来说，设计人员需要综合业务系统的企业模型，得到全企业范围内的数据视图，通过抽象划分逻辑模型的数据主题范围。

（2）主题定义

通过数据主题的分解和重构，进行主题的定义，包括定义实体、实体之间的关系，对应的存储粒度、存储期限等。在 ODS 中，通过对实体的归并，保证实体之间的一致性和唯一性。

（3）形成逻辑模型说明书

需要在 ODS 逻辑模型说明书中对数据范围、主题定义、实体和实体之间的关系进行详细地描述。在 ODS 逻辑结构说明书中需要对数据范围、主题定义、实体和相关属性的定义

进行精确、详尽地描述。同时需要详细说明数据的存储周期、存储方式等。ODS 逻辑模型需要解决数据的粒度层次划分，关于粒度层次的划分直接决定了 ODS 的数据量和查询的灵活性。一般来说，ODS 中的数据是从生产业务系统中取出的细节性数据，数据粒度与业务源系统保持一致。

10.3.2 ODS 物理模型设计

ODS 物理模型设计是对数据的索引策略、数据存放位置和数据的存储分配进行定义。物理模型设计人员需要了解数据的使用频率、数据规模以及响应时间要求等。同时理解外部存储设备的特性，如分块原则、设备的 I/O 特性等内容。

其中数据的索引策略是为了提高数据的存取效率。特别是在数据仓库中，设计人员应该考虑为数据存储建立专用或者多样的索引，因为数据仓库中的数据是不经常更新的，数据存储相对稳定。

数据存放位置主要考虑将不同类别的数据存放到不同的存储设备中。例如，一些重要程度高、对响应时间要求较高的数据应该存放在高速存储设备上，如硬盘；对一些存取频率较低和响应时间要求不高的数据应该放在低速存储设备上，如磁带和磁盘中。

数据的存储分配主要是确定块的大小、缓冲区的大小和个数等内容。通过对存储分配的参数指定，实现数据的物理优化。

小结

- 对于企业来说，ODS 系统可以解决很多问题。例如，ODS 拥有较少的历史数据，而尽可能接近实时地监控企业目前的运转情况，提供企业内部或者外部的信息以支持决策分析，提供实时的全局信息以便于制定未来的发展战略。
- ODS 系统作为企业运营数据共享的平台，应该集成各个业务系统的数据，支持跨系统的数据应用，有效地提升数据的质量。
- ODS 系统是商业智能架构的重要组成部分之一，它可以实现跨系统的数据整合。
- ODS 系统的架构设计可以分成几个层次：源数据增量层、标准增量层、基础数据层、共性加工层。每个应用系统都独立设计各自的数据集市。
- 源数据增量层存放各个业务系统的增量文件，可以由 ETL 工具进行增量抽取。源数据增量层可以将数据保存一周左右。
- 标准增量层是介于源数据增量层和基础数据层之间的模型，它的数据结构是贴源的，是经过清洗和标准化后的数据。
- 基础数据层是 ODS 系统的核心，对业务数据进行轻度的整合，该模型贴近源系统，同时保证数据的标准化。该层需要保留必要的历史数据，可能是几个月，也可能是若干年。
- 共性加工层是 ODS 系统的重要组成部分之一，目的是提高数据查询的效率，对查询请求频率较高的数据做进一步的整合，方便对共性基础指标进行统计分析，该层只包含基本的汇总数据。共性加工层将共性指标提炼出来，减少系统的重复处理。
- 关于 ODS 逻辑模型的设计过程，主要包括逻辑结构定义、存储周期定义和存储粒度定义。

逻辑结构定义主要包括定义各个实体的概念特性、实体和实体之间的关系等。

存储周期定义主要是指数据在 ODS 中的存储期限。例如，有些数据在 ODS 中保存一段时间后再加载到数据仓库即删除，有一些数据可能会在 ODS 中长期保存。

存储粒度定义是指数据在 ODS 中存储的细节程度。关于粒度层次的划分决定了 ODS 中的数据量和查询的灵活度。

- ODS 物理模型设计是对数据的索引策略、数据存放位置和数据的存储分配进行定义。物理模型设计人员需要了解数据的使用频率、数据规模以及响应时间要求等。同时理解外部存储设备的特性，如分块原则、设备的 I/O 特性等内容。

第 11 章　商业智能—数据集市架构和案例

本章目标

通过前几章的学习，我们已经掌握了商业智能的几个基本组成部分，包括数据仓库的定义、数据仓库产生的背景、数据仓库的主要特征、数据仓库面临的挑战和技术特性，ODS 的定义、ODS 的系统目标和业务目标、关于某商业银行 ODS 系统的数据架构规划、某商业银行 ODS 系统案例、ODS 逻辑模型设计和 ODS 物理模型设计等内容。下面我们主要讲解关于数据集市的架构和案例。

通过本章的学习，读者将掌握：

- 数据集市的概念
- 关于数据集市的误区
- 关于数据集市的主要应用
- 数据集市概念模型设计
- 数据集市逻辑模型设计
- 数据集市物理模型设计
- 数据集市的架构模式
- 某商业银行的数据集市架构解决方案

11.1　数据集市概述

11.1.1　数据集市概念

数据集市的概念在前面已经做了定义。简单地说，数据集市是一种较小的和集中的数据仓库。业务系统的数据经过数据仓库流入到不同的部门，而这些部门级的数据仓库就称为数据集市。一般来说，每个部门都有各自的数据集市，它们之间可能相互关联，但本质上是相互独立的。数据仓库主要面向整个企业，而数据集市则面向各个部门。数据仓库的粒度相对较小，而数据集市的粒度一般是概括汇总级的。

11.1.2　关于数据集市的误区

我们分析一下关于数据集市的理解有哪些误区？如图 11-11 所示。

（1）数据量大小是区分数据集市和数据仓库的主要特征

数据量的大小不能作为区分数据集市和数据仓库的主要特征，因为有可能某个生产厂商数据仓库的数据量远远小于电信行业某个部门数据集市的数据量。

（2）数据集市是容易建立起来的

数据集市在很大程度上比数据仓库的复杂性略低一些，因为它只针对某一特定主题。但是因为数据集市可能会从多个数据源中提取数据，围绕数据的复杂问题会很高，因此数据集市不会很容易建立起来。

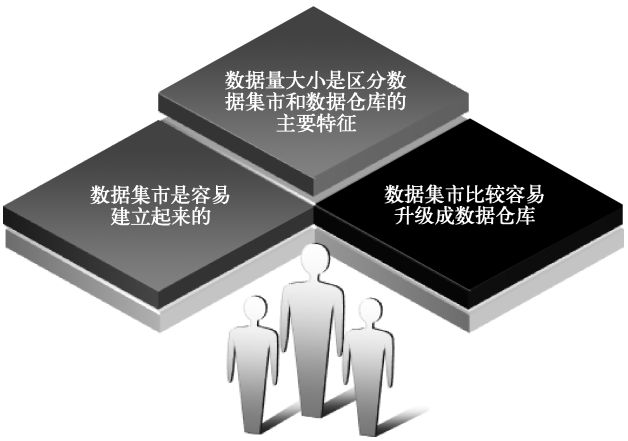


图 11-1　关于数据集市的理解有哪些误区

（3）数据集市比较容易升级成数据仓库

数据集市主要针对特定的业务需求，采用特殊的模型。当扩展和追加数据的时候，会增加信息孤岛，不能以企业全局的视角分析数据，因此数据集市升级到数据仓库会很困难。

11.1.3　关于数据集市的主要应用

关于数据集市的主要应用，包括监控预警、客户群分析、即席查询和自助报表。

（1）监控预警

数据集市的监控预警功能主要实现指标类、业务类相关数据的监控预警。

（2）客户群分析

数据集市的客户群分析是针对业务部门和客服部门的营销需求，对客户信息进行详细分析，为营销提供支撑。可以针对区域（如省、市、区、县、家庭、学校等）、客户属性（如职业、消费习惯等）进行客户群的细分。

（3）即席查询

数据集市的即席查询是基于数据集市业务逻辑视图，面向业务人员的查询工具，提供各种查询生成器的功能。

（4）自助报表

数据集市的自助报表一般是面向企业管理人员和业务人员使用的，可以提供各种报表预览和发布功能。可以提高业务部门、管理部门报表需求的响应速度。

11.2　数据集市模型设计

数据集市建模时通常采用“自顶向下”的方法，建模过程可以分成以下三个阶段：数

据集市概念模型设计、数据集市逻辑模型设计和数据集市物理模型设计。

1. 数据集市概念模型设计

数据集市概念模型设计是通过需求分析，明确需求涵盖的业务范围，然后对需求范围内的业务和业务之间的关系进行概括性的描述，通过对业务对象的归类，划分主题域。概念模型的设计是为逻辑模型设计做准备的。

2. 数据集市逻辑模型设计

数据集市逻辑模型设计是通过对概念模型的各个主题域进行细化，同时根据业务定义、分类和规则，定义实体并描述实体之间的关系，在实体关系的基础上明确各个实体的属性。实体间的对应、约束关系则来自于各业务过程中的规则，最后定义相应的事实表和维度表，组成星形逻辑模型。

3. 数据集市物理模型设计

数据集市物理模型的设计依赖于逻辑模型的完成，目的是提高数据分析的效率，针对具体的分析需求采取相应的优化策略。数据集市的主题分为两种类型：综合类主题和专业类主题。综合类主题是从整个企业的关键指标进行综合分析。专业类主题是从业务部门关心的指标进行分析。

数据集市的数据分为两种：一种是基于数据仓库的细节数据或者汇总数据进行统计分析，另一种是基于数据挖掘进行分析。

11.3 数据集市的架构模式

数据集市的架构模式主要分成库内数据集市和库外数据集市。

库内数据集市是部署在企业级数据仓库之内的，在数据仓库的汇总数据层和基础数据层基础上构建面向特定主题的数据集市。库内数据集市可以共享汇总数据层和基础数据层的数据，如图 11-2 所示。

库外数据集市是根据应用需求而形成的数据集合。库外集市一般是在数据仓库之外进行部署的，它具有专门的软硬件设备。库外数据集市的来源是数据仓库基础数据层和汇总数据层的数据，如图 11-3 所示。

图 11-2　库内数据集市

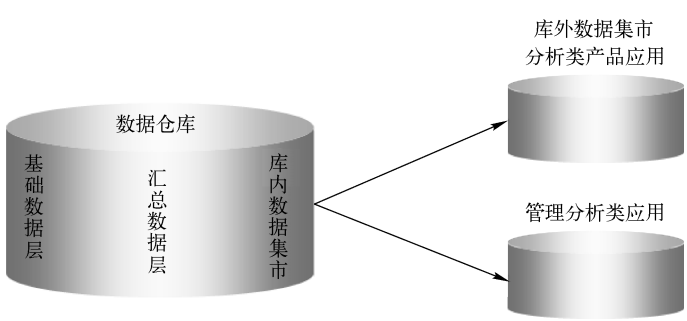


图 11-3　库外数据集市

11.4 某商业银行的数据集市架构解决方案

例如，某商业银行关于数据仓库建设已经初具规模，随着历史数据的累积，数据仓库可以满足各类分析需求，按照该银行的长期规划，数据集市的建设逐渐提上日程。它可以降低成本，提升效率，提高整体架构的安全性。

按照该银行的数据架构，数据集市的建设采用“自顶向下”的建设思路，即首先建设全行统一的数据仓库。数据仓库的数据来源于各类业务系统及外部数据，对全行数据进行整合，做到数据的完整、统一；再从业务层面，基于数据仓库建设各类应用的数据集市，数据集市的数据来源于数据仓库，避免重复的数据整合和转换工作，满足各类分析应用的需求，如图 11-4 所示。

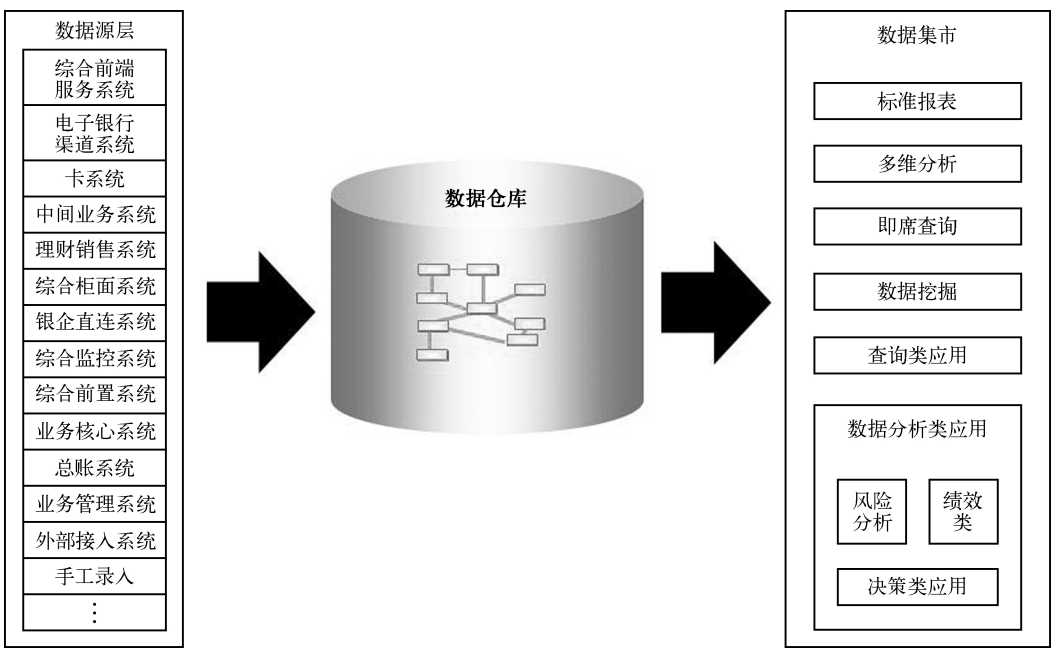


图 11-4　某商业银行的数据集市架构解决方案

小结

- 数据集市是一种较小的和集中的数据仓库。业务系统的数据经过数据仓库流入到不同的部门，而这些部门级的数据仓库就称为数据集市。一般来说，每个部门都有各自的数据集市，它们之间可能相互关联，但本质上是相互独立的。
- 关于数据集市的主要应用，包括监控预警、客户群分析、即席查询和自助报表。
- 数据集市建模时通常采用“自顶向下”的方法，建模过程可以分成以下三个阶段：数据集市概念模型设计、数据集市逻辑模型设计和数据集市物理模型设计。

- 数据集市概念模型设计是通过需求分析，明确需求涵盖的业务范围，然后对需求范围内的业务和业务之间的关系进行概括性的描述，通过对业务对象的归类，划分主题域。概念模型的设计是为逻辑模型设计做准备的。
- 数据集市逻辑模型设计是通过对概念模型的各个主题域进行细化，同时根据业务定义、分类和规则，定义实体并描述实体之间的关系，在实体关系的基础上明确各个实体的属性。
- 数据集市物理模型的设计依赖于逻辑模型的完成，目的是提高数据分析的效率，针对具体的分析需求采取相应的优化策略。数据集市的主题分为两种类型：综合类主题和专业类主题。综合类主题是从整个企业的关键指标进行综合分析。专业类主题是从业务部门关心的指标进行分析。

第 12 章　金融行业数据架构案例和商业智能

本章目标

通过前几章的学习，我们已经对企业总体规划、数据架构和商业智能有一个整体性的认识。本章将重点介绍金融行业数据架构的相关案例和商业智能内容。

学习本章后，读者将掌握：

- 金融行业背景概述
- 金融行业的数据架构
- 传统金融行业某系统的数据架构案例
- 互联网金融行业的数据架构案例
- 金融行业商业智能的背景和作用
- 金融行业如何实施商业智能
- 金融行业的业务流程和运营模式优化

12.1　金融行业背景

首先我们了解一下什么是金融。

我们可以简单地对金融进行定义：金融就是在我们的经济生活中，通过银行、证券机构等中介，从市场主体中募集资金，然后在借贷给其他市场主体的活动，可以把金融看做融资、投资和资金募集等三种经济活动，如图 12-1 所示。

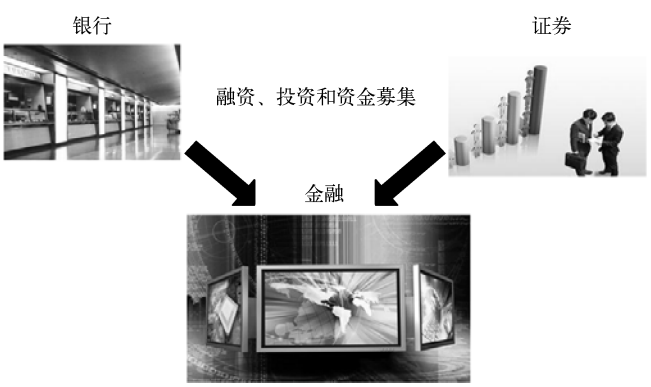


图 12-1　金融定义

对于商业银行来说，它有大量的客户群，可以吸收社会公众存款，资金实力非常雄厚，抗风险的能力比较强。同时银行有大量的客户信用数据，包括客户信用卡消费信息、贷款信息、还款信息和信用信息等。

随着互联网技术的进步，商业银行通过互联网融资会更有利，因为商业银行本身具有良

好的信用基础和声誉，各种贷款、股票和债券都可以通过互联网进行交易。同时也可以利用互联网技术解决信息不对称的问题。对于银行来说，借贷业务仍然是商业银行的核心业务，它的净利息收入占到 70% 左右。

目前一些互联网企业也在发展金融业，因为它拥有数量庞大的客户群，通过数据挖掘，分析客户的真实需求，然后提供所需的产品和服务。对于互联网金融来说，它没有改变商业的实质，而是仅仅改变了销售与服务的渠道，改善了客户的体验。例如，贷款、股票和各种债券可以通过互联网金融进行交易，它与传统的通过银行作为中介的交易方式不同，它既保证了资金按照供需双方直接交易，同时又不同于资本市场直接融资的另外一种融资模式。

我们可以把互联网和移动互联网统一称为互联网金融，如图 12–2 所示。它可以包括传统的商业银行、证券公司等金融机构的互联网化，通过互联网为客户提供各种金融服务。但是随之带来一些问题，例如，互联网企业发展金融业是否符合金融行业监管要求，是否能够承受各种风险，这是互联网企业目前面临的挑战和困难。

对于一些电商网站来说，它可以根据商品的点击频率以及商品与商品之间的关系，计算出用户感兴趣的商品的概率，然后在网页上进行直接推送，这种方式大大增加了购买成功的概率，也降低了广告宣传的成本。

