航空公司数据仓库模型设计与实现

Design and Implementation of Data Warehouse in Airlines

李福娟　刘仲英　（同济大学经济与管理学院 上海 200092）

摘要：本文以航空公司为背景，结合目前主要的数据仓库的设计理论和框架，提出了针对于航空公司的数据仓库的设计与实现过程。并对数据仓库的逻辑模型、实现方法和相关技术等问题进行了探讨。

关键词：企业数据仓库 数据集市 数据抽取转换导入(ETL) 航空公司

1　引言

随着"十五"期间民航主要业务信息系统相继建设完成，民航企业信息化已进入成熟阶段，航空公司的信息系统建设已经覆盖到企业经营的各个层面。目前电子客票已在国内广泛推广，航空公司之间的代码共享合作进一步加大，航空联盟合作的步伐逐步加快，以及内部管理的精细化要求，使得航空公司对信息化的要求不再停留在满足业务处理和日常管理方面，已经上升到决策分析层面。如对订座离港的数据分析和挖掘利用，对常旅客数据的忠诚度分析、代理人的价值分析、航班延误率正点率分析、市场分析、竞争分析等。而航空公司目前的信息化建设中还存在一些问题：异构数据环境无法支持有效的数据交换；系统与系统间数据差异较大；缺乏对历史数据的有效利用；缺乏对业务规则一致的理解、认识与利用；缺少先进的数据分析手段；缺乏全面完整的决策信息平台；因此航空公司非常关注能够有一个整合的、能提供决策分析的平台来解决目前存在的主要问题。

为此，考虑按照数据仓库的理念来建立航空公司企业级数据仓库(EDW)，对一些轻量级的、近期的数据并且粒度小的数据建立操作型数据存储(ODS)，面向用户分析的数据集市。从源系统到目标数据仓库之间使用 ETL 工具进行数据交换。最后再建立语义层，用前端表报工具提供给用户进行分析。

2　数据仓库的总体设计

数据仓库的设计主要是为整合航空公司现有的和未来的信息系统以便作为各种管理分析和决策支持类系统的可靠的数据来源。数据仓库的设计对整个数据仓库项目是至关重要的。设计是把需求和关键技术融合在一起的过程，设计方案一旦确定，整个数据仓库的基本特征也就确定了。为了从总体上把握数据仓库设计的好坏，我们给数据仓库设置了如下目标作为衡量数据仓库设计的标准。

2.1　数据仓库的设计原则

(1) 共享性。数据仓库应该实现全企业的数据共享，成为一个真正的企业级数据共享平台。这种数据共享决不是仅仅停留在"让全企业的用户都可以访问到数据仓库中的数据"这一层面上，而是应该做到"让全企业的用户不仅能够访问数据仓库中的数据，而且知道如何使用数据仓库中的数据"。

(2) 正确性。使数据仓库中的数据成为全企业的权威数据。这意味着当数据不一致的情况发生时，我们应该相信数据仓库中的数据是正确的，并以之为标准。

(3) 高效性。数据仓库的本质决定了它必须为大量的用户提供海量数据的查询和分析服务。在这种情况下，保证数据仓库能够高效响应所有用户的所有请求是非常重要的，如果做不到这一点，整个数据仓库就会失去其可用性，而这恰恰是引起数据仓库项目失败的最常见原因之一。

(4) 开放性。因为不知道数据仓库中的数据将来可能被用在什么方面，因此不能对其使用方式进行限制。数据仓库的应用可能是 B/S 方式的，也可能是 C/S 方式的；可能是一个通用的 Java 程序，也可能是一种专用的分析工具。数据仓库应该支持所有这些开发方式，并公布开发接口。

(5) 安全性。数据仓库必须有严格的访问权限控制机制。由于数据仓库对全企业开放,而不同的用户和部门所能够访问的数据和应用又是不同的,因此这一机制显得尤其重要。在开放的同时保证安全是数据仓库设计的首要目标之一。

(6) 扩展性。数据仓库必须能够不断调整和扩展,以适应企业业务的变化。从某种意义上来说,数据仓库建设是一个只有开始没有结束的过程。无论是数据仓库中的数据还是用户需求,都有可能发生变化。当这种变化发生时,数据仓库应该很容易作出相应的调整。从数据仓库设计的角度来讲,对于这种灵活性和扩展性的保证主要来自两个方面:模块化的设计和元数据驱动。

