****

****

**数据库新技术课程报告**

|  |  |
| --- | --- |
| **论文题目** | 数据仓库发展现状 |
| **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 绪论 3](#_Toc515897707)

[1.1 背景介绍 3](#_Toc515897708)

[1.2 诞生 3](#_Toc515897709)

[1.3 定义 4](#_Toc515897710)

# 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