在一些网银界面上，只有一些固定的营销广告，还没有真正地实现以客户为中心的交叉营销。实际上，我们完全可以根据客户大量的信息，如个人的资产情况和理财习惯，向客户推送个性化的产品和服务。一些电商企业其实也看准了这个方向，它们利用互联网平台，依靠用户的交易数据和信用数据，开展互联网上的融资业务。在这个过程中，借贷双方都避开了银行等金融中介，这就是所谓的金融脱媒现象，如图 12–3 所示。

图 12–2　互联网金融

图 12–3　金融脱媒

随着金融脱媒现象越来越凸显，对商业银行也提出了很高的要求，虽然商业银行积累了大量的客户信息、交易信息，但是在数据挖掘方面还有很大的提升空间。

例如，在一些个人网银页面，没有统一的界面可以一目了然地看到自己的负债情况，必须进入到不同的账户中查询余额。表面上是页面的问题，实质上是目前商业银行还是“以账户为中心”，没有真正做到“以客户为中心”，最理想的状态是让客户能够看到自身整个资产负债的情况，然后通过一步步钻取，看到每个账户的全貌和明细。所以说，银行的服务质量还有很大的提升空间。

根据以上的金融行业背景，对商业银行提出了更高的要求。面对这些要求，商业银行应该具备哪些能力呢？如图 12–4 所示，应该具备对客户的洞察力、精准营销和跨渠道客户管理的能力。

图 12-4　商业银行应该具备的能力

（1）对客户的洞察力

因为缺乏全企业统一客户视图以及有效利用这个视图的能力，很多银行一直都难以了解客户需求。商业银行可以利用数据仓库，通过数据分析和建模来了解银行客户需求。

（2）精准营销

商业银行可以通过数据仓库来分析客户，通过闭环营销，帮助银行利用每一次的互动来增强对客户的了解。

（3）跨渠道客户管理

客户通过各种渠道与商业银行以及其他金融机构进行互动。对于金融机构来说，需要考虑如何使用多渠道战略吸引客户，并且通过跨渠道战略去管理与客户的互动，从而丰富数据来源，获得更加深入的分析数据。

那么为了满足这些能力要求，商业银行应该具备什么样的数据架构呢？下面就来理解一下金融行业的数据架构。

12.2　金融行业的数据架构

金融行业的数据是推动商业银行等金融机构变革的主要推动力。目前来说，商业银行之间的竞争越来越激烈。商业银行的发展需要良性的差异化竞争，数据是竞争的基础条件。很多金融机构通过数据分析指导日常运营，为客户提供更好的服务和产品，同时降低商业银行运营的风险，获取竞争的优势。

金融机构每天都在产生大量的数据，包括各种文本、视频、图片、日志、音频和地理位置等信息，但是这些数据之间还存在着很多问题，如数据存在分割、标准不统一、难以共享等问题。

上述这些问题导致出现了大量的信息孤岛，从而难以利用这些宝贵的数据做出有效的决策分析。很早以前，商业银行的数据架构都是以统计报表为主。在信息化的建设过程中，各个系统之间的数据定义、数据采集流程缺少体系建设，信息之间难以共享。同一数据可能在

多个系统中重复录入和存储。优秀的数据架构在金融行业中显得尤为重要。

在所有行业中，银行的数据管理其实是比较困难的。

1）商业银行的 IT 系统建设较早，随着时间的流逝，系统变得越来越复杂。

2）商业银行对于数据的准确性要求是极高的，但很多银行的数据并没有统一的标准，不同系统之间的数据还存在不一致和不完整的现象。

3）关于商业银行的数据架构、数据治理和管控是非常重要的。

金融行业的数据架构一般包括以下几个部分：数据采集层、产品加工层和对外服务层，如图 12-5 所示。

图 12-5　金融行业的数据架构

从数据源开始，经过加载、集中、整合，以及对外服务这几个过程，可以将整个数据架构横向划分成：源数据区、基础区、产品加工区和产品服务区，如图 12-6 所示。各个区域都相对独立。

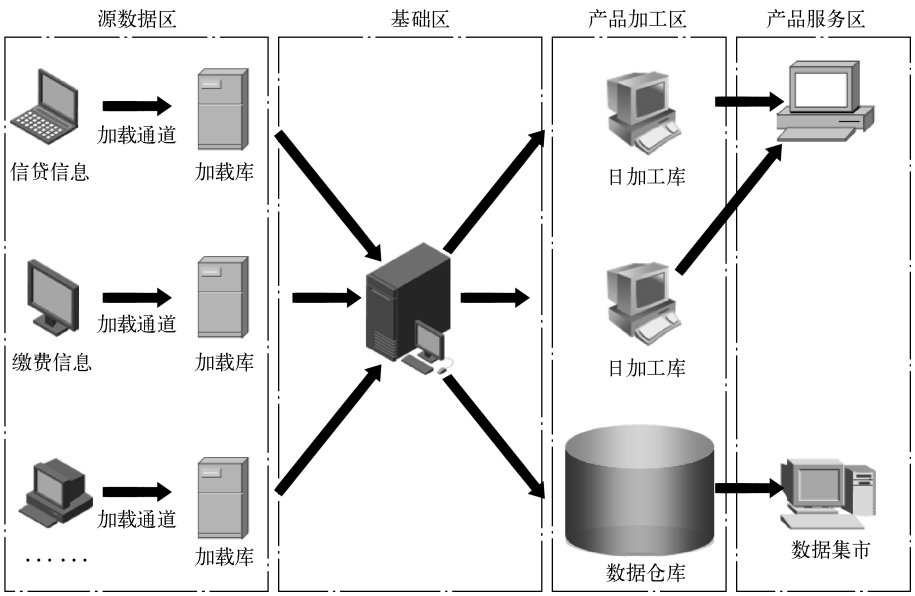


图 12-6　数据架构的横向划分

● 源数据区

在源数据区中，可以进行并行处理，设计多个加载通道，提高加载的并行度和加载效率。在数据加载入库之后，再进行逻辑校验，包括对错误数据的反馈，然后使用快速迁移技术，将数据迁移到基础区中，最后将加载库的数据清空，以备下一阶段的数据加载。

● 基础区

基础区中存储的是数据采集的信息，以满足对新增数据的采集、加载和整合，最后为产品加工做准备。

● 产品加工区

产品加工区主要面向应用，包括对数据类、解决方案类和服务类等产品的加工。产品加工区可以分成数据集中区和加工单元区。

（1）数据集中区

产品加工区的数据都来源于基础区，一般来说，产品加工的时间较长，为保证产品加工和数据加载都有相对独立的时间窗口，在产品加工区划出一个数据集中区，作为缓冲层，如图 12-7 所示。

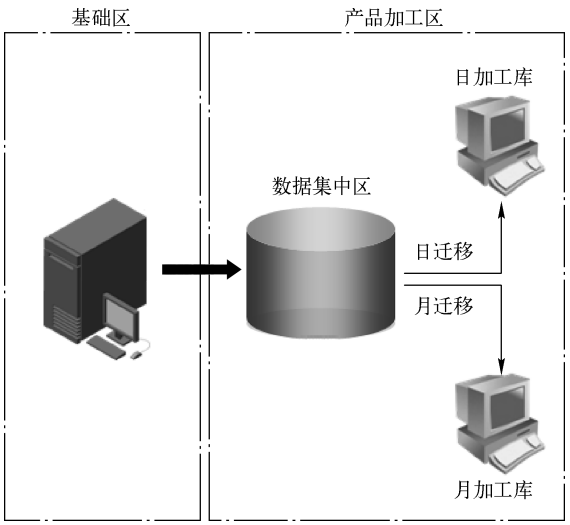


图 12-7　数据集中区

根据产品加工的需求，定期从基础区中抽取产品加工需要的数据，在产品加工之前建立一个数据集中区，目的是降低基础区和产品加工区之间的耦合性，同时根据加工频率的不同，将数据集中区分成日迁移和月迁移的数据。

对于日加工的数据，每天都需要根据数据加载量，完成当日的加载任务，同时为了避免不同产品之间加工过程的相互影响，可以为每类日加工建立相对独立的数据库实例。

对于月加工的数据，每月迁移一次数据，为了保证不同种类的产品数据加工的一致性。

在采用批量数据迁移的同时，考虑在数据迁移的时候暂停数据加工服务，尽量避免数据迁移和加工同时进行。

（2）加工单元区

在加工单元区中，提供各类产品加工的原子数据，如图 12-8 所示。

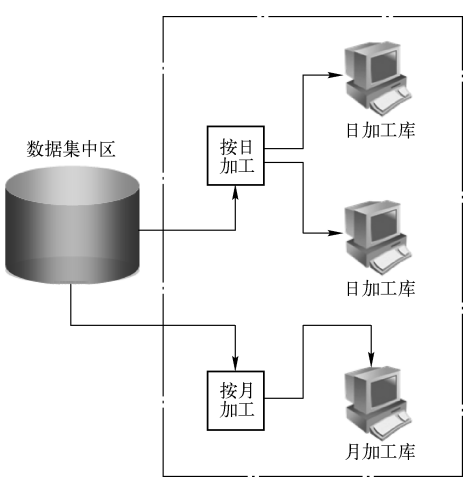


图 12-8　加工单元区

对于加工单元区来说，应该满足产品加工时

高性能的数据处理要求，从整体上来说，产品加工流程是批量的，并且利用并行处理技术，实现不同产品的加工需求。

● 产品服务区

产品服务区主要提供对外服务，存储数据类与工具类产品的数据，以及各种产品查询记录，如图 12-9 所示。

图 12-9　产品服务区

产品从整体上可以分为离线查询产品和实时查询产品，两类产品分别采用不同的数据组织形式，对于实时查询产品，可以快速地反馈查询结果。

产品服务区的数据按照产品更新频度又分为日更新和月更新两种类型，更新频度不同的产品提供服务的时间范围不同，日迁移产品可以提供全天的服务。各种产品查询的记录统一存储在产品服务区中，根据产品加工的需求，定期将查询记录迁移至产品加工区。

产品服务区的建设方案：

首先应该建设数据采集平台，将采集到的数据加载到基础库中，实现数据处理的批量化，可以利用多加载通道实现并行加载的功能。

其次，利用数据仓库技术进行多维分析和挖掘。一般来说，数据仓库包括数据获取层、数据存储层和前端应用层，如图 12-10 所示。

● 数据获取层

数据获取层把基础层相关的数据经过抽取、转换和清洗，按照统一的模式和不同的主题进行集成，装载到数据仓库中。

● 数据存储层

数据存储层主要包括数据仓库和数据集市。数据仓库整合系统全局的共享信息，它包含历史数据信息，记录了过去某一时间点到目前各个阶段的信息，通过这些信息，可以对企业的发展状况和未来趋势做出分析和预测。数据集市是为了特定的目的和范围，从数据仓库中独立出来的一部分数据。

● 前端应用层

前端应用层包括统计报表和数据挖掘，为用户访问数据仓库提供了手段。同时也预留了专业统计分析软件的接口。

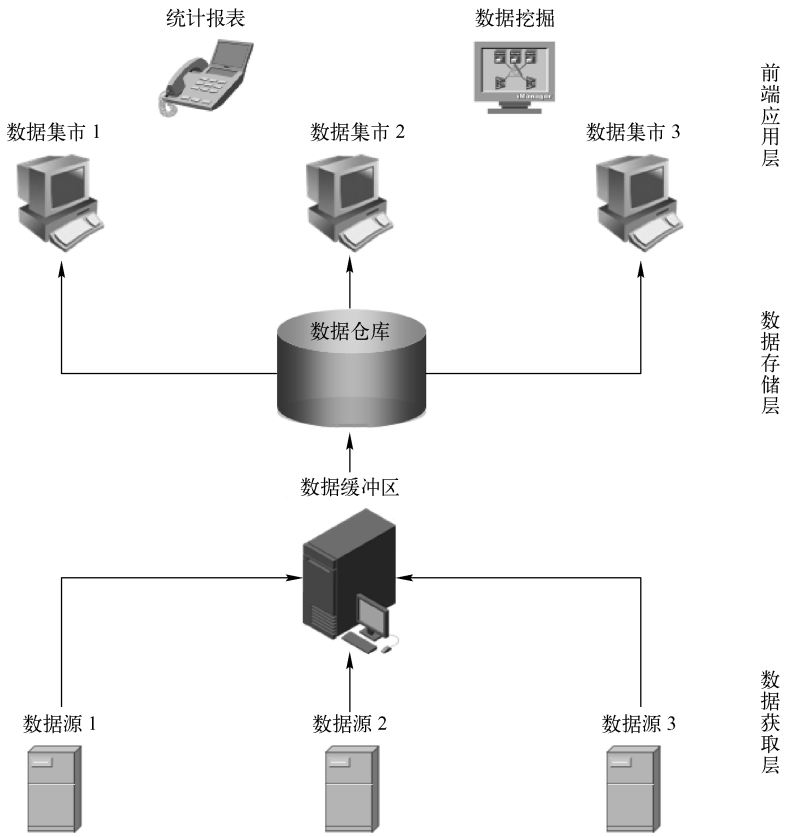


图 12–10　数据获取层、数据存储层和前端应用层

金融行业数据架构的特点，如图 12–11 所示。

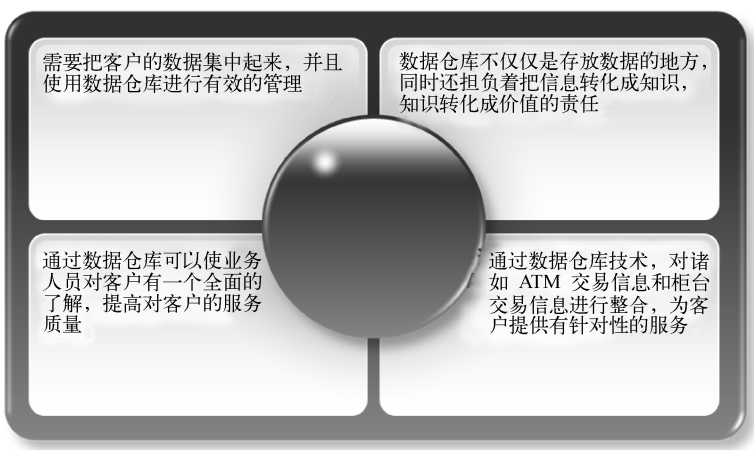


图 12–11　金融行业数据架构的特点

1）对于商业银行来说，为了全面了解客户的需求，并且提高服务的质量，需要把客户的数据集中起来，并且使用数据仓库进行有效的管理。

2）一般来说，银行使用数据库技术支持各种交易业务。数据仓库是把企业内部的分散

的数据库进行集成，形成统一的存储体系，相对于数据库来说，数据仓库不仅仅是存放数据的地方，同时还担负着把信息转化成知识，知识转化成价值的责任。

3）利用数据仓库技术可以为银行带来很多利益。例如，通过数据仓库可以使业务人员对客户有一个全面的了解，提高对客户的服务质量。

4）通过数据仓库技术，对诸如 ATM 交易信息和柜台交易信息进行整合，为客户提供有针对性的服务。

12.3 金融行业某系统的数据架构案例

12.3.1 传统金融行业某系统的数据架构案例

数据架构是企业架构的重要组成部分，帮助金融行业有效地分配、部署和使用数据，实现数据的合理组织和有效共享，从而保证数据在各个系统之间的一致性、完整性和有效性。

我们可以把传统金融行业某系统的数据架构分成以下几个部分：源数据层、内容管理、数据交换层、数据基础层、数据加工层和应用层，如图 12-12 所示。其中，源数据层提供产品加工和对外服务的所有数据。内容管理主要提供对非结构化数据存储、访问和管理的能力。数据交换层担负着系统内部各个数据库之间的数据交换任务。数据基础层进行格式校验及逻辑校验，形成唯一可信的数据源。数据加工层的数据来源为数据基础层，并将加工处理的数据提供给应用层。应用层可以包括查询类应用和分析类产品应用。

图 12-12 传统金融行业某系统的数据架构

1. 源数据层

源数据层提供产品加工和对外服务的所有数据。源数据层应该满足灵活和自动化的要求。它的特点主要包括以下几个方面：

1）需要描述源数据层采集哪些数据、数据源的类型和采集方式等内容。例如，数据源可以包括 Excel、数据库和通过网络爬虫得到的数据等。

2）需要描述数据源的内容格式，如结构化数据和非结构化数据。

3）需要描述数据源的频率特征。

举例来说，源数据层的主要特点见表 12-1。

表 12-1　源数据层的主要特点

数 据 来 源	采 集 内 容	数 据 格 式	数据采集方式
政府部门	行政处罚信息和奖励信息	结构化数据	接口方式
互联网	互联网信息	非结构化数据	网络爬虫方式
商业银行	客户身份信息、职业信息、居住信息、联络信息、客户概况信息等内容	结构化数据	接口方式
手工录入的数据	手工录入的信息	结构化数据	中间件方式

- 数据来源可以包括政府部门、互联网、商业银行和手工录入的数据等。
- 数据源的格式包括结构化数据、半结构化数据和非结构化数据。
- 数据采集方式包括接口方式、非接口方式、网络爬虫方式和 FTP 方式等。

其中对于接口方式，它主要是保证数据源端的数据质量，但是对于开发、调试、测试和技术方面的要求较高。对于非接口方式，特点是前期投入较少，对于技术方面要求不高，但是数据质量不能保证，对于人工的依赖较强。对于网络爬虫的采集方式，是从公网上获取非结构化数据，但收集的数据量较大，而单个数据的价值很低，投入的人力和技术也很大。对于 FTP 方式，是指通过大批量非结构化数据的上传进行采集，但是数据安全度较低，比较适合非结构化数据的上传。

2. 内容管理

除了从相关机构采集结构化的数据外，还可以从互联网或者其他渠道采集各种非结构化的数据。采集的非结构化数据包括：互联网信息、社交网络信息或者其他渠道提供的非结构化数据，如图 12-13 所示。

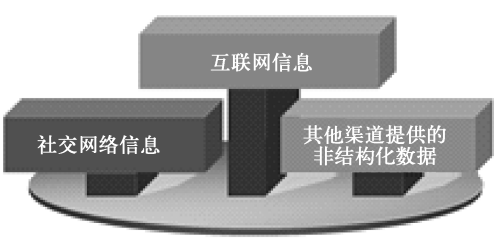


图 12-13　内容管理

一般来说，我们可以通过网络爬虫等技术收集各种非结构化数据，通过内容管理存储非结构化数据，建立非结构化数据的元数据信息，这些元数据信息可以存储在 Hadoop 平台中。其中非结构化元数据可能包括信息标签、摘要、索引和日志等。然后，在此基础上，与结构化数据进行关联，以供分析使用。这种方式实现了非结构化数据与结构化数据的整合，以供后续加工和使用，如图 12-14 所示。

