商业银行基于数据仓库的数据集市的解决方案

翟建国，李文生
北京邮电大学计算机科学与技术学院，北京（100876）
E-mail:zhaijg2008@gmail.com

摘　要： 在数据仓库建设的后期，由于企业的多个部门都开始使用数据仓库所提供的服务。而这些部门之间往往又对数据仓库有不同的要求，很多人就提出了在数据仓库基础上构建数据集市的概念。虽然在业务系统中拥有足够的信息来进行数据分析，但是很难从这些系统中提取有意义的信息而且速度也较慢。而且虽然这些系统可支持预定义报表，但却无法支持一个部门对数据进行历史的、集成的、智能的和易于访问的分析。因为业务数据分布在许多跨系统和平台的表中，而且通常是“脏的”，包含了不一致和无效的值，使得分析难于进行。数据集市可以合并不同系统的数据源来满足业务信息需求。本文针对某商业银行数据仓库建设本身的特点，提出了一整套切实可行的针对这家银行的数据集市建设的解决方案，并成功应用于实际项目中，希望对其他银行或者企业在建设自己的数据集市时具有参考意义。

关键字： 数据仓库，ODS，ETL，数据集市。

0. 引言

某商业银行数据仓库的建设目前已经初具规模，随着相关数据逐步积累，数据仓库的资源已经完全可以满足各类分析系统的数据需求，按照这家银行的技术架构，数据集市的建设已经迫在眉睫。

数据集市的建立可以解决各系统数据获取得问题，并在集市范围内得到最大的共享，数据集市可以根据各系统的数据需求，建设成一个高度整合的业务系统数据平台，经过整合后的数据便于应用的掘取、操作，发挥最大的效能。

数据集市的建立可以有效缓解数据仓库时间窗口的争用问题，加快各系统数据获取的速度。同时整合的数据集市不仅可以解决信息不对称等方面的问题，且整合后方便统一管理，降低费用，提升效能、增进安全，使得整体架构更加合理化、科学化。

1. 数据仓库技术

1.1 数据仓库介绍

数据仓库技术是数据库与人工智能两项计算机技术相结合的产物，是当今信息管理技术的主流，它利用人工智能中的机器学习、知识处理和神经网络等方法，从数据库中挖掘有用信息、发现知识、搞清数据库中大量数据之间的相互关联及众多数据之间存在的规律，可供使用者有效分析和利用数据库中的已存储的各种有用信息，以帮助决策者迅速做出决策，从而使企业产生巨大的经济效益。

数据仓库是“面向主题的、集成的、稳定的和随时间变化的数据集合，主要用于决策制定”（Bill Inmon，Building the Data Warehouse）。数据仓库的这些特点决定了它与传统的面向事物处理的数据库有着本质的不同。作为一个新兴的研究领域，数据仓库发展得很快，许多大学和公司都在这个领域进行着广泛深入的研究，其中尤以斯坦福大学、IBM Almaden 研究中心、威斯康辛大学、微软和 AT&T 的研究最具代表性。[1]

1.2 基于数据仓库建立数据集市

随着反洗钱系统、运行内控系统、审计系统等项目的启动，各个系统的技术架构必然都

统一规划为从数据仓库获取数据。

数据仓库的建设是以数据的整合、历史存储为主，数据本身无法直接满足各类应用的需求，且过多的应用直接访问数据仓库并发等需求对数据仓库造成的负荷将越来越重，按照总体架构科学化、合理化以及本着行内应用架构统一规划的大原则，因此必须要建设统一的数据集市，包括统一的 ETL 过程、统一的 ODS 和统一的数据集市模型，以满足外围系统数据需求，规范应用的系统框架。

按照这家银行技术架构，数据仓库的建设采用“自顶向下”的建设思路，即首先建设全行统一的数据仓库，在数据层面对全行数据进行整合，做到各类数据的完整、统一、准确、规范；再从业务层面，基于数据仓库建设各类应用的数据集市，满足应用的需求。