2.2 数据仓库的逻辑架构

上图 1 所示为一个针对航空公司的通用的、全面的企业级数据仓库的逻辑架构。下面我们将分别对数据仓库各部件的内容和作用进行说明。

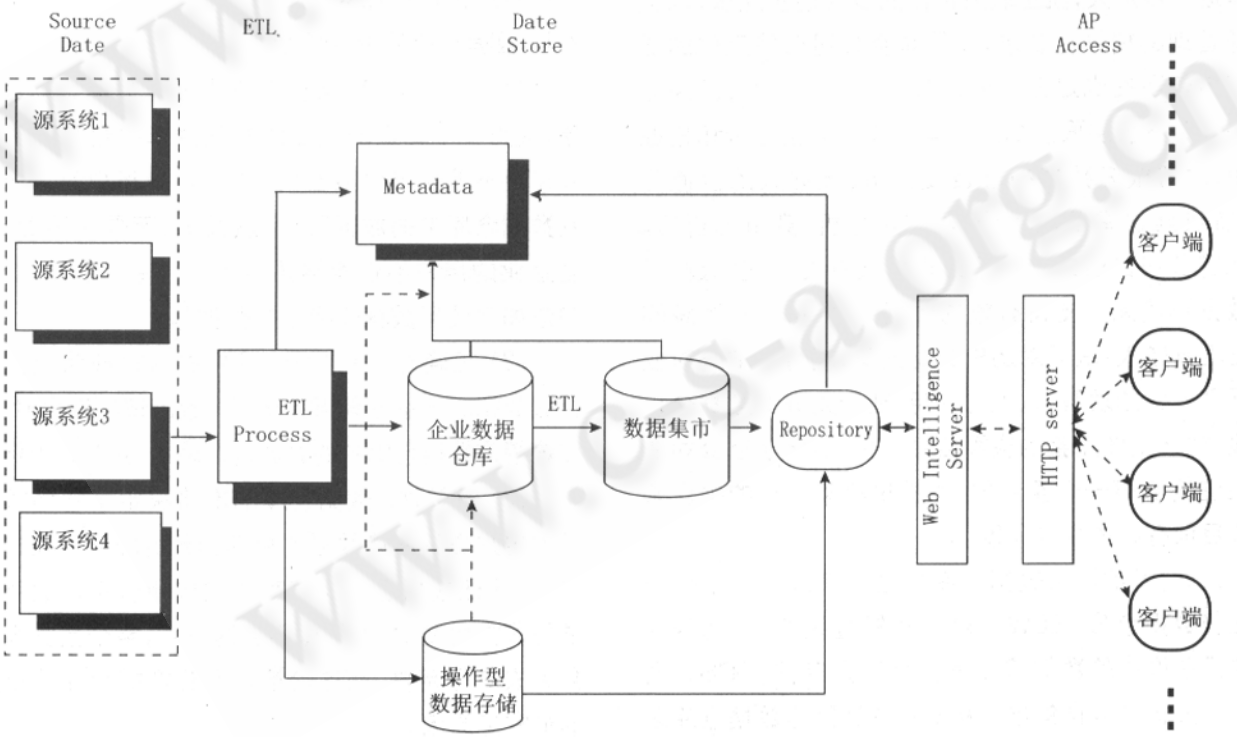

图 1　数据仓库的整体架构

(1) 源系统。指航空公司的源系统数据。这个数据可以是 ORACLE、DB2、SQL SERVER 中的关系型数据,也可以是 TXT、XML、EXCEL 等格式的非结构化数据。

(2) 数据转换流程(ETL Process)。主要指通过 ETL 工具直接获取源数据,然后对数据进行统一转换,最后加载数据到 EDW 的过程。从源系统到企业数据仓库、从操作型数据存储到企业数据仓库、从企业数据仓库到数据集市都可以通过 ETL 工具完成。每个数据表的 ETL 都被分为 Extract、Converting and Transform 和 Load 三个步骤,由 Job Scheduler 对各 Job 进行操作流程控制。

(3) 企业级数据仓库(Enterprise Data Warehouse,EDW)。作为经营分析的核心和唯一集成的数据源,保存分析粒度的数据。内含历史、当前快照及一定颗粒度数据的汇总。

(4) 面向主题的数据集市(Data Mart)。数据集市是为支持前端用户在性能、易用性、访问权限等方面的需求所采用的手段,根据应用需要而建立,数据集市的唯一数据源是 EDW,可以根据具体情况分别采用 View、Table、Cube 三种方式实现数据集市。

(5) 操作型数据存储 (Operational Data Storage)。ODS 是对运行数据存储的一组数据集合,它是以相近数据源结构,存储当前和近期数据为目的。往往存储量较小,相对于数据仓库存储量更小,并且不保