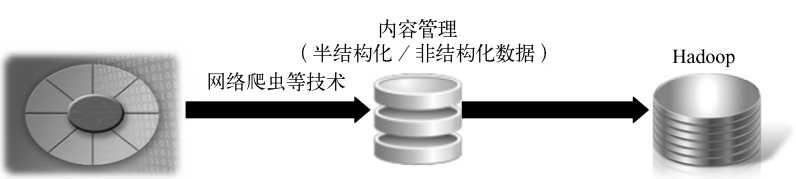


图 12-14　非结构化数据与结构化数据的整合

3. 数据交换层

（1）数据交换层的任务和功能

数据交换层承担着数据库之间的数据交换任务，同时也承担着外部文件和数据库之间的交换任务。数据交换层中的内部交换如图 12-15 所示。

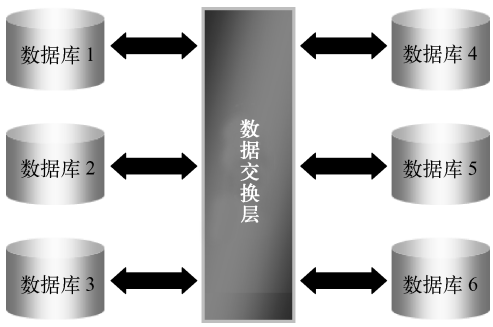


图 12-15　数据交换层中的内部交换

数据交换层中的外部交换如图 12-16 所示。

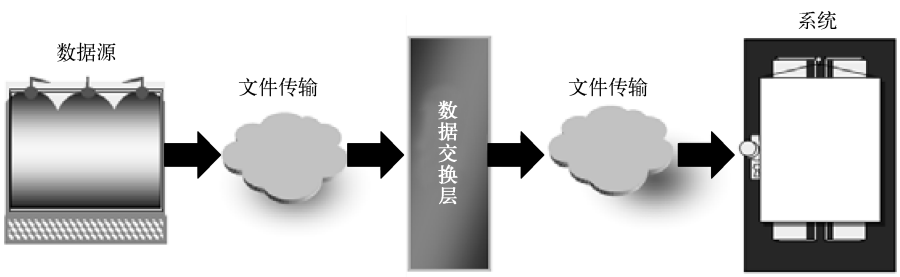


图 12-16　数据交换层中的外部交换

数据交换层具备数据抽取、质量检查、数据转换、数据加载四大功能，如图 12-17 所示。

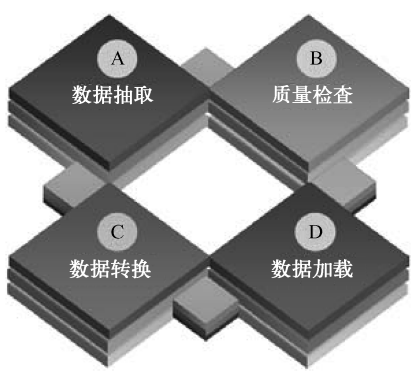


图 12-17　数据交换层功能

1）数据抽取。数据抽取是从源数据层获取数据，它可以实时或者定期地获取增量数据，通过数据库连接的方式，也可以通过文件交换的方式进行数据抽取，抽取的范围可以是结构化数据和非结构化数据。

2）质量检查。经过质量检查（见图 12-18），对数据进行清洗、取舍和去重，生成清洗后的数据文件，满足数据质量的基本要求。数据交换层的主要工作就是进行质量检查。不合格的文件是没有通过质量验证的数据。质量检查的内容包括数据的类型、格式和长度等内容。

3）数据转换。数据转换的功能是对数据质量清洗后的数据按照业务规则进行转换。

4）数据加载。数据加载的功能是创建可导入的文件，然后批量或者单条记录地导入到系统中。

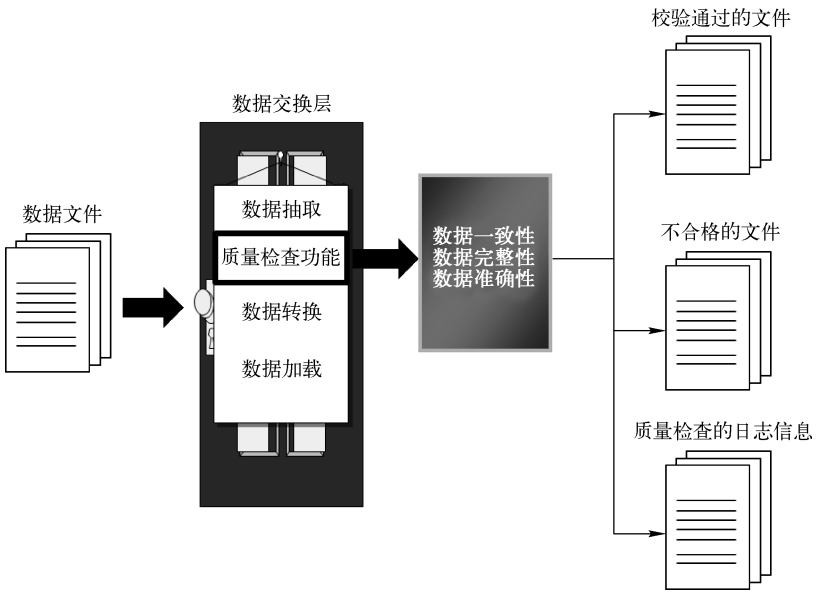


图 12-18　质量检查

（2）数据交换层的功能描述

1）数据交换层主要是数据交换的场所，它承担了各个层次之间的交换任务。

2）数据交换层支持外部交换的校验过程。

如图 12-19 所示，逻辑校验主要是缓冲区与加载区的数据进行关联校验，经过格式校验和逻辑校验之后，将数据加载到加载区中。

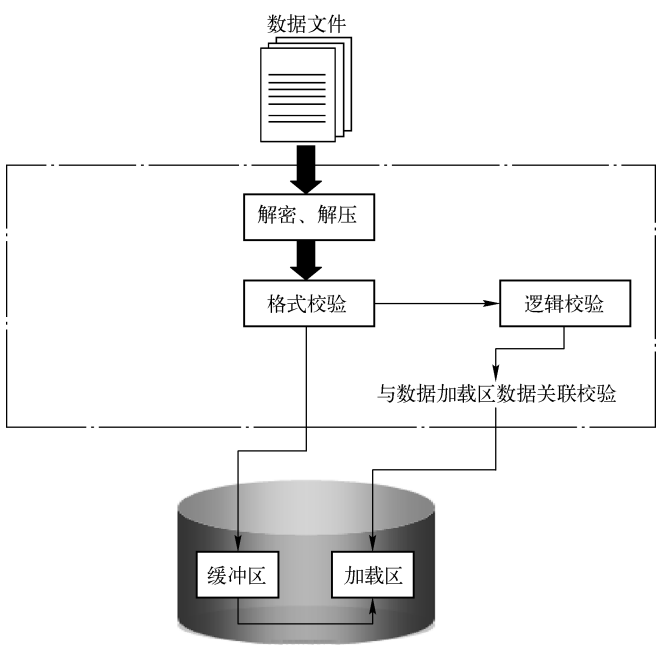


图 12-19　校验过程

3）数据交换层承担着内部系统和外部系统的数据交换任务。

如图 12-20 所示，对于主数据来说，可以将唯一身份信息通过数据交换层传输给外部系统。对于数据仓库来说，可以将质量检查结果通过数据交换层传输给外部系统。对于查询库来说，可以将查询记录通过数据交换层传输给外部系统。

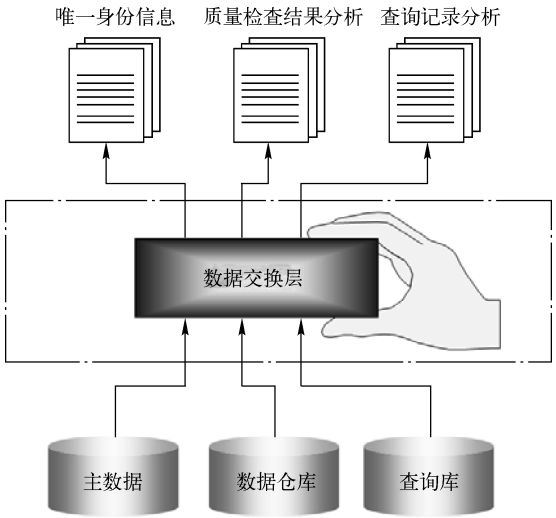


图 12-20　数据交换层支持内部系统和外部系统之间的数据交换

总之，数据交换层支持系统内部系统和外部系统之间的数据交换。

4）数据交换层支持系统内部的数据在各个数据库之间的流转。

5）数据交换层的订阅发布模式可以实现一源多目标的数据更新，如图 12-21 所示，当数据源发出一份数据文件后，根据订阅配置信息，将该数据文件传输到指定位置，然后根据不同的转换规则，把数据加载到不同的目标库中。

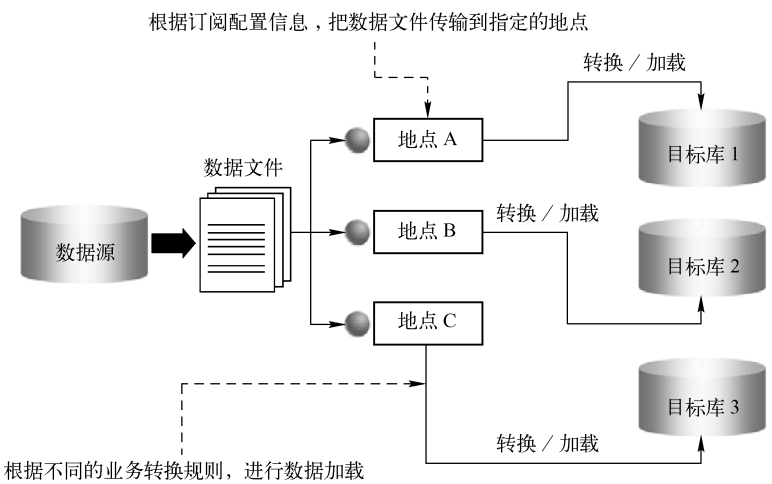


图 12-21　数据交换层的订阅发布模式

6）数据在传输过程中不进行任何加工的动作，如图 12-22 所示。同时确保数据传输与加工能够以流水线作业的方式进行，同时细化作业任务，分析作业任务之间的依赖关系，如图 12-23 所示。

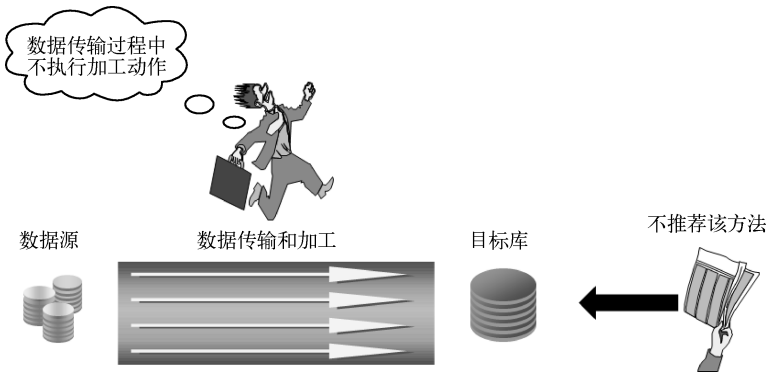


图 12-22　数据在传输过程中不进行加工的动作

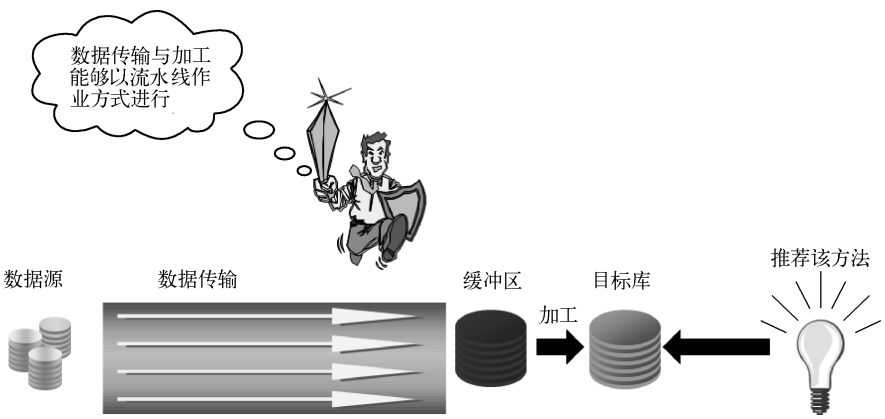


图 12-23　数据传输与加工以流水线作业的方式进行

4. 数据基础层

数据基础层是对抽取的数据进行格式校验和逻辑校验，它作为系统唯一可信的数据源。数据基础层包含三个部分：临时加载区、基础库和非结构化数据，如图 12-24 所示。

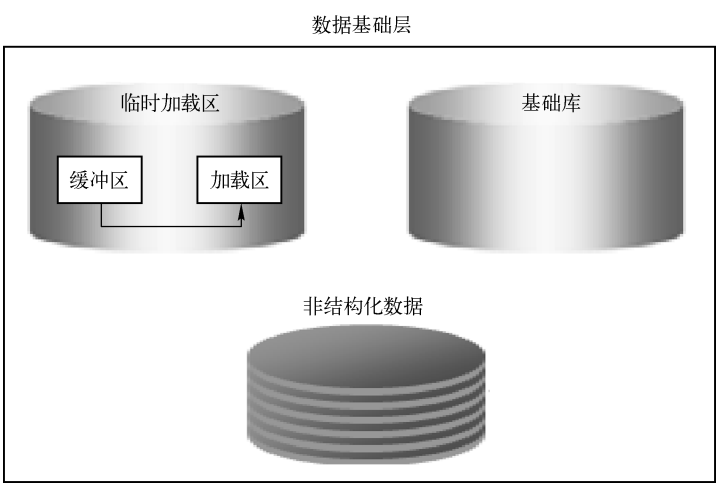


图 12-24　数据基础层

临时加载区作为校验数据进入系统的唯一途径，主要包括缓冲区和加载区。缓冲区是为数据交换设置的临时区域，为后续的逻辑校验做准备。而加载区主要完成格式校验和逻辑校验功能，如图 12-25 所示。

基础库存储的是系统唯一可信的数据源，存储的期限根据业务需求而定。它主要存储校验通过的数据。

5. 数据加工层

数据加工层的数据来源于数据基础层的基础库，然后将加工处理后的数据提供给应用层。数据加工层包括查询库、主数据和数据仓库，如图 12-26 所示。

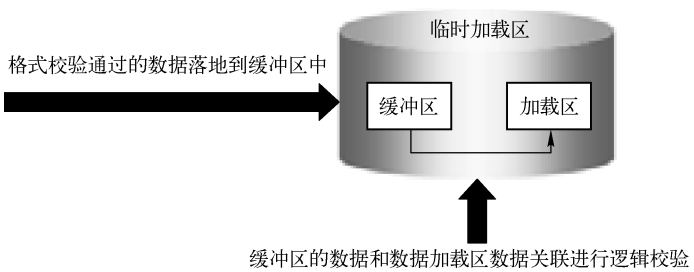


图 12-25　临时加载区

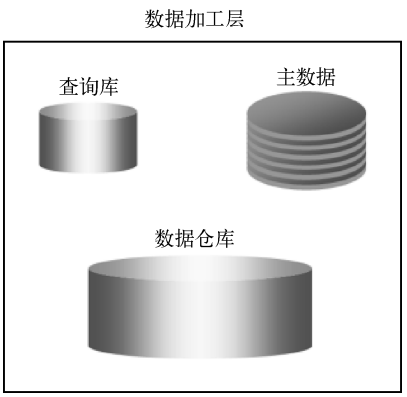


图 12-26　数据加工层

数据加工层的流程如图 12-27 所示。

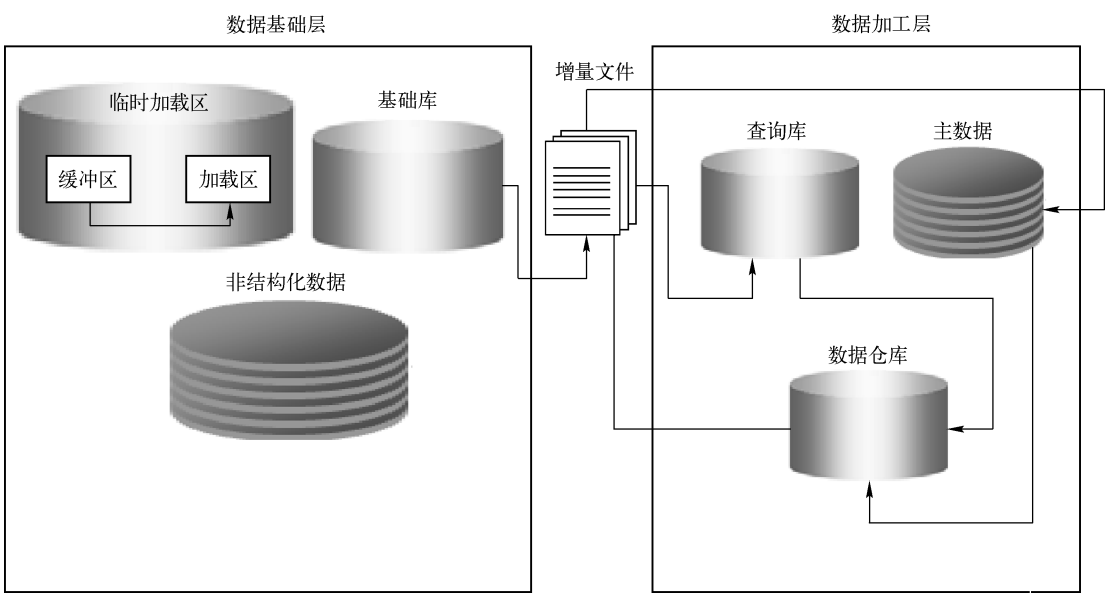


图 12-27　数据加工层的流程

（1）查询库

对于查询库来说，要求时效性高。基础库将数据导出成增量文件，加载到查询库中。

（2）主数据

主数据主要描述商业银行核心的信息，例如对于身份信息识别和归并的整合，尤其是当商业银行从以“账户为中心”向以“以客户为中心”转变的时候。对于客户身份信息的整合是非常重要的。主数据将整合后的结果再提供给数据仓库使用。