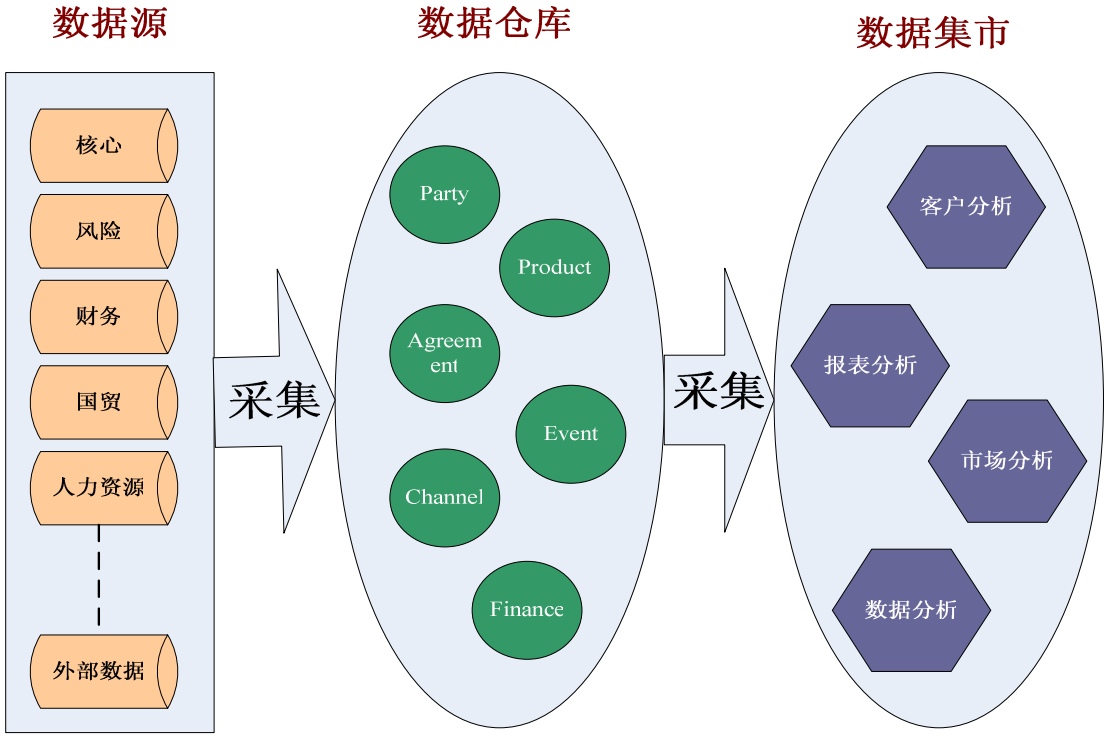

图 1 数据采集图

数据仓库的数据来源于各类业务系统及外部数据；数据集市的数据来源于数据仓库；各类应用的数据源来自数据集市。从数据的采集路径与环节看，路径清晰，环节简单，对原系统影响很小。数据集市的数据单一来源于数据仓库，系统的安全性比较高，并且避免了重复的数据整合与转换工作。[2]

