****

****

**数据库新技术课程报告**

|  |  |
| --- | --- |
| **论文题目** | 数据仓库技术报告 |
| **作者姓名** | 刘子凡 |
| **学 号** | DZ1533016 |
| **学科、专业名称** | 计算机科学与技术 |

**2018年5月18日**

# 目 录

[第一章 目 录 2](#_Toc515977139)

[第一章 数据仓库简介 3](#_Toc515977140)

[1.1 背景介绍 3](#_Toc515977141)

[1.2 诞生 3](#_Toc515977142)

[1.3 定义 4](#_Toc515977143)

[第二章 数据仓库体系结构与组成 6](#_Toc515977144)

[2.1 两层的体系结构 6](#_Toc515977145)

[2.2 三层的体系结构 8](#_Toc515977146)

[2.3 数据仓库的组成 9](#_Toc515977147)

[第三章 ETL过程 11](#_Toc515977148)

# 数据仓库简介

## 背景介绍

人类进入信息时代以来，特别是近20年来，数据规模日益扩大，数据量呈爆炸式增长。例如电商管理者需要从日积月累的交易信息中发现商品畅销的规律以及预测未来的趋势；医生需要从大量病人的电子病历中发现某些疾病的起因、症状；天文学家需要从天文望远镜获取的观测数据（其规模）中发现新的遥远天体及其运动规律等。这些数据的共同特点有：一是数据量巨大，一般都是TB级别以上；二是均以结构化的形式存储在数据库中，包含了大量潜在的、有价值的知识，有的已被发现，而有的还未被发现。这样的知识获取，仅仅依靠数据库管理系统的查询检索机制和统计分析方法，已经远远不能满足实际需求。为了有效地管理和利用这些海量数据、发现其中潜在的知识，数据仓库应运而生。

## 诞生

数据仓库是建立在传统事务型数据库基础之上，为企业决策支持系统以及数据挖掘系统提供数据源的技术。迄今为止，国外的数据仓库已经发展了二十几年，虽然国内起步较晚，但是发展十分迅速，目前已有众多的大型公司或企业正在建设或已经建设不同规模的数据仓库。

传统数据库和数据仓库的最根本区别在于其侧重点的不同。数据处理分为两大类：事务型处理，又称联机事务处理（Online Transaction Processing,OLTP）；分析型处理，又称联机分析处理（Online Analytical Processing,OLAP）。事务型处理以传统数据库为中心进行企业日常的业务处理；分析型处理以数据仓库为中心分析数据背后的关联和规律，为企业决策提供可靠、有效的科学依据。事务型处理和分析型处理的分离，划清了数据处理的分析型环境与事务型环境之间的界限，从而由原来以单一数据库为中心的数据环境，演变为以数据库为中心的事务处理系统和以数据仓库为基础的分析处理系统。企业的生产环境也从以数据库为中心演化为以数据库和数据仓库为中心。因此，在事务处理环境中直接构建分析处理应用是不合适的，要提高分析和决策的效率和有效性，分析性处理及其数据必须与事务型处理及其数据相分离，必须把分析型数据从事务型处理环境中提取出来，按照决策支持的需要重新组织，建立相对独立的分析处理环境，数据仓库正是为了构建这种新的分析处理环境而出现的一种数据存储、组织和处理技术。

## 定义

数据仓库的概念最早出现在20世纪80年代。1996年，被称为“数据仓库之父”的William H. Inmon系统地阐述了数据仓库的定义，即一个面向主题的、集成的、不可修改的且随时间变化的数据集合，用以支持管理人员的决策。

面向主题是相对于传统数据库的面向应用而言。所谓面向应用，是指系统实现过程中主要围绕着一些应用或功能，而面向主题则是考虑一个个的问题域，对问题域涉及的数据和分析数据所采用的功能给予同样的重视。数据仓库是面向在数据模型中已定义业务的主要主题域的，例如在“商场”这样一个领域中典型的主题域包括商品、供应商、顾客等。

集成是指数据仓库中的数据来自不同的数据源。由于历史的原因，个数据源的组织结构往往不同，在这些异构的数据导入到数据仓库之前，需经历一个集成过程。在数据仓库的所有特点中，这是最重要的。应用系统的设计人员经历多年制定出来的不同的设计策略有很多种不同的表示方法，在编码、命名习惯、属性和属性度量等方面往往是不一致的。例如“客户性别”的编码，在数据仓库中的编码为“男/女”还是“m/f”并不重要，重要的是无论是什么原始应用系统，在数据仓库中应该编码一致。如果应用系统中编码为“X/Y”，则在其导入数据仓库时就应进行转换。对所有应用都要考虑一致性，如命名习惯、键码结构、属性度量以及数据特点等。