对于身份信息整合来说，可以按照时间的先后顺序进行覆盖，或者采用全部保留的方式。对于疑似身份信息的整合，有可能需要经过人工判断。

主数据也可以存储商业银行的客户关联信息。

（3）数据仓库

数据仓库一般包括基础数据层、汇总数据层和库内集市层。数据仓库有以下两个特性，如图 12-28 所示。

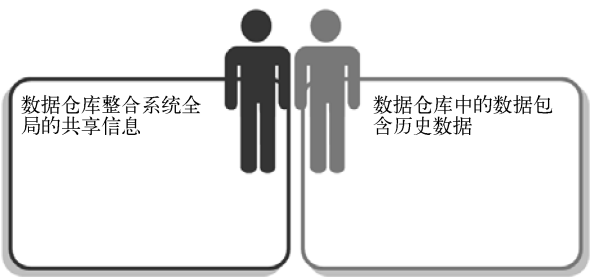


图 12-28　数据仓库的特性

1）数据仓库整合系统全局的共享信息。

可以收集、清洗、转换和存储各种操作型的数据源。

2）数据仓库中的数据包含历史数据。

它记录了系统从过去某一时间点到目前各个阶段的信息，通过对这些信息的分析，可以为企业的发展状况和未来趋势做出分析预测。其中数据仓库的数据包括基础库的数据、查询库的数据和主数据整合后的身份信息数据。

● 数据仓库基础数据层的特点

数据仓库基础数据层的数据是按照模型进行组织的。基础数据层的数据作为汇总层或者库内集市的数据源。基础数据层的数据一般不做删除。

● 数据仓库汇总数据层的特点

数据仓库汇总数据层主要是对基础数据层的数据进行轻度汇总，目的是减少共性的加工。

汇总数据层的建设是随着需求的增加而不断扩展的，对于汇总数据层的处理也是以创建中间表为主，目的是为后续数据加工使用做准备的，同时提高了数据仓库的性能。

● 库内集市层的特点

一般来说，数据集市层是根据应用需求而形成的数据集合，它支撑了各个部门的业务应用。每个部门都可以根据各自的需求，在集市上进行定义和维护。

数据集市可以分成分析类集市、研发类集市和管理类集市，如图 12-29 所示。

图 12-29　数据集市

● 分析类集市

分析类集市是通过数据挖掘的方法帮助企业提高业务运营效率，发现企业内部的规律和发展趋势。分析类集市可以包括文本分析、模拟分析、预测分析和可视化分析等，见表 12–2。

表 12–2　分析类集市

分析类集市	描　　述
文本分析	文本分析是对各种非结构化文本数据进行分析，将各种单词、短语赋予语义，我们通过词频统计，或者更复杂的过程进行分析。举例来说，情感分析是从大量的人群中挖掘出对某个企业或者机构的总体观点，同时提供客户对相关机构的各种评论和感受，使得企业或者机构可以更好地掌握客户感受，分析客户的真正需求
模拟分析	用先进的手段模拟业务流程、行为，帮助企业制定未来业务发展的方向
预测分析	分析历史和当前数据，预测企业未来的业务方向
可视化分析	通过图表、地图等各种可视化的形式，分析各种趋势

● 研发类集市

研发类集市是支撑各个业务部门的应用系统，主要用于支持研究分析类的工作，同时研发类集市也可以支持临时的抽数功能。

● 管理类集市

管理类集市是指为了提高运营管理而进行的整合分析。管理类集市包括：管理驾驶舱、固定报表、OLAP 分析、KPI 等，见表 12–3。

表 12–3　管理类集市

管理类集市	描　　述
管理驾驶舱	对高层人员关注的经营活动关键指标做定制化的展示，并且以各种直观的图表形式进行展示
固定报表	以固化报表的形式进行数据展示
OLAP 分析	以多维分析的方式帮助决策者发现问题、追溯问题根源和预测发展趋势
KPI	业务运营和绩效管理关键指标

其中基础数据库和数据仓库基础层的区别：

1）在组织形式上，基础库是贴数据源的数据，时效性较高，支持对基础产品的加工，为数据仓库提供数据源。

2）数据仓库基础层是按照第三范式的方式进行存储，时效性较低。数据仓库基础层支持汇总加工，同时支持高级分析。

6. 应用层

应用层包括查询类应用、分析类应用和管理类应用。应用层的数据可以批量加载，负责对外提供服务，同时查询记录可以回流到数据仓库的基础层，以支持分析类应用和管理类应用。

主数据的身份整合信息回流到数据仓库基础层，以支持分析类应用和管理类应用。

应用层的数据流转如图 12–30 所示。

查询类应用时效性较高，一些产品快照信息和查询记录可以返回给数据仓库。通过对产品数据的读写分离，可以最大限度地提高产品查询效率。

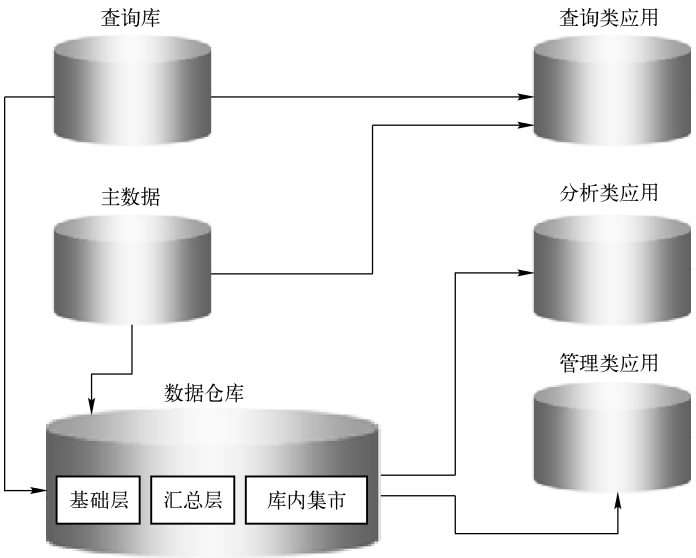


图 12-30　应用层的数据流转

12.3.2　互联网金融行业的数据架构

互联网金融实质上是互联网技术与金融服务的结合。互联网技术提高了金融行业的服务效率，拓宽了渠道和金融服务模式，但是没有改变金融行业的本质。互联网金融并没有改变传统金融行业中的存款、贷款、汇款和投资理财等功能。同时互联网金融企业依赖于传统商业银行提供的身份信息认证等服务。所以说互联网金融是对传统金融的有益补充。

但是互联网金融与传统的金融有一定的区别。例如，互联网金融非常重视客户的体验，特别是方便性和快捷性，但在安全性和严格性上还要不断提高。我们应该对互联网金融机构提出严格的监管要求，同时也对商业银行的创新战略提供新的思路。

在互联网金融的背景下，商业银行的创新思路包括以下几个方面，如图 12-31 所示。

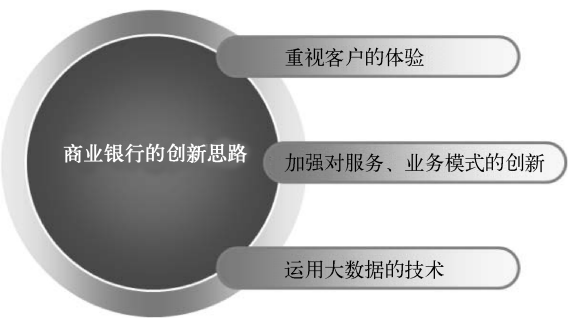


图 12-31　商业银行的创新思路

（1）重视客户的体验

商业银行要以客户为中心，从理解客户的角度设计金融产品和服务内容。同时优化银行内部工作流程，简化客户的操作，为客户提供方便快捷的高效服务。商业银行可以利用各种资源，例如移动终端、微博、微信和各种社交网站，开展全方位的客户营销。

（2）加强对服务、业务模式的创新

商业银行需要加强对服务、业务模式的创新，包括支付手段的创新、开发各种适合大小额支付的产品和对各种融资产品的创新等内容。

（3）运用大数据的技术

对于大数据技术的运用是提升商业银行核心竞争力的基础，它可以利用大数据技术优化业务流程，提升安全与风险的管理能力。

下面分析一下互联网金融和传统金融的区别，见表 12-4。

表 12-4　互联网金融和传统金融的区别

分类 / 项目	互联网金融	传 统 金 融
客户	面向所有互联网客户，包括银行客户和非银行客户	以银行客户为主
产品	包括所有的互联网金融产品和服务	以传统商业银行产品和服务为主
业务需求	业务需求变化较快	相对固定，同时有金融监管机构进行监督
渠道	所有与互联网相关的渠道	包括实体柜面、网上银行、ATM、手机银行等

互联网金融行业的 IT 架构主要包括应用架构、数据架构和技术架构，如图 12-32 所示。

图 12-32　互联网金融行业的 IT 架构

互联网金融行业的应用架构需要重点考虑技术的开放性，包括对大规模并发和快速响应需求的支持。数据架构主要考虑提高数据的智能程度，增强客户的体验度。技术架构强调建立一个安全的体系架构。

（1）互联网金融行业的应用架构

应用架构强调以客户体验为中心，数据驱动为主的原则。其中面向服务的架构设计，可以包括渠道层、业务操作层、产品服务层、决策支持层和基础应用层等几个部分。

（2）互联网金融行业的数据架构

数据架构主要强调数据的一致性和实时性，可以考虑对结构化数据、半结构化数据和非机构化数据的存储。同时考虑使用分布式云计算技术，以满足对海量数据存储、计算和多用户并发的使用。在大数据技术的使用上，可以考虑使用分布式文件系统、NoSQL 数据库、流数据处理技术等。

（3）互联网金融行业的技术架构

技术架构主要强调构建一个安全架构体系，主要包括合规、治理、人员、运维和各种流程监控等。同时业务可以扩展到云平台、虚拟化环境和社交网络平台。

12.4 金融行业的商业智能

12.4.1 金融行业商业智能的背景和作用

在当前市场竞争激烈和商业银行业务转型的大背景下，商业银行正面临着各种机遇和挑战。利用商业智能技术，可以大大提高商业银行的服务水平和内部管理水平。特别是在数据大集中的背景下，商业智能已经成为商业银行信息化建设的必然选择之一。

商业智能（BI）是对各种信息收集、管理和分析的过程，目的是使企业的决策者能够获得知识和洞察力。商业智能一般由数据仓库、数据集市、数据挖掘和在线分析等部分组成。商业智能提高了企业和商业银行的管理水平，强化了对风险管理和产品的创新能力。

同时商业智能可以更好地帮助企业抓住机遇，应对市场挑战。商业智能的作用主要体现在以下几个方面：对客户的信息进行整合，商业银行的风险管理能力将会得到提高，商业银行可以实现内部的精细化管理，帮助商业银行发现有价值的客户群体，如图 12-33 所示。

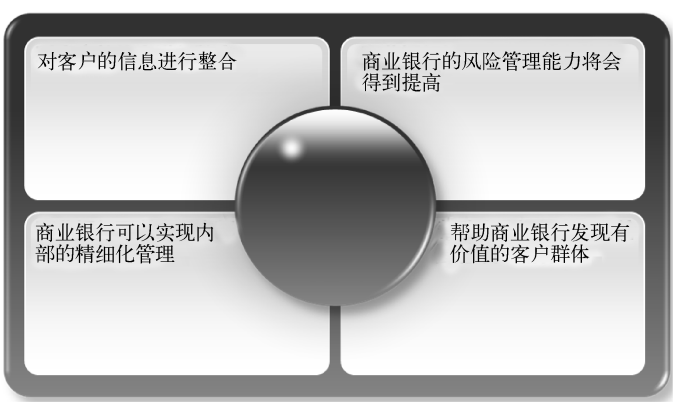


图 12-33 商业智能的作用

（1）对客户的信息进行整合

通过商业智能技术，可以对客户的信息进行整合，能够反映客户信息的全貌，使得管理者可以从产品类型、行业、机构等不同的角度对客户关心的各类指标进行分析，使得分析更具有针对性。

（2）商业银行的风险管理能力将会得到提高

通过商业智能技术，商业银行的风险管理能力将会得到大幅度提高，例如各种的操作风险、客户信用风险、市场风险和业务运营风险等将会得到有效控制。通过集中数据，使得风险管理机构能够全面掌握数据，可以根据对历史数据的分析，实现当前业务的预警和风险评级。

（3）商业银行可以实现内部的精细化管理

通过商业智能技术，商业银行可以实现内部的精细化管理，使得各种绩效考核和成本管理更准确，同时能够在产品、客户、机构等各条业务线上对指标进行量化。

（4）帮助商业银行发现有价值的客户群体

通过商业智能技术，可以帮助银行发现有价值的客户群体，针对这些客户的价值度和贡献度，有针对性地设计出更好的金融产品，从而更好地为客户服务，同时实现利润的最大化。

12.4.2 金融行业如何实施商业智能

金融行业商业智能的实施离不开高层领导的重视，同时需要投入大量的资源。在制定整体规划的同时，需要明确各个阶段的实施重点。可以按照商业智能的实施方法论开展工作，包括建立数据仓库、数据集市、元数据管理系统、OLAP 等。

在实施商业智能的同时，同样需要业务部门和技术部门的广泛合作，开发出适合业务发展的商业智能应用系统。

金融行业实施商业智能主要有以下几个方面的内容：

1）商业智能的实施需要由业务进行推动，首先应该明确业务发展的方向，制定出各个阶段商业智能实施的重点，为商业智能大规模的应用提供经验，同时短期内可以促进业务的发展，增强下一阶段工作的信心。对于商业银行来说，首先应该完成对客户信息的整合，形成基础数据，然后建立数据仓库、数据集市、OLAP 分析等基础架构。通过对各种业务应用的实施过程，形成完备的技术架构，逐步建立起具有实施能力的团队。

2）在商业智能的实施过程中，需要重视对数据的清洗和整合，为数据仓库的建设打下基础。在此基础上，需要注重对数据资源的整体规划，制定出实施步骤，保证实施的长效性。同时推动业务流程的改进，完善业务活动环节，发挥商业智能的价值。

金融行业可以将商业智能系统划分成以下几个层次：数据源层、数据模型层、可视化组件层和交付展示层，如图 12-34 所示。

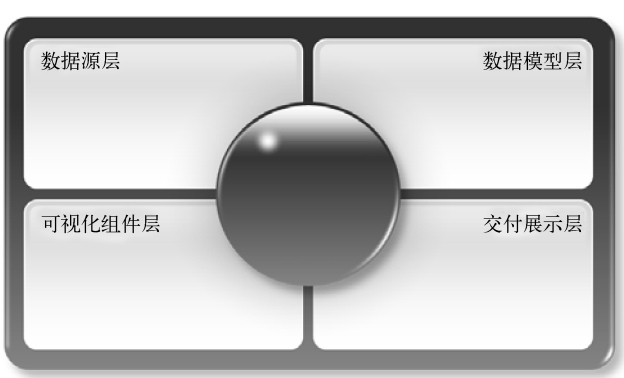


图 12-34　金融行业的商业智能系统划分

- 数据源层

主要支持各种数据源，例如 Hadoop、NoSQL、文本、Excel、CSV 等。

- 数据模型层

主要针对各类数据源、大数据集群和集成的企业应用模型，同时支持 OLAP 立方体模型。

● 可视化组件层

可视化组件层主要包含管理驾驶舱、报表、多维分析、数据集成和数据挖掘，如图 12-35 所示。

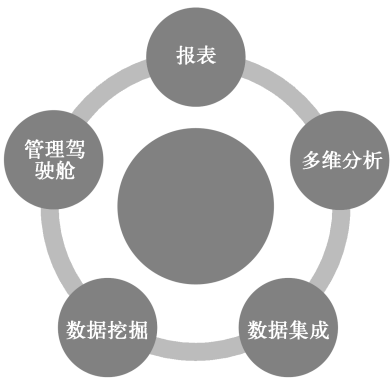


图 12-35　可视化组件层

可视化组件层各组成部分的功能及用户见表 12-5。

表 12-5　可视化组件层各组成部分的功能及用户

名　　称	具 体 功 能	主 要 用 户
管理驾驶舱	通过各种指标体系，实时反映该金融机构的运行状态，将采集的数据形象化和具体化	业务用户
报表	即席分析、操作报表	业务用户
多维分析	高级分析、多维度探查数据	业务用户、高级用户
数据集成	高性能数据集成、大数据清洗	高级用户、开发人员、DBA
数据挖掘	高级预测分析	业务用户、超级用户

● 交付展示层

交付展示层提供面向用户和集成商的全面接口，主要包括 Web、移动终端、打印、电子邮件等数据输出支持。一般来说，交付展示层可以提供丰富的二次开发接口及应用服务接口。

12.4.3　金融行业的业务流程和运营模式优化

在商业银行中，业务部门会提出各种需求，同时 IT 部门会根据计划对各种需求进行立项。当系统设计、开发完毕，一直到上线后，IT 部门会根据业务部门提出的问题进行修改和优化。目前来说，多数商业银行很少关注对业务流程和运营模式的优化，然而业务流程和运营模式的优化可以促进商业银行业务的发展。那么如何进行优化呢？

我们整理一下整体的思路：

通过对金融行业的环境分析，对战略的理解和核心业务流程的描述，同时参考行业内先进的实践经验，提出对商业银行业务流程的优化和改进意见，如图 12-36 所示。

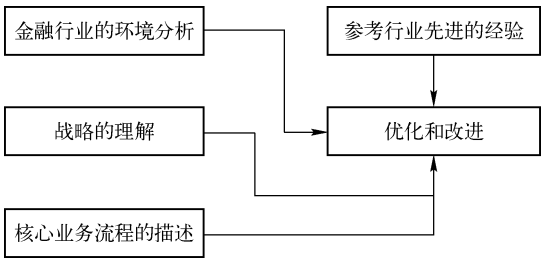


图 12-36　金融行业的业务流程和运营模式优化

● 金融行业的环境分析

主要了解金融行业的整体发展趋势、技术发展水平和竞争态势，以及该金融机构在市场竞争中将要面临的机遇和风险等内容。

● 对金融行业战略的理解

主要理解金融行业的发展方向、战略目标。

● 核心业务流程的描述

识别关键的业务流程，找出业务流程和运营模式需要改进的地方。

● 优化和改进

参考行业内先进的经验和业务现状，优化和改进该金融机构的业务流程和运营模式。

1. 对金融行业的环境分析

金融行业的环境分析主要包括行业的发展趋势、技术变革等几个方面。

1）对金融机构来说，如何能够为客户提供丰富的产品和服务是首要问题。例如，随着社会老龄化的到来，一些金融机构可以为老人提供风险较低的理财产品，以满足客户对生活的基本需求。同时也可以考虑其他的消费群体，为年轻人提供更方便和快捷的移动金融服务等。