因此这家银行数据集市的建立必须基于数据仓库完成。

2. 总体解决方案

2.1 系统逻辑框架

2.1.1 原业务系统逻辑框架

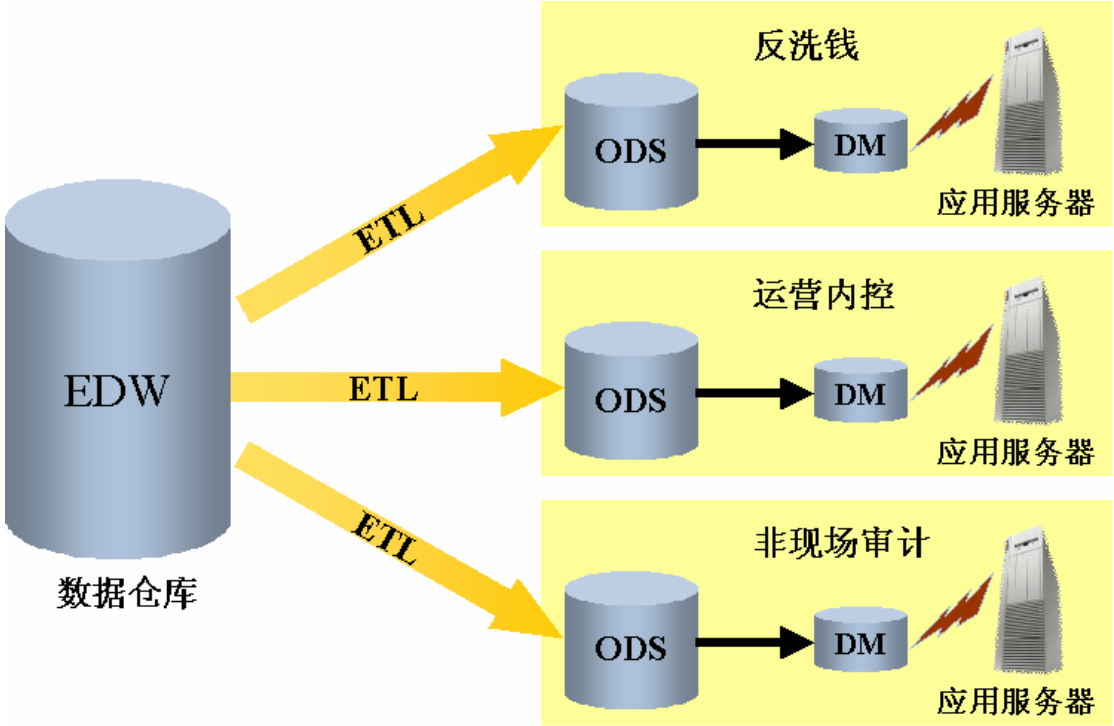

图 2 原业务系统逻辑框架

原业务系统都具有一个 ODS 数据库和 DM 数据集市，ODS 用以装载从数据仓库采集的数据，保证数据完整后再导入 DM 中供系统分析和使用，ODS 的作用是当从数据仓库中采集的数据由于各种原先出现部分数据缺失后者错误等情况时不会影响到现在应用系统的正常使用。

存在的问题：

1）数据仓库窗口征用，各业务系统数据来源都来自数据仓库，且数据量大，将大量占用数据仓库留给业务系统抽取数据的时间窗口，并且各应用抽取数据中存在部分重复部分，若能将 ODS 进行整合，势必减少数据抽取总量，减少数据抽取总时间花费；

2）繁多的 ETL 过程，对于每个应用必须单独开发 ETL 工程为应用供应数据，当应用不断增多，开发和维护 ETL 工程将变成不小的负担；

3）ODS 数据冗余，各业务系统都拥有一个 ODS 数据空间，且以面向应用的形式存储，造成空间的冗余；

4）各 DM 存储的数据未最大程度利用，各业务系统需要用到一些相近甚至相同的数据，包括衍生数据，所以各业务系统的 DM 中必然都存储着这些数据的”副本”，如果能够互相访问这些数据，那么可以节省存储空间。