不可修改是指与面向应用的书屋数据库需要对数据进行频繁的插入、更新操作不同，数据仓库中对数据的操作仅限于数据的初始导入和查询。操作型数据一般是一次访问和处理一条记录，也可以对操作型数据进行更新。但数据仓库中的数据则不然，通常是一起载入与访问，在数据仓库中并不进行一般意义上的数据更新。

随时间变化是指数据仓库以维的形式对数据进行组织，时间维是数据仓库中很重要的一个维度，并且数据仓库中的数据的时间跨度较大，从几年甚至到几十年，称为历史数据。数据仓库中数据随时间变化的特性表现在以下几个方面：

①数据仓库中数据的时间期限要远远长于操作型数据库中数据的时间期限。操作型数据库中数据的时间期限一般是60~90天，而数据仓库中数据的时间期限通常是5~10年。

②操作型数据库中含有“当前值”的数据，这些数据的准确性在访问时是有效的，同样当前值的数据可被更新。而数据仓库中的数据仅仅是一系列某一时刻生成的复杂快照。

③操作型数据的键码结构可能包含也可能不包含时间元素，如年、月和日，而数据仓库的键码结构总是包含某一时间元素。

数据仓库是企业决策系统的基础，因为在数据仓库中只有单一集成的数据源，并且数据是可以访问的，所以与传统数据库比，在数据仓库中分析决策人员的工作将容易很多。

# 数据仓库体系结构与组成

## 两层的体系结构

有数据仓库的定义可知，它是将企业各个业务系统中与分析有关的数据集成在一起，同时数据仓库面向的应用是分析型数据，因此形成了DB-DW两层数据仓库体系结构，如图2-1所示。



图2-1 DB-DW两层体系结构

其中，业务系统作为主要的分析数据来源，其数据格式主要是表的形式。实际中，由于要保证业务系统的正常运行，一般不直接在业务系统中进行数据的查询和抽取，而是采取备份库或者文件传输的方式进行数据仓库的数据抽取。外部数据源是指数据来源与企业的外部，描述企业运营的外部环境与企业经营分析有关的数据，如各个企业的市场份额等，外部数据作为经营分析的补充，对企业经营决策的正确性起着十分重要的作用。因此应保证外部数据的实时性和准确性。外部数据源具有多样性的特点，如年报等都可以作为外部数据源，同时外部数据源的格式也不统一，如文本、表格和图像等。因此对外部数据源及其数据格式等都要在数据仓库的元数据中进行记录，同时元数据中还应对外部数据的可信程度有一定的评价。

由于数据仓库中的数据源不统一，同时元数据的存储形式也不相同，因此有必要在数据进入数据仓库前先将数据存放在一个统一的暂存区中，引入数据暂存区的主要作用在于：

①统一不同数据源的数据格式，将不同数据源中不同的数据格式转换成同一的数据格式，供数据仓库统一处理。

②进行数据的初步检查，在数据进入数据仓库之前，先对数据进行初步检查，鉴于不影响数据仓库的处理时间，这里的检查将仅涉及比较粗略的数据检查，如记录数量、关键字段是否丢失等，对于错误的数据暂不导入数据仓库，这样对进入数据仓库的数据质量有一定的保证，但是更复杂的数据清洁工作，如字段格式的统一以及数据内容的清洗这种单一记录级的处理工作则应该在数据抽取中完成。

数据暂存区可以多种存储形式实现，如文件目录或者数据库表。

数据仓库中保存了大量的历史数据，同时数据仓库面向的是整个企业的分析应用，但在实际应用中不同部门的用户可能只使用其中的一部分数据，从处理速度和效率的角度出发，可以将这部分数据在逻辑或物理上进行分离，使用户无需到数据仓库的海量数据中查询，只在与本部门有关的数据子集上操作，这样就形成了数据集市（Data Mart）的概念，它是指面向企业中的某个部门（主题）在逻辑上或物理上划分出来的数据仓库的数据子集。将数据仓库按照数据的应用划分为多个数据集市，有利于数据仓库的负载均衡，保证应用的执行效率。同时，由于数据集市具有统一的数据来源——数据仓库，遵循统一的数据模型，保证了各个不同数据集市中的数据统一。