2）随着资本市场的成熟，金融脱媒现象越发明显，特别是利率市场化的调整，大幅度降低了商业银行的利润空间。

3）随着技术的进步，客户获取信息的渠道越来越多，特别是互联网和社交网络的发展，增加了金融机构和客户之间的信息不对称。这要求金融机构通过各种渠道采集完整的客户信息，减少这种不对称性，提高金融机构的决策分析能力和风险管控能力。

2. 对金融行业先进经验的分析

通过对国内外金融机构先进经验的分析，以市场作为驱动力，强化对产品的创新能力和对外服务能力。对核心的竞争优势进行分析，从而提高自身的能力。例如，很多国外先进的金融机构，它们的业务模式可以包括决策分析、行业解决方案等。它们基于数据匹配和整合技术，为客户提供各种专业化和个性化产品和服务。

3. 核心业务流程的描述

对于金融机构来说，它的核心业务流程是数据采集、产品加工、产品研发和对外服务。其中数据采集和产品加工是数据流动的过程。产品研发是从产品的设计、研发，一直到产品上线的过程。对外服务是客户申请服务到服务终止的过程。

对于商业银行来说，如何提高对客户的服务能力和工作效率，降低运营成本，提升产品研发和对外服务的核心竞争力是业务优先关注的地方。

4. 优化和改进

关于金融行业业务流程和运营模式的优化和改进措施主要包括以下几种手段：对数据采集、产品加工的优化，对产品服务的优化，对产品研发流程的优化等等，如图 12–37 所示。

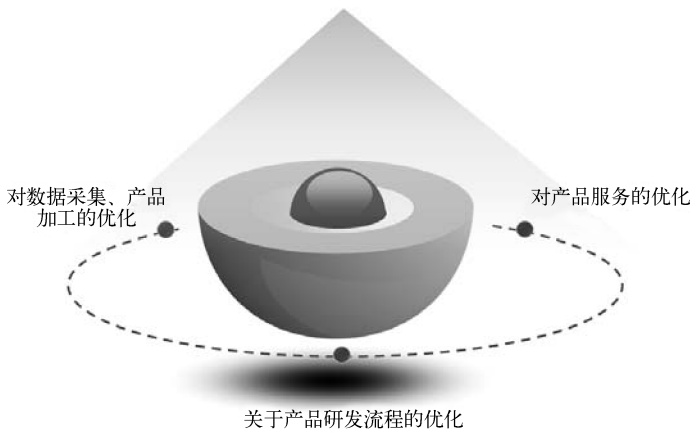


图 12–37　金融行业业务流程和运营模式的优化和改进措施

（1）对数据采集、产品加工的优化

目前很多商业银行的数据采集、产品加工的扩展性都不高。对数据自动化处理能力、数据质量和采集策略的管理能力普遍较低。

我们可以在数据源规划、调度监控管理和校验等几个方面进行改进和优化。

1）参考数据源业务发生的频率，提高数据采集的灵活性。将数据采集、数据加工和对外服务进行综合考虑，以实现业务之间的平衡。

2）通过调度监控的管理，实现各个作业任务之间的协调，使不同的业务环节围绕在统一体系下。解决办法是建立数据采集和调度监控机制，加强产品加工的能力。同时可以收集宏观的产品需求信息、产品的反馈信息等，然后对客户群进行细分。

3）整合业务的流程，提高自动化程度，减少手工干预的工作。加强数据质量、查询匹配、数据整合等关键环节的能力。我们也可以把数据质量管理工作前移，保证数据入库之前的质量，可以采用抽样统计与逐条数据校验的方式，规避系统性的数据错误，作为数据质量提升的策略之一。

举例来说，我们可以根据历史数据的报送情况，动态调整抽样和统计的规则，借鉴国外身份信息的整合经验，以自然人为单位，作为整合的对象，利用个人姓名、证件号码、地址信息和电话号码进行整合。很多金融机构因为质量管理手段单一，并且以逐条记录校验为主，所以效率很低。可以通过建立数据质量跟踪和反馈机制，同时提供相应的激励措施等方法提高效率。

（2）对产品服务的优化

1）首先对产品服务的现状进行分析。例如，判断一些金融机构是否建立了以市场化为主的产品服务流程体系。在产品服务的各个环节之间是否存在信息共享。金融机构需要清晰的服务定位。通过不断优化产品体系，改善客户的产品体验度，完善服务规范体系等手段达到优化的目的。

2）通过多维度的分析金融机构的特点，对客户群进行细分，提供有针对性的差异化服务，以满足金融机构在不同业务场景下的信息需求。对于金融机构来说，它们应该重点分析对外可以提供哪些服务，如何保证对外产品服务的标准化。同时可以为客户提供灵活的查询引擎，支持产品的组装等方面。

（3）对产品研发流程的优化

产品研发流程主要包括：理解产品设计的功能，产品研发进度情况，市场的动向和设计变更决策等。关于产品研发流程的优化是最富有挑战性的工作之一。

小结

- 金融就是在我们的经济生活中，通过银行、证券机构等中介，从市场主体中募集资金，然后在借贷给其他市场主体的活动，可以把金融看做融资、投资和资金募集等三种经济活动。
- 对于商业银行来说，它有大量的客户群，可以吸收社会公众存款，资金实力非常雄厚，抗风险的能力比较强。同时银行有大量的客户信用数据，包括各种客户信用卡消费信息、贷款信息、还款信息和信用信息等。
- 随着互联网技术的进步，商业银行通过互联网融资会更有利，因为商业银行本身具有良好的信用基础和声誉，各种贷款、股票和债券都可以通过互联网进行交易。同时也可以利用互联网技术解决信息不对称的问题。对于银行来说，借贷业务仍然是银行的核心业务，它的净利息收入占到 70% 左右。
- 商业银行应该具备的能力主要包括对客户的洞察力、精准营销和跨渠道客户管理。
- 金融行业的数据架构一般包括以下几个部分：数据采集层、产品加工层和对外服务层。
- 从数据源开始，经过加载、集中、整合，以及对外服务这几个过程，可以将整个数据架构横向划分成：源数据区、基础区、产品加工区和产品服务区。各个区域都相对独立。
- 数据架构是企业架构的重要组成部分，帮助金融行业有效地分配、部署和使用数据，实现数据的合理组织和有效共享，从而保证数据在各个系统之间的一致性、完整性和有效性。
- 在当前市场竞争激烈和商业银行业务转型的大背景下，商业银行正面临着各种机遇和挑战。利用商业智能技术，可以大大提高商业银行的服务水平和内部管理水平。特别是在数据大集中的背景下，商业智能已经成为商业银行信息化建设的必然选择之一。
- 商业智能（BI）是对各种信息收集、管理和分析的过程，目的是使企业的决策者能够获得知识和洞察力。商业智能一般由数据仓库、数据集市、数据挖掘和在线分析等部分组成。商业智能提高了企业和银行的管理水平，强化了对风险管理和产品的创新能力。
- 金融行业商业智能的实施离不开高层领导的重视，同时需要投入大量的资源。在制定整体规划的同时，需要明确各个阶段的实施重点。可以按照商业智能的实施方法论开

展工作，包括建立数据仓库、数据集市、元数据管理系统、OLAP 等。

- 金融行业的环境分析主要包括行业的发展趋势、技术变革等几个方面。
- 通过对国内外金融机构先进经验的分析，以市场作为驱动力，强化对产品的创新能力和对外服务能力。对核心的竞争优势进行分析，从而提高自身的能力。
- 对于金融机构来说，它的核心业务流程是数据采集、产品加工、产品研发和对外服务。其中数据采集和产品加工是数据流动的过程。产品研发是从产品的设计、研发，一直到产品上线的过程。对外服务是客户申请服务到服务终止的过程。

第 13 章　电力行业数据架构和商业智能案例

本章目标

通过前一章的学习，读者已经掌握了金融行业背景概述、金融行业的数据架构、传统金融行业某系统的数据架构案例、互联网金融行业的数据架构案例、金融行业的商业智能概述、金融行业商业智能的背景和作用、金融行业如何实施商业智能、金融行业的业务流程和运营模式优化等内容。

学习本章后，读者将掌握：

- 电力行业面临的挑战
- 建设电力行业企业级数据仓库的因素和策略
- 电力行业商业智能的数据架构
- 电力行业商业智能系统开发流程
- 数据仓库运维内容
- 电力行业数据仓库的建设方法
- 商业智能运维组织架构
- 针对电力行业的数据管理
- 关于电力行业的数据质量管理
- 关于电力行业的数据标准管理
- 关于电力行业的数据安全管理

13.1　电力行业商业智能

1. 电力行业面临的挑战

电力行业主要面临着业务挑战和技术挑战，如图 13-1 所示。

（1）业务挑战

1）电力行业的分析系统一般仅提供简单的报表功能，功能单一，高层人员无法从全局的角度对各条业务线进行多层次的综合分析。

2）对于各个分析系统来说，它们又集中于各自的领域，不具备跨业务的分析能力，存在着数据不一致的现象，不能有效地发挥电力行业数据资产的价值。

（2）技术挑战

1）分析型系统与业务生产系统耦合性较强，缺乏对全局业务分析的支持，对于相同业务数据，可能会存在不同的版本。

2）各个业务系统管理着各自的数据，数据的业务含义在各个部门之间可能存在不一致的解释，数据质量也相对较低。

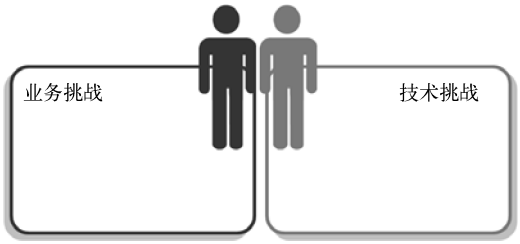


图 13–1　电力行业主要面临着业务挑战和技术挑战

2. 建设电力行业企业级数据仓库的因素和策略

建设电力行业企业级数据仓库的因素主要包括业务因素和技术因素，如图 13–2 所示。

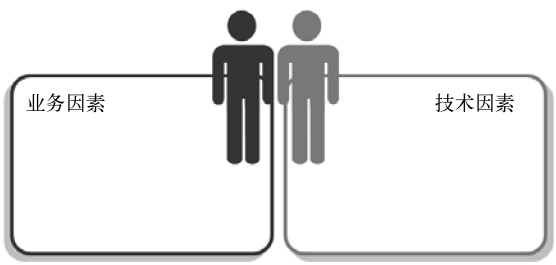


图 13–2　建设电力行业企业级数据仓库的因素

● 业务因素

在业务上，缺乏统一的报表与指标规范体系，缺少明确的数据责任体系。

● 技术因素

缺少规范的数据架构，导致数据分布的不合理和模型的不一致。同时数据管理不规范，缺乏企业级的数据整合和管控机制。

3. 电力行业企业级数据仓库的建设策略

1）电力行业对数据分析的需求有一定的差异性，对于分析应用，允许各个省市存在个性化的内容。

2）对于电力行业企业级数据仓库的核心模型，应该有一个统一的数据标准，它可以帮助各个省市建立统一的数据管理体系，通过试点地区的成功经验推广，减少其他省市数据仓库实施的风险。

4. 电力行业商业智能的数据架构

电力行业商业智能的数据架构包括源数据层、数据抽取层、数据存储层、数据访问层和用户访问层。

● 源数据层

主要包括各个业务系统的数据。

● 数据抽取层

主要包括抽取、清洗、转换和加载。

● 数据存储层

主要包括 ODS、数据仓库和数据集市。

● 数据访问层

主要工作流程包括用户应用通过 Web 浏览器提交数据请求，Web 浏览器通过 Internet 发送 HTTP 请求给 Web 服务器。数据请求发送给应用服务器。获得数据后以 HTTP response 的形式发送给用户。

- 用户访问层

主要包括：报表、查询、在线分析和知识发现等。

电力行业商业智能的数据架构的实现如图 13-3 所示。

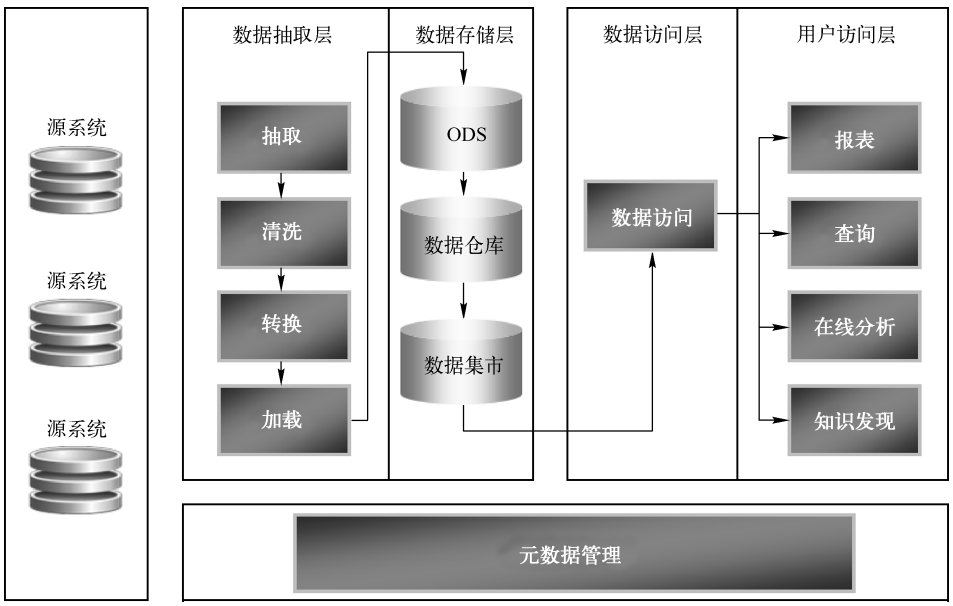


图 13-3　电力行业商业智能的数据架构的实现

5. 电力行业商业智能系统开发流程

电力行业商业智能系统的开发流程主要包括计划，分析，设计及开发，测试，部署，如图 13-4 所示。

图 13-4　电力行业商业智能系统的开发流程

- 计划

计划包括复查期望的目标，评估系统现状能力，定义系统建设方案等内容。

- 分析

分析包括对高层需求的确认，定义数据分析的需求，建立概念模型，评估系统建设风险，定义开发和执行环境的需求，制定 UAT 计划和性能测试计划等内容。

● 设计及开发

设计及开发包括制定报表开发规范，建立逻辑模型和物理模型，设计 ETL 的开发流程，部署 ETL 开发程序的测试环境等内容。

● 测试

完成对商业智能的产品测试、性能测试和 UAT 测试等。

● 部署

评估部署条件，完成数据转换，最后发布应用程序。

6. 数据仓库运维内容

电力行业数据仓库系统的运维内容主要包括：备份与恢复，归档与恢复，系统监控，容量规划，性能管理，如图 13-5 所示。

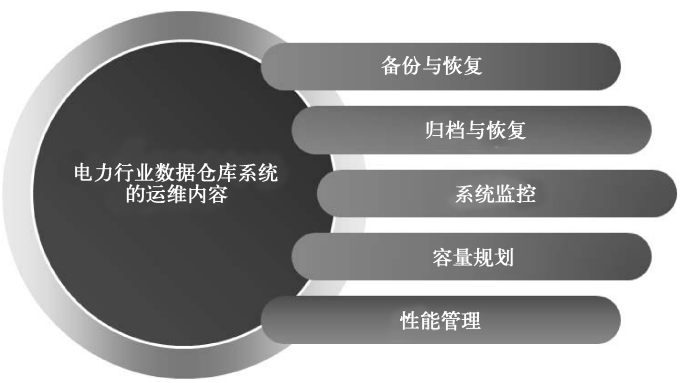


图 13-5　电力行业数据仓库系统的运维内容

● 备份与恢复

数据仓库的定期备份与恢复是数据仓库运维的重要环节之一，它需要满足用户对于业务恢复执行频率与速度的要求。这些流程必须满足用户的可用性需求和数据的线性增长要求。

● 归档与恢复

对于数据仓库运维人员来说，数据的归档活动经常被忽略，但是数据量不断增加，使得数据仓库需要增加额外的存储设备，增加了系统的复杂性。正是上述原因，使得数据仓库不能永久地保存数据，需要将历史数据归档到离线存储设备上。

● 系统监控

对于数据仓库来说，系统的监控工作更加复杂，很多数据仓库系统的建设都忽略了对数据库使用情况的监控，这些监控信息可以帮助系统管理员对数据库进行调整，以满足对现在和未来数据容量的需求。

● 容量规划

对于数据仓库来说，CPU、内存、硬盘和网络等硬件资源的容量计算是非常关键的工作。在数据仓库系统中，硬件开销最大。特别是服务器、存储等基础设施的成本很大，硬件成本直接影响了企业 IT 系统的总体成本，所以做好系统容量的计算是降低 IT 系统的成本，提高运营绩效的重要途径之一。