2.1.2 架设统一数据集市后的系统逻辑框架

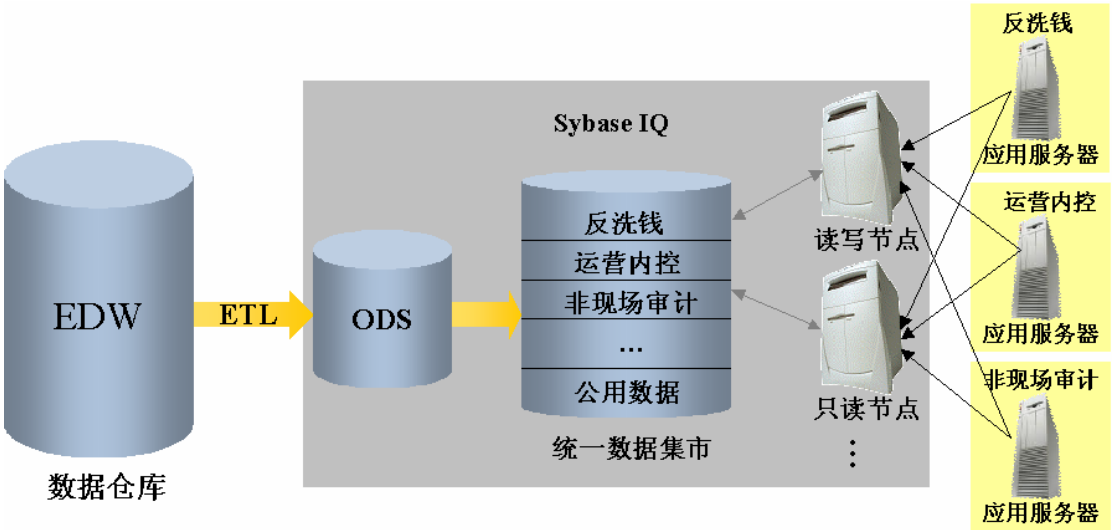

图 3 统一数据集市的系统逻辑框架

建设统一的 ODS 逻辑数据结构，从数据仓库中采集的数据首先将进入统一的 ODS 数据空间中，数据验证通过完整性和一致性后再进入统一数据集市，统一数据集市将数据在逻辑上分为各应用系统特有数据和应用系统公用数据，在物理上存储在统一的数据空间中，而应用系统的数据源直接应用于统一数据集市上。

此框架中统一数据集市采用 SybaseIQ，通过其特有的功能，可以设置多个数据库引擎（节点）连接同一个数据存储，而各个逻辑数据部分通过不同的用户进行访问，由于 SybaseIQ 仅支持 1 个可读写节点，所以需要将应用系统中的写操作和读操作分别连接相应的节点以达到良好的性能，后期可以扩展多个只读节点，解决 CPU 和内存的瓶颈问题，使前端应用系统的查询等功能得到更快的响应。

解决的问题：

1）减少数据仓库窗口占用时间，整合 ODS 以后，装载数据的过程被统一成 1 个，当后续增加新的系统应用时，只需要扩展 ODS 模型结构，整个 ETL 过程仍然是唯一的；

2）减少 ODS 的冗余，整合 ODS 后，原先的多个 ODS 重复数据占用的存储空间被释放，这样集中而单一的数据通过转换供应给各应用

3）最大程度的利用数据资源，整合 DM 以后，将划出公用数据部分，供所有应用系统共同使用，这部分数据在存储中仅有 1 份，结合各应用特色部分数据，确保了各应用系统的数据是完整的，在 SybaseIQ 中，各部分数据通过不同的用户来区隔，通过灵活配置用户权限来达到安全性的要求，这种结构同时可以很好的利用 SybaseIQ 在 I/O 方面的优势；

4）良好的扩展性，当后续增加应用时，仅需要扩展 ODS 和统一数据集市以及 ETL 过程，并不会影响到原有系统。

2.2 数据集市模型概述

统一数据集市模型将基于反洗钱、运营内控、非现场审计 3 个应用的数据需求进行概要设计公用数据模型，未来在应用的扩充上以数据需求为驱动在模型上进行丰富和扩展，公用数据模型设计的原则是以后续应用的增加很好的融入这部分模型为目的。