留历史。并经常用作 EDW 的临时存储区域。

(6) 前端展现工具。可采用 Business Objects、BRIO、海波龙等作为前端报表产生工具。

(7) Metadata 管理工具。Metadata 管理工具可以利用元数据指导和帮助最终用户使用数据仓库的信息，同时元数据也可用于帮助技术人员管理数据仓库的操作。

2.3 数据仓库的设计

首先需要通过数据交换工具(ETL 工具)将源系统的数据抽取、转换、导入到数据仓库，为各业务系统之间提供一个统一的数据交换通道，使数据交换更加准确、便捷、高效、通畅；依据企业数据仓库所具有的强大的数据存储和数据处理能力，能为各应用系统存储历史数据，完成海量数据的处理，比如复杂的业务统计功能、各种数据模型的计算、以及数据挖掘工作，为业务系统减负。同时，数据仓库的建设过程中，建立一系列的企业级数据标准。这些技术标准将规范和约束未来的系统建设工作，使公司的各个系统形成一个整体。

企业数据仓库与 ODS、DM 的关系如图 2 所示。

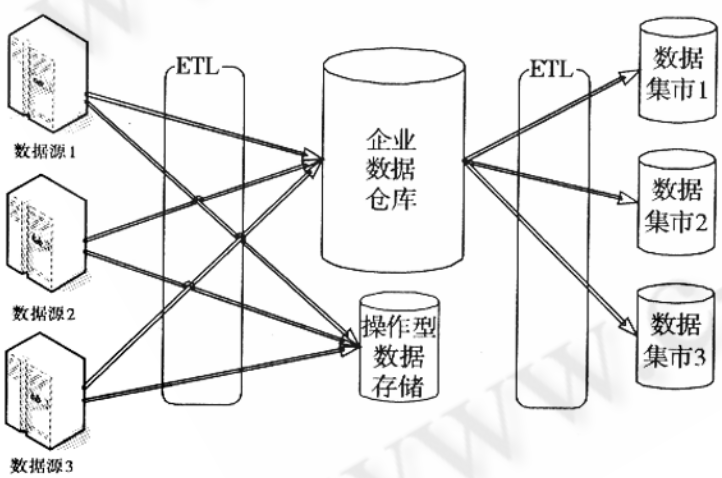

图 2　EDW、ODS 与 DM 的关系图

其中 ODS 的设计中考虑它是对当前和近期运行数据的存储，主要面对用户一些临时性或短期的查询和分析需求，因此考虑用关系型的数据结构。企业数据仓库经常采用的数据模型有：星型模型、雪花型多维模型、树型多维模型、增量型多维模型等。数据集市是面向主题的小的数据仓库，它的主要模型也是星型模型、雪花型多维模型等。从数据组织形式的角度，如果说 ODS 是一个错综复杂的“整体蜘蛛网”，那么 EDW 和 DM 就是一棵结构分明的数据树。EDW 和 DM 的拓朴结构从数据主题开始，经过若干层的子主题，到最后包含在不同子主题中的“数据单元”为止。ODS、EDW 和 DM 数据是数据仓库的三大主体数据存储区域，它们的数据模型和存储策略合在一起就构成了数据仓库核心模型中的数据模型。在数据仓库的设计阶段，所有关于数据仓库数据模型的内容都应该记录在相应的文档中。随着环境的变化和业务的发展，数据仓库的数据模型也会发生变化。因此从长远的角度讲，数据仓库的数据模型应该是记录在元数据中的，而不是记录在某个固定的数据模型文件中。

3 数据仓库在航空公司中的实现

按照航空公司的数据特点，有很多自行开发的信息系统。各个信息系统之间由于建设时间不同，标准不同，应用对象不同，导致数据有很多不一致、不完善、不统一的地方。通过数据仓库的建立，建立了公司的核心信息资源，使企业能更综合、深入地挖掘和利用企业信息资源，从而在客户服务、市场营销、经营决策、安全管理、成本控制等方面进行跨上一个新的台阶，大大提升公司的核心竞争力。

在内容上，数据仓库首先是一个整个公司的核心数据中心，同时也是由一个个数据集市组成的，即以特定业务主题的小数据仓库。在某航空公司内部，将分布构建如下面向主题的数据集市，主要包括：销售数据集市、生产运行数据集市、客户数据集市、财务数据集市、货运数据集市、人力资源数据集市等；

3.1 数据交换

ETL(Extract Transaction Load)模块涵盖了对数据交换处理的全部过程，其中从源系统至 ODS、源系统至 EDW，EDW 至 DM 的数据加载过程是整个数据仓库的主要内容之一，也是最复杂并且工作量最大的部分。

ETL 模块是对业务逻辑转换的直接体现，只有经过 ETL 处理后保存到数据库中的数据才是有价值和有效的数据。ETL 模块包括数据抽取、数据清洗、数据转换、预计算、数据汇总和数据加载等各类处理过程。