可以看出DB-DW两层的数据仓库体系结构是一种管道过滤器的结构，数据从数据源进入数据仓库到展示给最终用户，都有一定的关联关系，因此要保证数据仓库中数据处理的合理调度，则需要通过数据仓库的元数据完成。

## 三层的体系结构

数据仓库的提出使得操作型处理和分析型处理得以分离，从而形成了DB-DW两层的体系结构，但是在企业的业务处理中存在介于操作型和分析型之间的需求，需要对短期的历史数据进行分析，同时要求较快的响应速度，这种分析无法在操作型数据库中完成，因为其保存的是数据的瞬态信息，但是如果通过数据仓库完成，由于数据仓库保存了大量的历史数据，在响应时间上无法满足需求，因此提出了操作型数据存储（Operational Data Store，ODS）的概念，ODS数据可以概括为面向主题的、集成的、当前或接近当前的和不断变化的数据。其中，面向主题和集成的特点与数据仓库的概念相似；“当前”是指数据在存取时刻是最新的；“接近当前”是指存取的数据是最近一段时间得到的；而“不断变化的”是指ODS数据可以联机改变，包括增加、删除和更新等操作。



图2-2 DB-ODS-DW三层体系结构

面向主题和集成的特点使得ODS数据在部分特征上很接近数据仓库的数据，但是ODS和数据仓库之间也存在重要的差别，主要体现在：

①数据的内容不同。数据仓库中历史数据是指长期保存并可重复查询的数据，既保存细节数据，也保存综合数据，而且一般是静态数据（延迟>24小时）。而ODS一般只保存细节数据，而且ODS数据是可以更新的，即变化的，ODS中保存的历史数据也是近期的（延迟>1秒）。

②数据量不同。ODS保存的数据量要远远小于数据仓库的数据量。

③面向的应用不同。数据仓库用于长期的趋势分析或决策支持，而ODS主要支持企业的全局OLTP和即时决策分析应用。

引入ODS后，原来的DB-DW两层体系结构将被拓展为DB-ODS-DW三层体系结构，如图2-2所示。

在DB-ODS-DW三层的体系结构中，ODS的作用可以概括为：

①为数据仓库提供数据，减少数据仓库数据抽取的复杂性。由ODS的定义可知，它具有面向主题和集成两个特点，因此来自各业务系统的源数据首先进入ODS，在进入ODS时完成数据清洁和集成的工作，这样再向数据仓库提供的数据就是清洁和统一的，减轻了数据仓库中数据抽取的工作。

②即时的OLAP分析。由于在业务系统中需要对近期或当前的数据进行分析，如果该应用放在数据仓库中完成，由于数据仓库相应的处理环节较多，同时数据仓库保存了大量的历史数据，如果要完成这些需求势必造成留给数据仓库的数据处理时间减少，不利于数据仓库的建设，所以将这部分任务分配给ODS，由于ODS保存了近期的数据，可以完成用户的即时分析需求。

③全局的OLTP操作。由于ODS数据的集成性，整合了企业中不同业务系统的数据，同时ODS数据是可更新的，因此ODS可以提供面向企业全局的OLTP操作。

## 数据仓库的组成

数据仓库的大小一般都在TB级以上。传统的关系型数据库是针对OLTP的，并不适用于数据量大且复杂度高的数据仓库。因此，数据仓库系统必须能够满足：

①扩充数据仓库的数据。

②每日对数据仓库系统的管理和维护。

③允许用户增加需求。

要充分满足上述三点并不容易，尤其是数据仓库最终以自动的数据分析处理为目标。扩充数据仓库数据是非常重要的，其重点是抽取、整理并转换数据以及以适当的方式展现给用户分析使用。数据仓库并不是一个只读系统，虽然说事实数据导入数据仓库之后就不会再更新，但是如果用户需求改变，例如希望以不同的方式浏览相同的分析结果，则仍然会修订索引数据。每日对数据仓库系统的管理和维护与传统的OLTP系统完全不同，因为数据仓库的数据量远比OLTP系统大得多，所以需要更积极的管理方式。例如添加或删除数据、将数据仓库数据存入备存介质、由备存介质加载数据等。因此可以这样理解，数据仓库是一个持续更新的系统，以满足用户新的决策需求。允许用户增加需求的能力似乎是设计数据仓库系统时最困难的工作，因为每个用户都有各种不同的需求。数据仓库除了能够允许更新现有的需求之外，还应该可以增加新的分析主题。