● 性能管理

性能是数据仓库架构中每个组件都需要考虑的问题。在架构过程中需要考虑系统的性能问题，例如系统负载、索引构建、大文件传输、用户查询响应时间、备份与恢复时长等。

13.2 电力行业相关商业智能案例

1. 电力行业数据仓库的建设方法

数据仓库开发应实施以全局的观点为基础，业务需求为导向的滚动式开发方法，如图13-6所示。

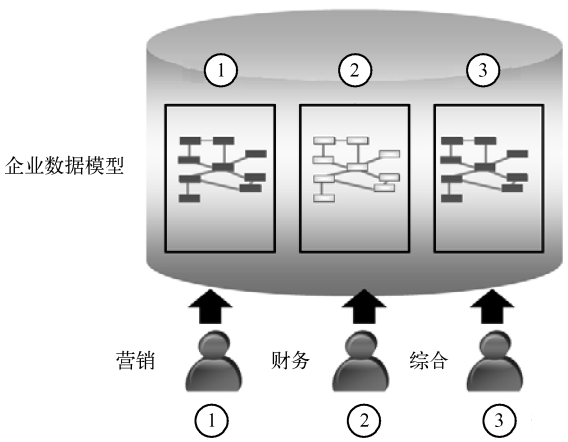


图13-6 电力行业数据仓库的建设方法

关于省市级的数据仓库演进方法是以数据仓库分析能力和数据整合能力的提高为主线索，提升数据管控能力，改进数据质量。

1）首先采用Quick Win（速赢）方式，建立领导查询系统，如图13-7所示。

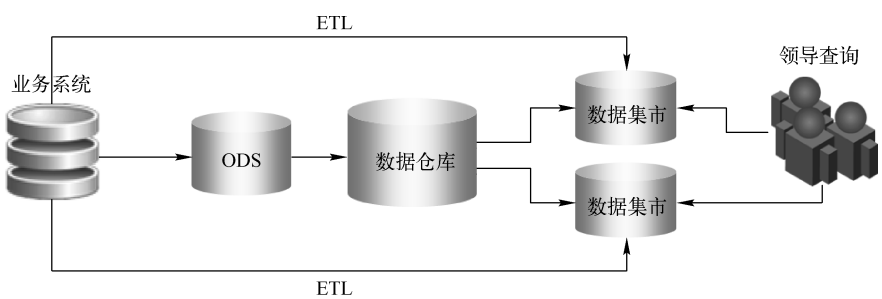


图13-7 Quick Win（速赢）方式

2）然后，建立数据仓库，并且对数据仓库不断地进行完善和改进。挑选重要的主题进行数据仓库建设，提供联机分析及综合报表，如图13-8所示。

3）数据仓库优化。在优化阶段，数据仓库已经基本建成。在此阶段，数据仓库可以提供更全面的数据分析以及数据展现功能，包括对数据进行更深层次的挖掘，如图13-9所示。

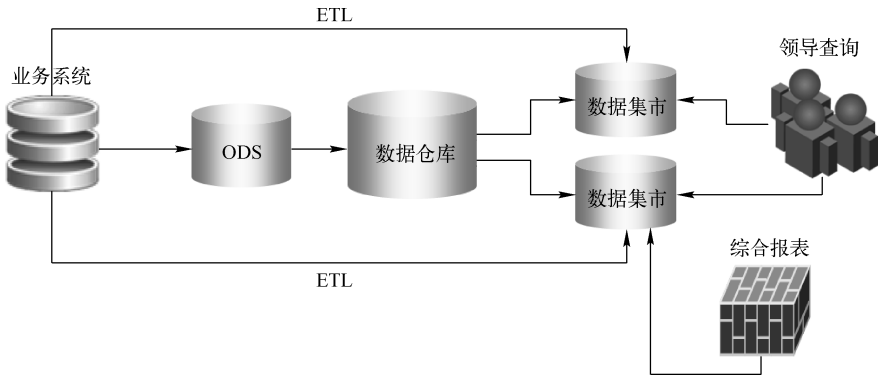


图 13-8　建立数据仓库并不断完善和改进

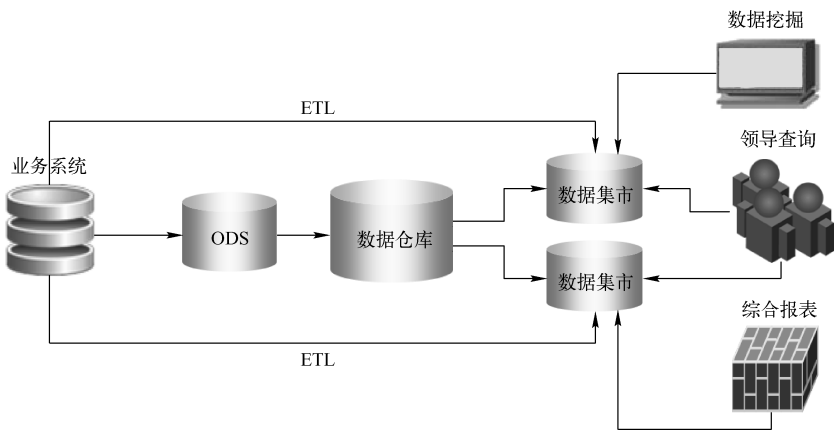


图 13-9　数据仓库优化

2. Quick Win（速赢）阶段的工作任务和效果

（1）阶段任务

功能：实现综合分析和领导查询。

技术：建立数据仓库技术架构，包括开发环境、执行环境和运维环境。

管理：初步统一编码，使数据集市中的标准一致，对于地市级上报的指标统一口径。

（2）效果

该阶段基本实现综合分析和统计功能，包括指标的查询和统计，表现方式主要是普通报表、图形和仪表盘等。但是对明细数据的分析能力有限，缺乏丰富的多维分析能力，从整个架构上看，只有数据集市，没有建立企业级的数据仓库。整体的架构在这个阶段基本形成。数据集市中的数据可以自动更新。

3. 数据仓库建立及完善阶段的工作任务和效果

（1）阶段任务

功能：完善综合分析和领导查询。

技术：在数据仓库中建立客户、产品、财务主题域的物理模型，将数据源转化后进入到数据仓库中，实现数据仓库数据到数据集市的转换，对于数据集市中的指标，逐步转向由数据仓库计算得来。

管理：建立数据质量管理团队、方法和流程，对数据质量进行分析，实施数据安全的分级策略，使用户对数据具有不同的访问权限。同时建立数据标准管理团队，对数据标准进行管理维护，初步具有应对数据标准需求的能力。在此基础上，建立数据仓库运维架构，包括组织、流程、方法等内容。

（2）效果

综合分析和统计中的指标可以从数据仓库中统计得来，指标的准确度和自动化程度得到提高。企业级数据仓库初步形成，并且具备一定的数据整合能力，为分析提供明细和汇总的数据。例如，通过对电量、电费、电价的分析，提高电量的需求预测和价格制定能力。同时可以全面了解客户的电力消费和缴费情况，帮助制定相关的政策和服务措施。它可以基于OLAP 分析技术做更深入的数据分析，数据质量逐步得到改善，保证数据仓库系统运行时的高可用性。

4. 数据仓库优化阶段的工作任务和达到的效果

（1）阶段任务

功能：完善综合分析及领导查询，完成剩余的数据分析功能 。

技术：在数据仓库中建立其他主题域的物理模型，建立数据源到数据仓库的映射关系，将数据源进行转换后再送入到数据仓库中。在此基础上，建立其他的数据集市，并且实现数据仓库到数据集市的转换，使集市中的指标，全部转向由数据仓库计算得来。

管理：优化数据标准维护流程、数据质量管理流程，同时优化数据仓库运维能力，建立数据生命周期。

（2）效果

对于综合分析和更多的指标可以从数据仓库中统计得来，指标的准确度和自动化程度得到优化和提高。企业级数据仓库已经形成，具备数据整合能力，为数据分析提供充分支持。数据质量进一步改善，在源头对数据质量进行管理，使运维效率得到提高。

其中，电力行业商业智能组织架构如图 13-10 所示，主要包括项目领导小组、项目管理办公室、专家组、质量监控组、项目经理/项目实施管理团队、架构设计组和开发测试组。

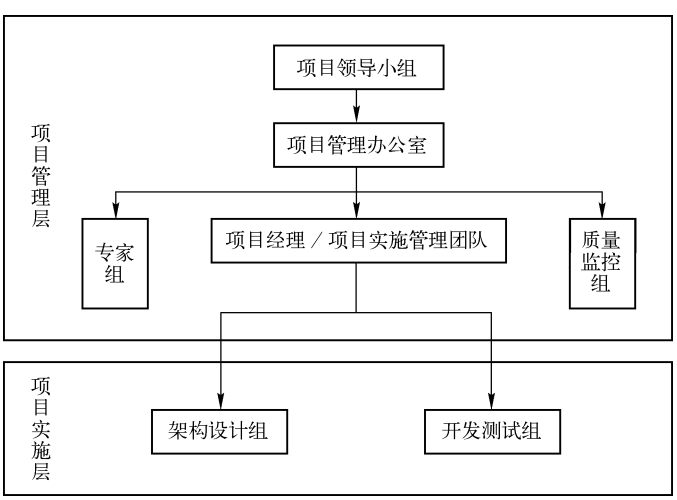


图 13-10　电力行业商业智能组织架构

电力行业商业智能的任务流程如图 13-11 所示，主要包括计划阶段、分析阶段、设计阶段、开发阶段、测试阶段和部署阶段。

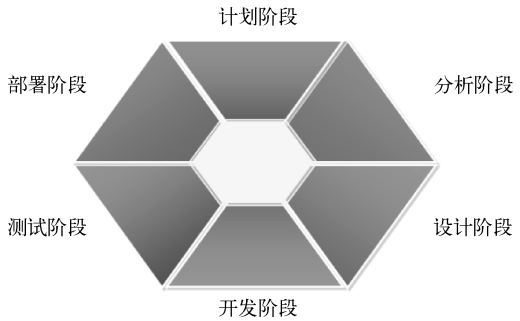


图 13-11　电力行业商业智能的任务流程

1）计划阶段

计划阶段主要包括定义期望目标、评估现状能力、定义方案和定义交付策略，如图 13-12 所示。

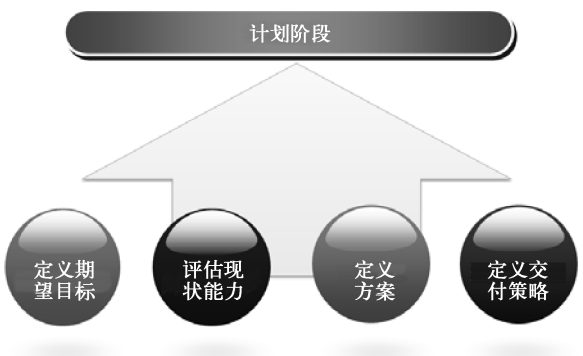


图 13-12　计划阶段

● 定义期望目标

包括愿景及业务目标，确认目标业务流程，定义方案蓝图等。

● 评估现状能力

评估当前业务流程和当前的能力、性能、风险等内容，分析当前技术架构、组织架构、数据管控的现状。

● 定义方案

定义应用解决方案、技术解决方案、业务流程变更解决方案和运维解决方案。

● 定义交付策略

定义开发策略、测试策略、试点策略、部署策略、元数据管理和数据管控策略。

2）分析阶段

分析阶段主要包括定义数据分析需求、建立概念数据模型、定义用户访问需求、评估风险、定义开发运行环境需求、制定 UAT 计划和性能测试计划，如图 13-13 所示。

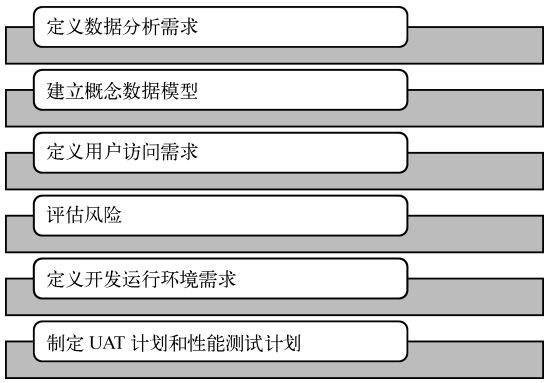


图 13-13　分析阶段

3）设计阶段

设计阶段主要包括定义 ETL 技术整合方案和报表详细规范，建立逻辑数据模型，对设计开发和运行环境的准备，制定测试计划，如图 13-14 所示。

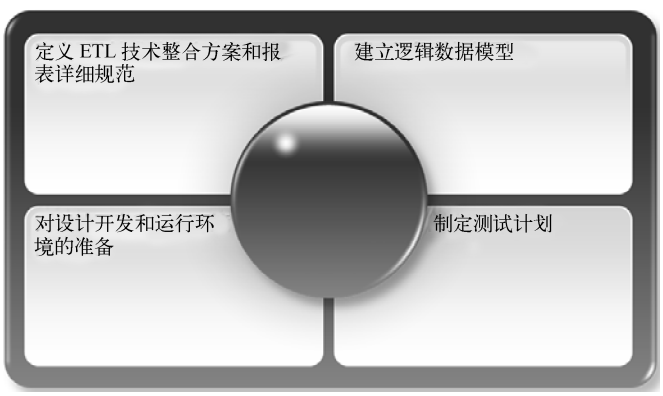


图 13-14　设计阶段

4）开发阶段

开发阶段主要包括制定 ETL 开发流程，前台组件开发，物理数据模型开发，开发、运行环境的准备，ETL、报表组件测试计划，如图 13-15 所示。

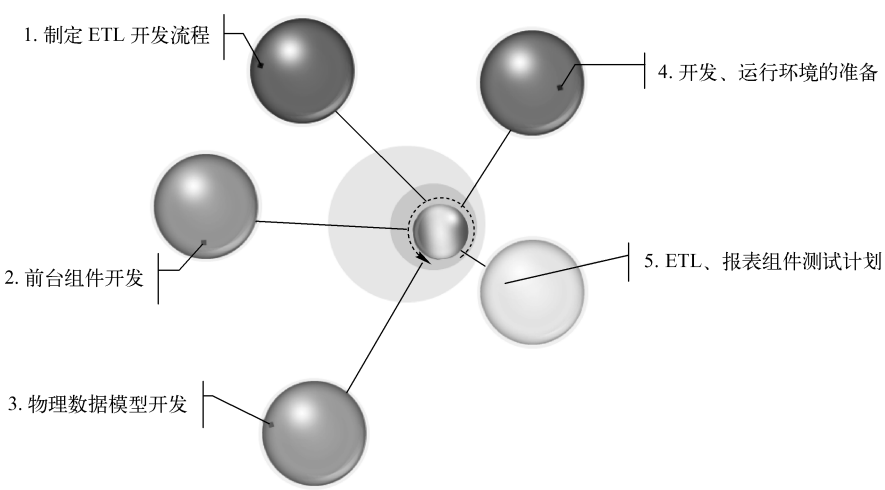


图 13-15　开发阶段

其中 ETL 开发流程包括数据映射、逻辑设计、调度设计、编码等。

5）测试阶段

测试阶段主要包括组件测试、产品测试、性能测试、UAT 测试，如图 13-16 所示。

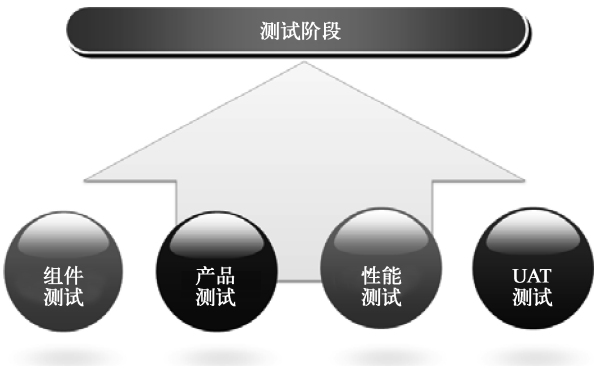


图 13-16　测试阶段

● 组件测试

组件测试包括编写组件测试的脚本、发布测试环境、执行组件测试，最后根据测试结果及时通报错误并修复。

● 产品测试

产品测试包括确认产品测试的周期、编写测试脚本、发布测试环境、执行产品测试，最后根据测试结果及时通报错误并修复。

● 性能测试

性能测试包括确认性能测试周期、编写测试脚本、发布性能测试环境、执行性能测试，最后根据测试结果及时通报错误并修复。

● UAT 测试

主要包括对用户培训手册、测试脚本、测试场景、测试策略和测试用户的准备。

6）部署阶段

部署阶段主要包括评估部署条件、完成数据转换和部署测试、发布应用，如图 13-17 所示。

图 13-17　部署阶段

其中评估部署条件主要是评估应用程序、技术架构、部署站点和基础架构的准备情况，同时制定对偶发事件的应急处理机制，详细列出每一阶段的检查点。完成数据转换主要包括清洗数据、创建数据备份、执行数据转换，最后验证转换后数据的正确性。完成部署测试主要包括执行部署测试、验证结果、结果反馈、错误修复，最后将部署结果通知开发和实施团队。

5. 电力行业商业智能运维组织架构

电力行业商业智能运维组织架构层次一般为运维中心、服务支持、技术支持等团队。例如，服务支持包括设施支持人员、流程管理人员；技术支持团队包括商业智能（BI）支持

人员、网络支持人员、存储支持人员、操作系统支持人员，如图 13-18 所示。

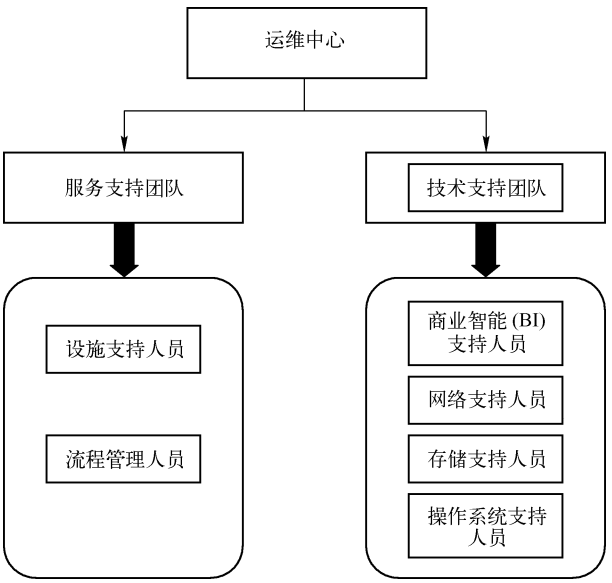


图 13-18　商业智能运维组织架构

6. 电力行业商业智能基础环境搭建

（1）网络容量规划方法

关于电力行业数据仓库的网络容量规划，可以分成以下三个阶段：业务需求规划、制定容量规划和容量规划执行。

第一阶段：业务需求规划。

第一阶段主要包括识别关键业务，识别造成影响的技术因素，制定数据收集清单，制定基础设施配置清单，识别约束条件和限制条件，安装和配置数据收集工具，对确认的数据指标进行收集等内容，如图 13-19 所示。