数据集市模型分为 4 个部分：

公用数据部分，主要包括：客户信息、帐户信息、交易流水等公用属性部分，以及地址

(国家、地区、省份)、产品、机构和代码表等；

运营内控特有数据部分，主要包括：指标数据、预警结果、预警规则配置数据等；

非现场审计特有数据部分，主要包括：统计数据、审计预警等；

反洗钱特有数据部分，主要包括：指标结果、衍生数据等。

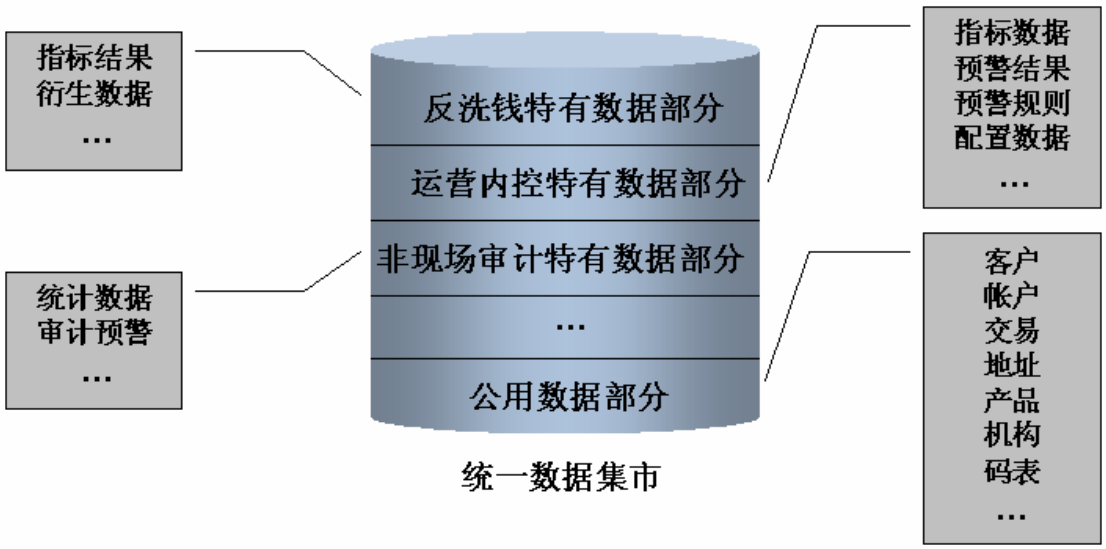

图 4 数据集市模型

2.3 系统数据流程

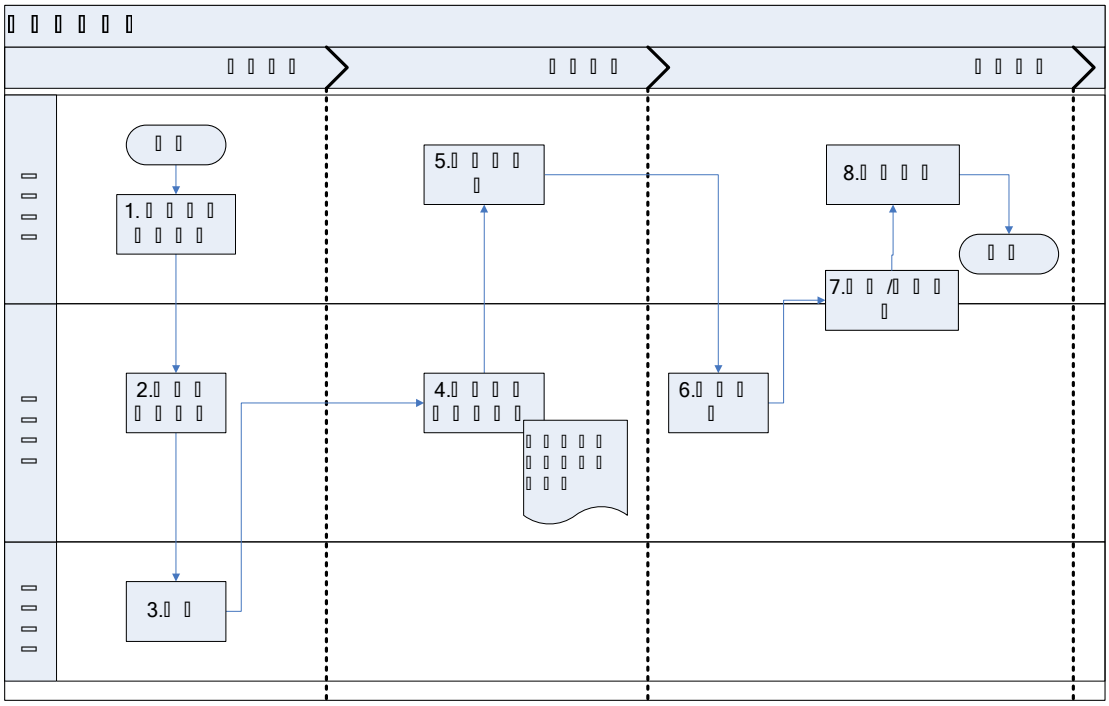

图 5 系统数据流程图

数据集市的建立是由实际的业务需求驱动的，随着各类业务需求的提出，数据集市的建立会逐渐完善。

数据集市会根据数据仓库的实际情况，对业务需求进行分析，明确业务满足度；

在数据集市需求分析期间，数据仓库提供充分的支持，并验证分析结果（业务模型与数

据仓库模型的映射）的正确性；

如果部分数据需求在当前数据仓库中没有，而在源系统中存在，建议数据仓库对这部分数据进行加载，使得获取数据的统一来源是数据仓库，在架构上做到统一；

数据集市按照分析结果，将数据整合进数据集市模型，并由业务系统验证数据的正确性；

进入生产阶段，数据集市每日会根据各个业务系统的综合需求进行数据的抽取。

2.4 数据存储方案

从数据仓库采集数据到 ODS 过程中，将尽量减少转换等操作，以减少数据仓库的征用时间，从 ODS 到统一数据集市过程中再进行衍生数据的计算等，之后数据集市保存所有应用的当前数据，而 ODS 中的数据在保存一段时间后将被删除。