数据仓库系统应该具有以下功能：

①抽取数据与加载数据。

②整理并转换数据为数据仓库适用的格式。

③备份与备存数据。

④管理所有的查询，将他们导向适当的数据源。

具体来说，数据仓库系统的管理器可细分为三种，即：

①加载管理器，负责抽取并加载数据，在加载数据之前与过程中执行简单地转换。

②仓库管理器，负责转换并管理数据仓库数据、备份与备存数据。

③查询管理器，负责引导并管理数据仓库的查询。

# ETL过程

业务系统的数据经过抽取、转换并加载到数据仓库所在的中心存储库的过程称为ETL（Extract，Transform and Load）过程，制定这一过程的策略称为ETL策略，而完成ETL过程的工具则是ETL工具。相对于数据仓库的表而言，业务系统数据库中的表称为源表，业务系统数据库称为源数据库，数据仓库中绝大部分数据都来自用于业务系统数据库。在构建数据仓库的过程中，ETL的实施是一项烦琐、冗长而艰巨的任务，因为它关系到数据仓库的数据质量问题。如果导入的数据漏洞百出，那么对于决策者而言无疑会造成困扰。

## 主要流程

随着应用和系统环境的不同，数据的抽取、转换和加载具有不同的特点。

（1）预处理是正式开始作业之前的准备工作，包括清空工作区、检查过渡、准备区。在需要直接访问操作型数据源系统时，应检查远程数据库服务器状态，并核对目标数据加载状态，以核算出加载作业的参数，如加载数据的时间间隔和范围（是24小时的数据，还是前3天的数据）。

（2）启动数据抽取的批处理，执行数据提取、清洁和转换等过程。

（3）因为维表有事实表所参照的主键，所以需要先完成对维表的加载，生成维表主键，并作为以后加载事实表所需的外键。在加载维表时，有时需要处理好缓慢变化的维，并可能涉及版本号的处理问题。

（4）加载事实表。这中间也涉及键查找的问题，即从有关维表中找到相应的主键，并以此作为事实表的外键。

（5）事实表加载完成后，再对数据立方体体系进行刷新，以保障数据立方体与其基础数据同步。

（6）设计具有完善的出错处理机制和作业控制日志系统，以监测和协调整个加载过程。

## 数据抽取

数据抽取是ETL处理的首要任务，解决的主要问题是确定需要抽取的数据，并采用适当的方式抽取。

源数据进入数据仓库是通过数据抽取完成的，从一个或多个源数据库中通过记录选取进行数据复制的过程。抽取过程将记录写入ODS或临时存储区以备进一步处理。

数据抽取的主要功能包括：

①数据提取。主要是确定要导入数据仓库的数据。

②数据清洁。检查数据源中存在矛盾的数据，按照用户确认的清洁规则对数据进行修改。

③数据转换。主要是将数据源的数据转换成数据仓库要求的格式，其中包括数据格式的转换，例如将数据源中的日期字段转换成数据仓库要求的字符形式；数据内容的转换主要是将同一含义的字段用统一的形式表达；数据模式的转换是由于数据仓库系统和业务系统面向的数据操作不同，所以在数据模式上也存在不同，例如电信业务的出账表的主键包括用户标识和费用项，但是在数据仓库中用户主体的用户账务信息采用用户标识作为主键，将不同费用项的费用作为字段，这样就需要在数据抽取时进行不同数据模式的转换。

④生成衍生数据。由于数据仓库保存了大量的历史数据同时需要保证查询的效率，需要对用户经常进行的查询通过预处理操作来提高查询效率，生成衍生数据。衍生数据既包括某些数值数据的预运算，例如平均值和汇总等，也包括某些分类字段的生成，例如对用户费用的分档信息等。

数据抽取的重要组成部分是变化数据捕获（Change Data Capture，CDC）。实现CDC的方法包括时间戳、读取DBMS系统的日志文件、使用源系统中的触发器或者自行开发CDC程序来检查日志文件等。时间戳是最简单的，也是相当普遍的方法，前提是所有的源表都有时间戳。例如超市的业务系统数据库中几乎每个表都有“插入记录日期”和“更新记录日期”两个字段，这是规范建库的一个基本标准。

在多数情况下，数据源与数据仓库并不处于同一个数据服务器中，它们往往是独立的，并处在远程系统中。数据抽取往往是以远程、分布式的方式进行，并涉及各种各样的方法和手段，