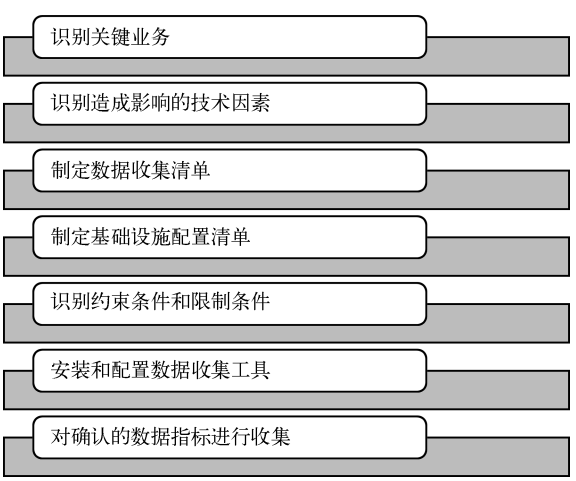


图 13-19　业务需求规划

第二阶段：制定容量规划。

第二阶段主要包括确定容量规划方法，使用不同的容量模型，决定当前和未来容量管理的优先级，提出容量管理的改进计划，如图 13-20 所示。

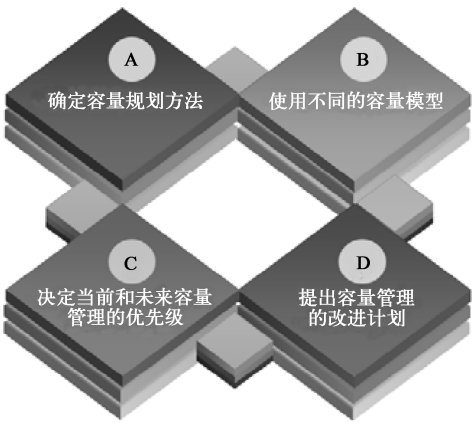


图 13-20　制定容量规划

第三阶段：容量规划执行。

第三阶段主要包括审核容量规划方法，建立沟通机制；建立容量规划的行动方案；执行容量规划；跟踪容量规划的结果，及时调整规划，如图 13-21 所示。

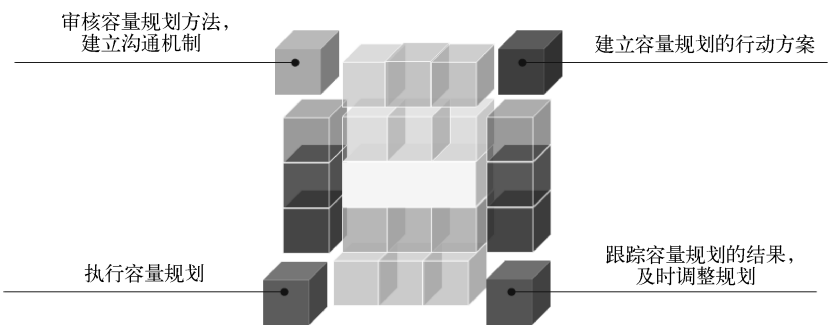


图 13-21　容量规划执行

（2）数据中心服务器逻辑拓扑图

关于电力公司数据中心服务器逻辑拓扑如图 13-22 所示。

（3）数据仓库相关存储估算

数据仓库存储容量包括 4 个部分：数据仓库容量、数据集市容量、ODS 容量和备份空间，如图 13-23 所示。

1）数据仓库容量：包括数据、索引和归档日志等信息。

2）数据集市容量：包括数据、索引和归档日志等信息。

3）ODS 容量：包括数据、索引和归档日志等信息。

4）备份空间：主要包括数据仓库、数据集市、ODS 在磁盘阵列上的备份空间。

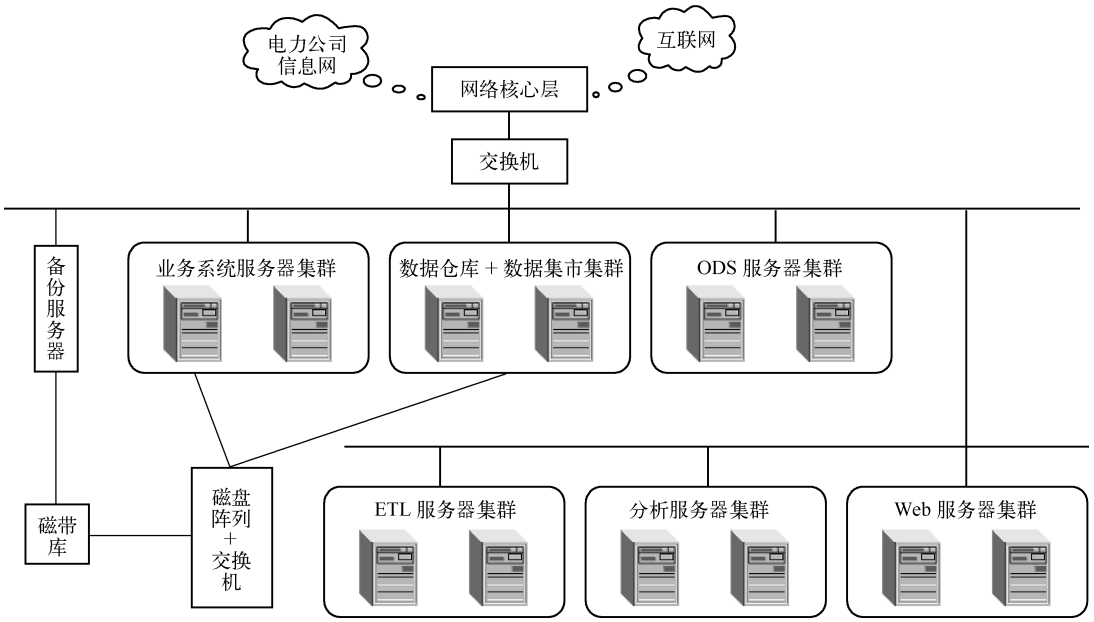


图 13-22　电力公司数据中心服务器逻辑拓扑图

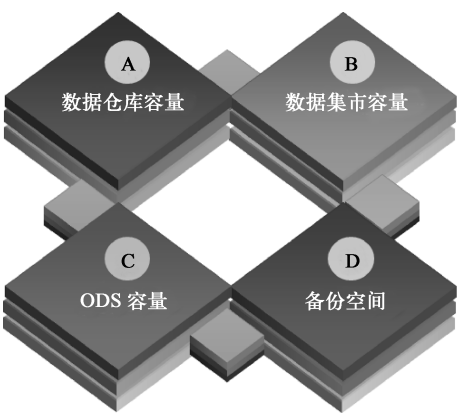


图 13-23　数据仓库总体存储

对于数据仓库的容量估算，举例见表 13-1。

表 13-1　数据仓库容量估算

编号	空 间 用 途	固 定 容 量	运 算 说 明
1	数据库系统		
2	数据库系统软件	4 GB	固定
3	数据库系统数据	4 GB	固定
4	数据库回滚段	16 GB	固定
5	数据库临时表空间	10 GB	固定
6	数据仓库数据		

（续）

编号	空 间 用 途	固 定 容 量	运 算 说 明
A1	目前数据	1 × K（K 为业务数据总量）	数据仓库是企业级范围内经过整合后的存储体，容量应该等于或者大于业务数据的总量
A2	目前索引	0. 3 × K	数据仓库的索引一般为数据仓库数据量的 30% 左右
A3	数据仓库目前数据总量	A1 + A2	
A4	每年增长数据总量	A3 × N%	N 为业务数据年增长率
A5	10 年的数据仓库总量	A3 + A4 × 10	

对于数据集市的容量估算，举例见表 13-2。

表 13-2　数据集市容量估算

编号	空 间 用 途	固 定 容 量	运 算 说 明
1	数据库系统		
2	数据库系统软件	4 GB	固定
3	数据库系统数据	4 GB	固定
4	数据库回滚段	16 GB	固定
5	数据库其他数据	10 GB	固定
6	数据库备份临时空间	16 GB	固定
7	数据集市数据		
B1	目前数据	0. 4 × A1	数据集市的当前数据容量约等于数据仓库当前数据的 40%
B2	目前索引	0. 5 × B1	数据集市的索引约占数据量的 50%
B3	数据集市当前数据总量	B1 + B2	
B4	每年增长数据总量	B3 × N%	
B5	10 年的集市总量	B3 + B4 × 10	

对于 ODS 的容量估算，举例见表 13-3。

表 13-3　ODS 容量估算

编　　号	空 间 用 途	固 定 容 量	运 算 说 明
1	数据库系统		
2	数据库系统软件	4 GB	固定
3	数据库系统数据	4 GB	固定
4	数据库回滚段	16 GB	固定
5	数据库其他数据	10 GB	固定
6	数据库备份临时空间	16 GB	固定
7	ODS 数据		
C1	目前数据	K × 5%	5%：日数据变动量占业务数据总量百分比
C2	目前索引	0. 2 × C1	ODS 的索引约占数据量的 20%
C3	ODS 当前数据总量	C1 + C2	
C4	每年增长数据总量	C3 × N%	
C5	10 年的 ODS 总量	C3 + C4 × 10	

关于磁盘备份空间的需求，见表 13-4，其中全备份不保存在磁盘阵列上。

表 13-4　磁盘备份空间的需求

应用服务器	容　　量	增量备份频率	增量备份数据量	一级备份（磁盘阵列）
数据仓库	A5	每日	A5 ×5%	A5 ×5% ×7 + A5 ×0
数据集市	B5	每日	B5 ×5%	B5 ×5% ×7 + B5 ×0
ODS	C5	每日	C5 ×5%	C5 ×5% ×7 + C5 ×0

（4）电力行业商业智能系统相关服务器描述见表 13-5。

表 13-5　电力行业商业智能系统相关服务器

服　务　器	目　　的	说　　明
数据仓库服务器	目的是提供数据仓库数据的存储、计算、查询、汇总等功能	要求数据的存储容量大，复杂的数据查询可能会影响 CPU、内存、I/O 的整体性能
数据集市服务器	目的是提供数据集市数据的存储、计算、查询、汇总等功能	数据访问较多，需要的存储容量较大，复杂的数据查询可能会影响 CPU、内存、I/O 的整体性能
ODS 服务器	目的是提供 ODS 数据的存储、计算、查询、汇总等功能	数据访问较多，需要的存储容量较小，复杂的数据查询可能会影响 CPU、内存、I/O 的整体性能
ETL 服务器	安装 ETL 软件，提供数据抽取、清洗、转换功能	因为聚合、计算、匹配等操作，所以需要高性能的 CPU 和内存
分析服务器	安装商业智能软件，同时提供各种分析、报表、查询等功能	因为有大量并发用户的请求和各种逻辑处理，所以需要高性能的 CPU 和内存
Web 服务器	目的是处理 Web 客户端的请求	因为有大量并发用户的请求和多个在线的 Web 服务请求，所以需要高性能的 CPU 和内存

7. 电力行业数据仓库建设难点

电力行业数据仓库建设的难点主要包括缺乏统一的数据规划、缺乏统一的数据管理标准体系、缺乏统一的编码管理、缺乏对数据仓库建设的验证过程等方面，如图 13-24 所示。

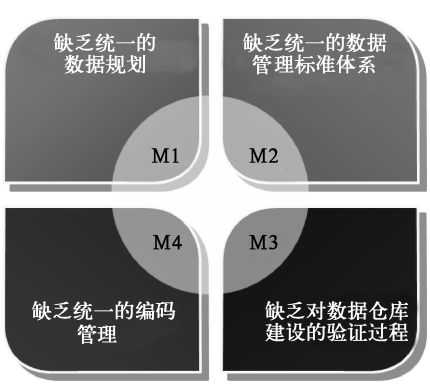


图 13-24　电力行业数据仓库建设难点

● 缺乏统一的数据规划

电力行业下级单位缺乏统一的标准体系。因为各自的建设，所以很容易形成信息孤岛。

● 缺乏统一的数据管理标准体系

电力行业总部层面缺乏统一的数据管理标准体系，没有对应的管理机构和方法去协调新增的数据需求。

● 缺乏统一的编码管理

电力行业缺乏统一的编码管理，导致数据存在不统一、不完整的现象。同时数据集成的成本很高，数据质量偏低。

● 缺乏对数据仓库建设的验证过程

电力行业普遍缺乏对数据仓库建设的验证过程，包括对试点单位的推广和建立相应的管理机制等，提高了整个电力行业数据仓库建设的风险概率。

8. 数据仓库的总体建设策略建议

1）电力行业省级单位对数据分析的需求具有一定差异性。除了有整个电力行业共性统一的内容，也允许存在个性化的内容，我们在技术架构统一的前提下，允许不同的省级单位使用不同的平台软件。

2）电力公司总部对数据仓库的建设应该有一个统一的数据标准体系，它可以帮助省级单位建立各自的数据管理体系，保证总部和省级单位数据的可用性。

3）可以通过对试点省级单位的成功推广，减少其他单位实施数据仓库的风险，也就是通过典型成功案例经验的指导，在全国范围内进行数据仓库建设。

总之，电力行业数据仓库的实施策略是以降低风险为原则，通过试点建设积累经验和方法，形成统一的数据模型标准、管控方法和数据仓库体系架构，然后向其他省级单位推广。这样可以保证整个电力行业数据仓库建设的有序开展。

举例来说，首先通过试点的建设，对数据模型进行规划，提供逻辑模型和物理模型，制定数据标准管理机制，建立数据仓库体系架构和数据质量管理策略。

然后经过一系列的经验验证，形成统一的数据模型标准、数据仓库统一体系架构以及各种数据标准管理机制等。

最后进行宣传推广和执行督导。数据仓库的开发流程是以业务需求驱动为导向的滚动式开发，以全局观点为基础的不断完善的闭环流程，如图 13–25 所示。

图 13–25　数据仓库的总体建设策略建议

13.3 电力行业数据架构

电力总公司 ODS 的功能与省电力公司的 ODS 相同，主要区别在于数据源的不同。总公司的 ODS 数据源主要是部署在电力公司总部的业务系统数据源。电力总公司数据仓库的数据源主要包括电力公司总部业务系统的明细数据、省电力公司数据仓库上报的数据等。

电力总公司的数据集市主要基于企业宏观发展的分析应用，包括可以跨系统、跨省市地对数据进行全面宏观的分析，同时也聚焦于企业的管理。省电力公司的数据仓库主要基于对省级单位的数据整合和历史数据存储。这些数据主要是细节性的、低级别的信息。根据分析需求，建立汇总数据。同时为数据集市提供整合后的、高质量的数据。

省电力公司的数据集市是针对特定的、某个主题域的数据集合。这些数据可以快速地被访问。电力行业总体的数据架构如图 13-26 所示。

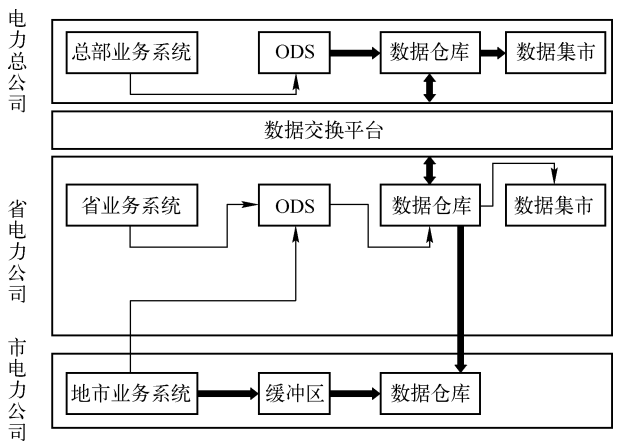


图 13-26　电力行业总体的数据架构

1. 针对电力行业的数据管理

数据管理是数据架构的基础，它决定了数据的可用性和价值。

1）数据管理保证数据的质量，确保数据的可用性。

2）数据管理将数据、数据使用者、数据的管理机构整合到一起。

3）数据管理的内容主要包括数据质量管理、数据标准管理和数据安全管理，如图 13-27 所示。

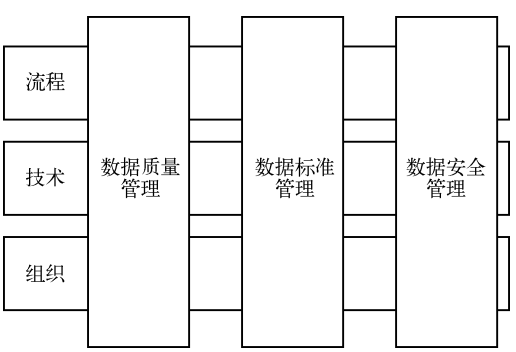


图 13-27　针对电力行业的数据管理

4）数据管理主要考虑流程、技术和组织。

2. 数据质量管理

数据质量管理主要包含以下几个方面的内容：

（1）数据质量管理定义

通过制定电力公司数据质量的衡量指标，评估数据在使用过程中的质量问题。寻找数据质量产生的根源，利用相关的工作流程解决数据质量问题，以保证电力公司的数据质量不断提高。

（2）数据质量管理目标

分析业务需求对于数据质量管理的要求，降低因为数据质量问题而导致的决策风险，通过数据质量的不断提升和改进，建立相应的管理机制和数据质量问题处理流程。

（3）数据质量管理原则

1）数据质量管理需要数据创建人员、使用人员和维护人员之间的通力合作。

2）电力公司应该学习先进的外部经验，了解实施数据质量管理的必要流程。

3）可以选择部分主题进行数据质量管理试点工作。

（4）数据质量管理工作内容

1）制定数据质量管理策略，满足业务分析对数据质量的要求。

2）根据数据质量管理目标，制定数据质量管理方法。

3）执行数据质量管理流程，推进数据质量管理的分布实施。根据业务管理主题分类、分阶段进行推广。

电力公司面临的数据质量问题分类见表 13–6。

表 13–6　电力公司面临的数据质量问题分类

数据质量问题分类	说　明	示　例
数据是否完整	判断是否有足够的信息能够满足决策需求，每条信息是否完整	所有的地址是否都有邮编、个人信息中是否都有联系方式等
数据是否能够正确反映现实	数据是否能够符合实际情况	客户的邮政编码是否与目前的家庭地址一致
有无冗余数据	是否有数条记录表示同一个实体	某生产系统保留了关于设备的多个有效记录

（5）数据质量的指标类型

数据质量的指标类型见表 13–7。

表 13–7　数据质量的指标类型

指标类型	说　明	衡量标准
完整性	实体的每个属性都有明确的值，不存在“空”或“未知”的属性	字段的空值率
相关性	对于数据库中的某些实体，它们的存在可能要依赖于其他的实体	外键无对应主键的比率
唯一性	一个表中的一组属性值是唯一的	主键的重复率
有效性	实体属性的值在有效范围之内	异常值比率

（续）

指标类型	说　明	衡量标准
及时性	是否满足应用对数据的时间要求	满足时间要求的比率
非重复记录	是否存在多条记录代表同一个实体的现象	数据非重复记录比率
真实性	数据库中的实体必须与现实世界中的对象保持一致	真实数据比率
精确性	数据精度是否满足业务需求	满足业务需求对精度要求的比率
一致性	多个系统内一致数据的百分比	关于数据不同存储的比率
可理解性	数据本身的含义是否明确	数据理解的比率
可获得性	数据是否可获得，以满足业务的需求	数据可获得记录的比率