数据集市数据获取以增量方式为主，对无法判断增量的数据以全量方式获取；

数据在 ODS 中保留 N 天的时间（具体保留时间需要视应用需求而定），以达到对业务系统数据提供的缓冲，同时避免各个系统补数据时反复从数据仓库中抽取数据；

每日从数据仓库获取数据在 ODS 中以切片方式保存，在系统频繁增删的情况下可以保持系统的高性能，避免产生大量数据碎片；

ODS 的数据空间要留有足够的冗余，以满足业务系统全量初始化的需求。

2.5 数据仓库到 ODS 的 ETL 数据采集方案

在应用业务数据需求通过数据仓库验证后，数据集市小组需要开发从数据仓库采集数据到 ODS 数据模型的 ETL 工程，通过优化和调试，确保 ETL 过程是高效率、高稳定的，主要目标减少对数据仓库时间窗口的占用，ETL 过程将减少转换的过程。为了便于 ETL 过程批量加载，设计和开发 ETL 过程应避免依赖性，这样在调度策略上可以更加灵活。

随着基于数据集市的应用的增多，对于资源以及时间窗口争用的矛盾会变得突出,在 ETL 过程调度策略上根据实际情况对各应用进行优先级排序，在 ETL 过程和 ODS 到数据集市的转换过程中避免应用间的依赖性，减少耦合性，这样通过 2 个方面加快应用获取数据的效率：

批量优先加载优先级高的应用所需要数据的那部分 ETL 过程；

采用在 ETL 过程和 ODS 到数据集市的转换过程分步并行的方式，以保障优先级高的应用对于数据需求时间点的要求。

例如：由于运营内控应用对于数据的需求时间点要求最高，可以优先加载运营内控需求数据的这部分 ETL 过程，一旦满足运营内控需求部分的源数据到达 ODS，便激活 ODS 到数据集市转换过程关于运营内控部分的脚本，而 ETL 过程并不停止，执行脚本的过程可以充分利用 ETL 数据加载过程中闲置的 CPU 和内存资源。

2.6 ODS 到数据集市转换方案

2.6.1 公用数据部分

由于公用数据部分不涉及业务逻辑、算法等，也不包涵过多的转换等操作，所以 ODS 到数据集市公用数据模型部分转换脚本由数据集市小组统一进行开发和验证，结果与应用之间进行确认。

2.6.2 应用特有数据部分

各应用特有数据模型部分转换涉及到业务逻辑和算法，数据集市公布 ODS 数据接口和数据集市模型接口，由各应用定制完成特有数据部分的转换，在这个过程中数据集市小组提供各应用足够的支持。

2.7 外部接口方案

2.7.1 数据仓库接口

数据仓库中数据的获取采用数据不落地方式，利用 ETL 抽取工具（行内统一采购 DataStage）将数据直接抽取进入 ODS 数据模型，经过整合的 ETL 过程避免开发和维护的紊乱。

2.7.2 外部系统接口

ODS 和统一数据集市可以建设在 SybaseIQ 内，可以更好的提高数据加载的效率。应用系统获取数据将很便捷和高效的通过脚本来实现。

2.8 实施方案

目前这家银行需要应用统一数据集市的项目有 3 个，分别是反洗钱、运营内控、非现场审计项目，所以在数据集市的建置方法上有 2 种方案供选择：

采用统一规划统一实施的原则建置，由各应用提交数据需求和数据模型，数据集市小组进行全局分析和考虑，设计规划 ODS 模型和数据集市模型，这样做的好处是模型的设计会更加成熟可靠，缺点是所有应用项目必须等待数据集市建置完成后才能付诸实施，项目的延期必定造成损失。