其中数据清洗是整个数据仓库的数据入口，典型的数据清洗任务有：数据验证（剔除不符合检验条件的数据），数据映射（使数据源的数据在进入数据仓库之前各项数据属性具有统一的标准）。在数据清洗完成后存在一个数据整理和转换的过程，这一过程需要对数据进行处理，使之能够适应应用模块的需要。数据转换过程可能非常复杂。要进行多维查询和分析，就必须根据不同的维度对企业的数据进行汇总，并将结果保留在 DM 中，这一过程就是 ETL 中的汇总过程。在经历了数据的抽取、清洗、整理、预计算和汇总后，我们需要将所获得的结果载入 ODS、EDW 和 DM 中。这个过程应该是定时进行的，并且不同主题的数据加载任务有各自不同的执行时间表。

ETL 的数据加载过程分两种方式：

（1）全量加载。如第一次运行 Job 需要将所有的数据加载到数据中心，执行全量加载，将目标表清空，将所有记录加载到目标表中。

（2）增量加载。从第二次运行 Job 开始，每次进行增量加载，增量加载的方式有两种：时间戳，CDC。在该航空公司的数据仓库实施过程中，这两种方式都采用到。

3.2 数据存储

数据存储包括了 ODS、EDW 和 DM 三个存储单元。在该航空公司中，ODS 存储了业务系统中直接与运营相关的、最详细的数据，包括各个系统中的静态表、运营相关的数据表，而中间结果表、查询结果表和管理日志将不会导入到 ODS 模块。

EDW 是数据仓库中高附加值数据的集合，是对 ODS 数据的进一步整合，通过 ETL 模块，采用预计算、汇总和统计等操作，在对 ODS 中的数据进行处理后根据不同的业务主题进行分类。主题是数据仓库中数据的最小逻辑单元（由若干张表组成）。EDW 模块将提供多个数据主题来组织数据，比如航班运营、票据结算、财务、客户等来重新组织数据。这些数据主题直接为前端的数据分析提供数据支持，前端应用可能查询这些数据主题中数据来生成报表或者提取数据。而对于特定的数据分析和数据挖掘应用，如果需要的数据比较复杂，范围比较广，可以为特定的应用建立一个主题，专门存放经过特定处理的数据，来完成特定的分析应用。在 EDW 模块中，数据的逻辑结构（数据模型）采用“星型拓扑结构”。EDW 主要存储大容量的数据，不直接面对最终用户访问，我们分布构建如下面向主题的数据集市，主要包括：销售数据集市、生产运行数据集市、客户数据集市、财务数据集市、货运数据集市、人力资源数据集市等，直接面对客户访问。图 3 给出了企业数据仓库关于航班运行分析的星型结构图。

3.3 数据展现

商业智能应用是以数据仓库为基础，利用多维分析和数据挖掘技术，对数据进行深层次的分析和挖掘，为公司的客户服务、市场营销、经营决策、安全管理、成本控制等管理决策提供支持。在该航空公司，商业智能应用采用 BO 报表工具，主要实现销售分析、代理人分析、运行分析、飞行员乘务员分析、客户分析等。

4 结论

数据仓库技术自从数据仓库专家 W. H. Inmon 在其著作《Building the Data Warehouse》一书中给出权威定义后已经发展数年。期间国内很多企业包括航空公司都在实施数据仓库项目，但总体来说还是实施失败的居多数。失败的主要原因有：把数据仓库看成是一个纯粹的技术项目；数据仓库的数据模型不清，最后建成的整个数据仓库只不过是一个数据和应用的大集中，用户也没有享受到数据共享所能够带来的种种好处；速度奇慢，导致用户无法忍受而拒绝使用数据仓库；使用不成熟的技术或者在一开始就开发复杂性过高的应用，导致数据仓库实施周期过长，开发完成后由于业务环境发生显著变化而无法使用等等。

数据仓库系统是一个关系复杂、组件众多、牵涉面广泛的系统，要管理好这样一个系统，完善的管理手段是非常重要的。如，建立和完善起相应的一系列业务处理和管理流程；数据标准和规范也是确保数据仓库管理工作能够按计划进行的重要手段。

参考文献

1 李秀、廖磷、刘文煌，基于 Web 的数据仓库系统的研究[J]，计算机工程，2001，27(11)：44～46.
2 王红，航空公司客运收益管理系统数据仓库的设计[J]，计算机应用与软件，2004，21(6)：49～50.
3 Bill Inmom，数据仓库，机械工业出版社.
4 彭木根，数据仓库技术与实现，电子工业出版社.