（6）数据质量的分类

数据质量的分类见表 13-8。

表 13-8　数据质量分类标准

数据质量分类标准	说　明	示　例
一致性	当有多条记录存在时，信息及含义是否保持一致	关于设备的信息在生产系统和财务系统中是否一致
时效性	从数据的创建到使用，是否满足用户对时效性的要求	数据在业务系统中从产生到使用，是否满足用户对时效性的要求
可访问性	数据是否可以被用户访问	数据是否进入到数据仓库中，并且能够被决策分析者使用和访问
可用性	数据是否是可用的和易于理解的	一个报告是否容易理解，不会产生歧义

（7）数据质量管理工作说明

数据质量管理工作的流程是数据分析人员或者管理维护人员定期提交数据质量报告，报告内容可能不断增加，随着质量管理工作的开展，报告内容将落实到各个环节中，但是数据质量管理不能代替系统的测试工作。

3. 数据标准管理

（1）数据标准管理定义

制定和维护电力公司业务经营所涉及的数据的标准。主要包括：制定标准、审核标准、执行标准、反馈数据标准。数据标准管理的对象是所有业务经营管理的数据，不包括参数型数据。

（2）数据标准管理工作目标

1）完善数据标准。

2）使用数据标准。

3）反馈数据标准。

4）更新数据标准。

5）制定相应的数据标准管理机制，包括相应的岗位职责、工作模板等。

（3）数据标准管理指导原则

参考国内外相关行业的标准，同时结合电力公司的实际需求，要求数据标准能够在一段时间内相对稳定，满足电力公司各个部门对数据标准的要求，而不是频繁地更改与修订。

（4）数据标准管理工作内容

数据标准管理主要工作内容包括：制定并公布数据标准，制定数据标准管理方法、管理流程、岗位职责、工作模板等。

4. 数据安全管理

（1）数据安全管理定义

电力公司数据安全管理的定义是对敏感数据建立一套完整的数据安全分级和授权机制。

（2）数据安全分级的工作目标

通过建立一套完整的数据安全分级标准，明确数据使用者和数据安全人员的工作职责及权限，同时建立相关的数据使用授权机制。

（3）建立数据安全机制的指导原则

结合相关的法律、法规和电力行业内部的标准，开展关于数据安全分级和授权的工作。根据数据使用者的职责，定义使用者的权限。该流程是包括制定、审核、颁布、执行、反馈和修正在内的闭环工作过程。

（4）数据安全分级的工作内容

通过制定相关的数据安全标准和政策，定义和维护数据的安全分级标准，建立标准的维护和更新流程，为数据的应用和管理提供安全保障。主要内容包括建立数据安全分级和数据使用授权机制，实现数据访问的安全性，同时对数据安全分级和授权机制的流程进行调整和优化。

（5）数据安全级别的划分

数据安全级别的划分见表 13–9。

表 13–9　数据的安全级别

密　级	定　义	示　例
绝密	关系到国家安全或者包含商业机密的信息，要求信息具有高度机密性、准确性、完整性、可靠性和可用性。	例如，涉及国家安全的机密信息；电力行业的战略规划、购并计划、财务信息等内容
机密	涉及电力行业运作的信息，要求保证机密性、准确性、完整性、可靠性和可用性	例如，各种产品和系统的源代码，未公开的监管数据和各种审计报告等
内部	可以在电力企业内部共享的信息，但是不能对公众开放的数据和信息，要求保证数据的完整性、准确性、可靠性、可用性	例如，业务操作流程、会议备忘录、内部通讯录等
公开	经过审核后，通过电力企业发布渠道向外公开的数据和信息，需要保证信息的完整性和准确性	电力企业网站发布的信息和公开报告

小结

（1）电力行业主要面临着如下业务挑战和技术挑战。

- 业务挑战

1）电力行业的分析系统一般仅提供简单的报表功能，功能单一，高层人员无法从全局的角度对各条业务线进行多层次的综合分析。

2）对于各个分析系统来说，它们又集中于各自的领域，不具备跨业务的分析能力，存在着数据不一致的现象，不能有效地发挥电力行业数据资产的价值。

● 技术挑战

1）分析型系统与业务生产系统耦合性较强，缺乏对全局业务分析的支持，对于相同业务数据，可能会存在不同的版本。

2）各个业务系统管理各自的数据，数据的业务含义在各个部门之间可能存在不一致的解释，数据质量也相对较低。

（2）电力行业商业智能的数据架构包括源数据层、数据抽取层、数据存储层、数据访问层和用户访问层。

（3）建设电力行业企业级数据仓库的因素主要包括业务因素和技术因素。

● 业务因素

在业务上，缺乏统一的报表与指标规范体系，缺少明确的数据责任体系。

● 技术因素

缺少规范的数据架构，导致数据分布的不合理和模型的不一致。同时数据管理不规范，缺乏企业级的数据整合和管控机制。

（4）数据仓库开发应实施以全局的观点为基础，业务需求为导向的滚动式开发方法。

（5）电力行业商业智能系统的开发流程：

● 计划
● 分析
● 设计及开发
● 测试
● 部署

（6）电力行业数据仓库系统的运维内容：

● 备份与恢复
● 归档与恢复
● 系统监控
● 容量规划
● 性能管理

（7）数据仓库的总体建设策略建议：

1）电力公司省级单位对数据分析的需求具有一定差异性。除了有整个电力公司共性统一的内容，也允许存在个性化的内容，我们在技术架构统一的前提下，允许不同的省级单位使用不同的平台软件。

2）电力公司总部对数据仓库的建设应该有一个统一的数据标准体系，它可以帮助省级单位建立各自的数据管理体系，保证总部和省级单位数据的可用性。

3）可以通过对试点省级单位的成功推广，减少其他单位实施数据仓库的风险，也就是通过典型成功案例经验的指导，在全国范围内进行数据仓库建设。

（8）电力总公司 ODS 的功能与省电力公司的 ODS 相同，主要区别在于数据源的不同。

总公司的 ODS 数据源主要是部署在电力公司总部的业务系统数据源。电力总公司的数据仓库的数据源主要包括电力公司总部业务系统的明细数据、省电力公司数据仓库上报的数据等。

（9）针对电力行业的数据管理：

数据管理是数据架构的基础，它决定了数据的可用性和价值。

1）数据管理保证数据的质量，确保数据的可用性。

2）数据管理将数据、数据使用者、数据的管理机构整合到一起。

3）数据管理的内容主要包括数据质量管理、数据标准管理和数据安全管理。

4）数据管理主要考虑流程、技术和组织。

（10）数据质量的管理主要包含以下几个方面的内容：数据质量管理定义、数据质量管理目标、数据质量管理原则、数据质量管理工作内容等。

技 术 词 汇

1）企业战略：企业战略是对企业发展目标，包括达成目标的方法和途径的总体谋划。

2）企业业务战略：企业的业务战略是指企业拥有的所有资产，通过多种方式进行有效的运营，以实现利润的最大化和资本的增值。它强调了企业在各自的生产领域中的发展之道，包括如何创造价值，并且以更好的服务去满足客户，这是企业业务战略的核心和重点。

3）企业 IT 战略：企业的 IT 战略是指在充分研究企业的发展愿景、业务策略和管理的基础上，形成信息系统的远景、组成架构、逻辑关系等，以支撑企业战略目标的实现。

4）企业架构：企业架构实质上就是对企业多角度的一种描述，它反映了企业的业务流程、技术的组织和安排，是对企业关键性业务和技术的整体性描述。

5）IT 架构：IT 架构是对企业系统的 IT 规划，是建立企业信息化系统的综合性的蓝图，IT 架构可以帮助企业获得最优的投资回报，同时实现业务和技术接口之间的标准化，保证企业运营和企业战略之间的一致性，IT 架构又是承接 IT 战略与 IT 项目执行的桥梁，它主要包含应用架构、数据架构和技术架构。

6）业务架构：广义的业务架构包括产品、销售、财务、人力资源、客户服务等企业核心的业务功能和职责。并且将企业战略转化成企业运营的目标和形式，同时明确相关人员、企业资源、IT 资源和服务如何协调和部署的。我们可以说由企业战略决定了业务架构的模式，同时业务架构又是企业战略实现的手段。而狭义的业务架构包含了企业运营活动中的业务策略、组织、关键业务流程、组织架构以及人员结构等内容。

7）数据架构：数据架构是数据在信息系统中的布局与流向的框架和与数据相关的架构组件的摆放。数据是指系统所处理的所有信息和数据。而架构组件负责数据的存储、交互和应用等功能。主要内容包括数据的流向，是指数据从源系统经过各类处理、加工而到达目标系统的过程。数据架构的核心包括对数据层次的划分、数据的分布、各层次的数据模型和数据的转换等。数据架构是企业架构中最重要的组成部分之一。

8）数据分类：数据分类是按照选定的属性（或特征）区分分类对象，将具有某种共同属性（或特征）的分类对象集合在一起的过程。

9）数据大类：数据大类是从宏观的角度理解企业全局的业务情况。

10）数据小类：数据小类是在同一大类内，按照业务的特性做进一步的细分。

11）数据模型：数据模型是对数据特征的抽象，它一般分为概念模型、逻辑模型和物理模型。概念模型是以数据分类的形式体现，而逻辑模型以 ER 图的形式体现。

12）概念模型：概念模型是从业务的角度对数据进行抽象，包括业务层面上主题域的划分，以及各个主题域下的数据分类和基于分类的非功能属性。

13）逻辑数据模型：逻辑数据模型是用来发现、记录和沟通业务的详细“蓝图”，由一系列表和实体详细描述组成，是通用的业务语言，便于业务与业务之间的功能理解，遵循第三范式，包括主题域的设计、基本实体的设计和主要属性的设计，是 IT 人员和业务人员沟通的工具和桥梁。

14）物理模型：物理模型是对逻辑模型针对具体实现环境的物理化，可以不遵循第三范式，主要包括实体属性的物理化，属性的长度、类型、主键、外键、索引等详细设计。物理模型主要是描述模型实体的细节，对列的属性进行明确的定义。物理模型的建设过程是在逻辑模型的基础上，为应用生产环境选取一个合适的物理结构的过程，包括存储结构和存储方法。

15）数据分布：数据分布主要分析业务数据在多个系统之间和多个环节之间的分布情况。

16）数据流转：数据流转是描述业务分类在各个逻辑库之间的流转情况。

17）数据归档：数据归档是定期将基础数据存储、应用的数据进行归档保存，它的目的是为了保存原始数据。原则上数据归档对中间数据或者临时数据不进行归档操作。

18）数据质量管理：数据质量管理是对每个阶段里可能引发的各种数据质量问题进行识别、监控和预警等一系列的活动，通过业务管控以及技术手段，保证数据的一致性、完整性和准确性，使其数据能够准确地反映当前的业务状况。

19）技术架构的定义：技术架构是 IT 架构中比较底层的架构，它定义了如何建立一个 IT 运行环境来支持数据架构和应用架构。技术架构主要描述业务、数据、应用服务部署的基础设施能力，通过技术架构可以建立一个 IT 平台，涉及对技术的采用、基础设施的建立、产品的选择、系统的管理等方面。

20）应用架构的定义：应用架构是对实现业务能力、支撑业务发展的应用功能结构化的描述方法。系统的应用架构可以从功能和应用两个不同的视角描述系统各组件构成以及组件之间的关系。功能组件模型侧重于业务功能角度，应用组件模型侧重于应用系统设计角度。

21）数据治理分析框架的定义：数据治理分析框架主要包含两个部分，一个是数据治理管控机制，如政策、组织、流程和技术工具，另一个是数据治理涉及的领域，如数据质量管理、数据标准管理、数据生命周期管理和元数据管理。

22）数据治理的定义：数据治理是一套包含策略、原则、组织结构、管理制度、流程，并由各种相关技术工具所支撑的管理框架。数据治理是对数据管理与应用行使权力和控制的活动集合，在数据管理与应用层面上进行规划、监督和控制。数据治理为数据管理、数据应用与服务提供保障。

23）数据治理现状分析框架：主要用于帮助系统对数据治理现状进行分析，一般包括数据治理机制和数据治理领域两个部分。

24）数据治理领域：数据治理领域可以包括数据质量管理、数据生命周期管理、数据标准管理和元数据管理。

25）数据生命周期管理：数据生命周期管理根据数据在生命周期各个阶段的使用情况和需求特点，采用技术手段，对数据的存储、迁移和销毁进行统一管理，以提高系统运行的效率。数据生命周期管理的目的是对数据进行统一管理，降低数据的安全隐患和存储压力。

26）元数据管理：元数据管理是描述数据的数据，它可以帮助企业了解数据、认识数据和管理数据。

27）数据标准管理：数据标准管理是一套完整的数据规范，是数据在使用和交换过程中，为了保持数据一致性和准确性而制定的规范，它主要包括数据分类、业务标准和技术标准的详细定义。数据标准是数据治理中基本的业务和技术层面的保障。

28）大数据：大数据就是通过快速的采集、挖掘和分析，从大数据量的、多样化的数据中提取价值。形象地说，大数据就是“沙里淘金”的过程。

29）商业智能：商业智能就是利用数据仓库、数据分析和挖掘技术，以抽取、转换、查询、分析和预测为主的技术手段，帮助企业完成决策分析的一套解决方案。

30）数据仓库：数据仓库是一个面向主题的、集成的、非易失的、历史的、随着时间的流逝发生变化的数据集合，它主要用来支持企业管理人员的决策分析。

31）数据集市：数据集市就是满足特定的部门或者用户的需求，按照多维的方式进行存储，包括定义维度、需要计算的指标、维度的层次等，生成面向决策分析需求的数据立方体。数据仓库体系结构中增加了数据集市，数据集市可以看作部门级的小型数据仓库。

32）分析类数据集市是通过数据挖掘等方法帮助企业发现业务趋势，提高企业运营效率，深度挖掘数据的价值。分析类数据集市包括文本分析、数据挖掘、预测分析和可视化分析等。

33）管理类数据集市是指为了企业运营管理需要而进行的数据整合分析。管理类数据集市面向企业内部的人员，对于数据的实时性要求不高。主要包括管理驾驶舱、固定报表、OLAP 分析和 KPI。管理类数据集市主要支持对业务运营的分析。

34）研发类数据集市主要支撑各个业务部门的应用系统，满足分析需要的数据集合。

35）金融：金融就是在日常经济生活中，通过银行、证券机构等中介，从市场主体中募集资金，然后再借贷给其他市场主体的活动。可以把金融看作融资、投资和资金募集这 3 种经济活动。

36）ODS：ODS 是一个面向主题的、集成的、可变的、反映当前细节的数据集合。它主要用于支持企业处理业务应用和存储面向主题的、即时性的集成数据，为企业决策者提供当前细节性的数据，通常作为数据仓库的过渡阶段。

37）ETL：ETL 是数据抽取（Extract）、转换（Transform）、加载（Load ）的英文简写。它的一般过程是指：将源数据抽取出来，中间经过数据的清洗、转换，最后加载到目标表中。

38）OLTP：OLTP（在线联机事务处理）系统主要面向细节性的数据，存储的都是当前的数据，用来支持日常业务运作。这些数据都是可以更新的，数据处理量相对较小。

39）OLAP：OLAP（在线联机分析处理）系统主要是综合的、并且经过提炼的数据，它的数据主要是历史数据，不可修改，数据处理量相对较大，主要面向决策分析处理。

40）内容管理：内容管理主要提供对非结构化数据的存储、访问和管理的能力，包括一些凭证影像、所有格式的办公文档、XML、HTML、各类报表、图像和音频/视频信息等。

41）数据归档：数据归档就是将旧的以及不需要的数据，从数据库中复制到其他地方。

42）维度：是指人们观察事物的角度，如地区维度、时间维度、产品维度等。

43）层：根据维度细节程度的不同，划分数据在逻辑上的等级关系，用来描述维度的各个方面。例如，时间维度包括年、季度、月、日等层次，地区维度包括国家、省、市、县等层次。

44）维度的成员：维度的取值，即维度中的各个数据元素的取值。例如，地区维度中具体的成员有英国、法国、德国。

45）钻取：通过变换维度的层次，改变粒度的大小。它包括向上钻取（Drill Up）和向

下钻取（Drill Down）。向上钻取是将细节数据向上追溯到最高层次的汇总数据。向下钻取是将最高层次的汇总数据深入到最低层次的细节数据中。

46）旋转：通过变换维度的方向，重新安排维的位置，例如行列互换。

47）切片和切块：在一个或者多个维度上选取固定的值，分析其他维度上的度量数据。如果其他维度剩余两个，则是切片；如果是 3 个，则是切块。

48）度量：多维数据的取值。例如，销售额、利润。

49）ROLAP：是基于关系数据库的 OLAP，以关系型数据库为基础，对多维数据的存储。

50）MOLAP：是基于多维数据库的 OLAP，其中切片、切块是主要技术。

51）HOLAP：是基于关系型和多维矩阵型等混合型的 OLAP 实现。

52）数据挖掘：数据挖掘（Data Mining）就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的但又是潜在有用的信息和知识的过程。

53）流数据：流数据是一组顺序、大量、快速、连续到达的数据序列。一般情况下，数据流可视为一个随时间延续而无限增长的动态数据集合。

参 考 文 献

[1] 王飞，刘国峰．能深入浅出——Cognos，Informatica 技术与应用［M］．北京：机械工业出版社，2012.

[2] 维克托·迈尔－舍恩伯格，肯尼恩·库克耶．大数据时代［M］．周涛，等译．杭州：浙江人民出版社，2013.

在线互动交流平台
官方微博：http://weibo.com/cmpjsj
豆 瓣 网：http://site.douban.com/139085/
读者信箱：cmp_itbook@163.com

数据架构与商业智能

作者简介

王飞，吉林大学硕士毕业，曾在电力行业从事多年的数据仓库架构设计、数据模型设计、数据库设计开发等工作，积累了丰富的项目经验和理论知识。目前在央行从事 IT 架构咨询的工作。主要著作包括《商业智能深入浅出——大数据时代下的架构规划与案例》，《商业智能深入浅出——Cognos，Informatica 技术与应用》等。

51CTO学院推荐

地址：北京市百万庄大街22号
邮政编码：100037
电话服务
服务咨询热线：010-88361066
读者购书热线：010-68326294
010-88379203
网络服务
机 工 官 网：www.cmpbook.com
机 工 官 博：weibo.com/cmp1952
金 书 网：www.golden-book.com
教育服务网：www.cmpedu.com
封面无防伪标均为盗版

机械工业出版社微信公众号
计算机分社微信服务号

上架指导 计算机/数据架构

ISBN 978-7-111-50289-0

策划编辑◎ 丁 诚 / 封面设计◎ 子时文化 ZiShi Culture

ISBN 978-7-111-50289-0
9 787111 502890
定价：69.00 元