采用循序渐进的原则建置，以某个具体应用为先行应用的方法，初步建置数据集市，以后续陆续添加应用的方式扩展数据集市，这样做的好处是可以结合目前项目的进度进行搭建，而不需要等待所有应用提交数据需求和数据模型的时间，缺点是初期模型势必无法做到非常成熟。即便是采用这种方式，由于模型设计等时间是无法缩短的，所以需要该行根据目前应用的具体情况选择最合适的先行应用。

无论采用何种构建方式，最终都可以达成成熟的模型，需要结合该行实际情况进行选择。以下分别讨论 2 种方式。

2.8.1 统一规划统一实施

统一设计和规划数据集市的开发和实施分成 3 个大阶段：

第 1 阶段，收集各应用数据需求和数据模型，从全局角度进行分析，并与数据仓库就数据需求内容进行沟通，需要数据仓库验证是否满足应用需求，数据集市小组结合各应用的数据需求和数据模型进行统一规划和设计数据集市和 ODS 的模型；

第 2 阶段，数据集市小组开发和调试从数据仓库采集数据到 ODS 的 ETL 过程以及 ODS 到数据集市公用模型部分的转换脚本，同时公布 ODS 以及数据集市的模型，由各应用定制从 ODS 加载数据到数据集市的脚本；

第 3 阶段，将各应用挂载到数据集市上进行测试，根据各应用的反馈进行部分修改，通过稳定性观察后最终达成验收并正式运行。

2.8.2 循序渐进构建

以一个具体应用为先行应用的方案，设计和建置统一数据集市，前期数据集市设计公用数据模型和该应用特有数据模型，后续再陆续添加其它应用，这样可以减少对多个应用项目所造成的影响。前期的重点落在合理的、科学的设计数据集市中公用数据模型部分，这样在后期添加应用时对于公用数据模型部分可以控制到基本不修改或者修改幅度很小，同时还要考虑到其它尚未确定的应用将来加入统一数据集市的情况，使得成果得到最大程度的利用。

3. 系统建设目标

第一，ETL 的整合，将每个应用单独的 ETL 过程进行整合，以利于 ETL 工程的开发和维护；第二，ODS 的整合，将每个应用独立存在的 ODS 进行整合，节省存储空间资源；第三，DM 数据模型的整合，使得多个应用的数据来源一致，更加符合先进的科学架构。基于以上 3 个整合点，将统一数据集市建设成高可用性、高健全性的应用数据平台

4. 结束语

本文针对某银行的数据仓库的具体情况，提出了一套建设这家银行统一数据集市的解决方案。重点阐述了系统的逻辑框架、数据仓库到 ODS 的 ETL 数据采集方案、ODS 到数据集市转换方案等。

通过在这家银行后来的数据集市建设实施实践后，证明本方案是完全切实可行的，而且整体效率也比原来的业务系统要高很多，希望本方案会对其他银行企业的数据集市建设带来一些借鉴、指导作用。

参考文献

[1] 王彦龙 企业级数据仓库原理、设计与实践电子工业出版社 2006 年
[2] 项目组资料 某银行数据集市建设的解决方案 2007 年
[3] （美）William H.Inmon 著 王志海等译 数据仓库（第四版） 机械工业出版社 2007 年

The Data Market Solution based on Data Warehouse Technology of the Commercial Bank

Zhai Jianguo, Li Wensheng
School of Computer Science and Technology, Beijing University of Posts and Telecommunications, Beijing(100876)

Abstract

At the last stage of building the data warehouse, many departments of the enterprises all began to employ the services supported from the data warehouse and they had different requirements for the data warehouse. So a lot of people proposed the conception of building the data markets based on the data warehouse. Although the operational systems have enough information to conduct data analysis, it is difficult to extract meaningful information from these systems and a slow speed to do it. And though these systems can support predefined statements, but it can not support a department of history of data, integrated, intelligent and easy-to-access analysis. Because the business data is distributed in many Inter-system and cross-platform tables, and often it is "dirty". It contains inconsistencies and invalid value, making analysis very difficult. Data mart system can be combined in different data sources to meet the business information needs. Aiming at the features of the data warehouse in a business bank,

this paper is intent on proposing a feasible method to build the whole data market for this bank and it is successfully used in building the data market project. I wish there will be some values for some banks or enterprises when they are intending to building their own data markets.
Keywords: Data Warehouse, ODS, ETL, Data Market