

基于数据仓库的管理信息系统建设

吴正宏 (中国石油化工总公司信息中心 100029)

摘要:九十年代出现的数据仓库系统有力地推动了决策支持的发展。本文介绍了数据仓库系统的作用和构成,并介绍了中国石油化工总公司机关在数据仓库系统建设方面所作的工作和展望。

关键词:数据仓库 信息系统 数据采集

一、数据仓库系统

数据仓库系统(Data Warehouse System)是为了利用现有的数据进行分析 and 推理,从而为决策提供依据而建立的综合体系。它由数据仓库、数据仓库管理系统和数据仓库工具三部分组成。

1. 为什么要建立数据仓库系统

建立数据仓库系统是为了利用现有的数据进行分析 and 推理,从而为决策提供依据。用计算机进行决策支持的基本概念早在六十年代末就提出,七十年代初美国 Michaels. Scott .Morton 发表论文“管理决策系统”标志着决策支持理论研究开始。随着社会的发展和计算机技术的进步,人们不满足于用计算机只作简单的数据处理和事务处理。进一步用现有的数据进行分析 and 推理从而为决策提供依据的需求导致了决策支持系统(DSS, Decision Support System)的出现。八十年代数据库(DB, Data Base)技术日趋成熟,特别是关系数据库(RDB, Relational Data Base)技术迅速发展,有关学者在研究 DSS 的过程中提出基于关系数据库的三库结构(数据库、模型库、方法库)和四库结构(数据库、模型库、方法库、知识库)。但据此概念开发的 DSS 往往只是简单的查询系统或报表系统,并不能给决策者提供辅助决策的信息。80年代很少见到有实际应用价值的 DSS 的原因如下:

(1)缺乏足够的数据。处于信息系统建设初期的公司或机构,各种数据库和基于数据库的联机事务处理(OLTP, On-Line Transaction Processing)系统刚刚开始建立,数据库和数据库中的数据都较少,没有足够的数据库支持 DSS。科学的决策必须以大量可靠的数据为基础, DSS 往往不仅需要当前的业务处理数据还需要历史的业务处理数据,不仅需要当前的各层次综合数据还需要历史的各层次综合数据。没有足够的数据库难以实现 DSS。

(2)缺乏适当的数据组织形式。即使有足够的数据库,如果这些数据都在一个个独立的业务系统中以表、帐的形式组织在一起。这些数据也只是一种处于原始状态的资源,是一些“信息孤岛”。多年来,国内外许多机构、公司把信息技术用于管理和经营以提高运作效率、增强竞争能力。这些基于关系数据库的 OLTP 系统的应用使得越来越多的关键性数据存入数据库,但是这些数据并没有得到充分的利用。对这些业务系统及其表、帐的数据结构不熟悉的管理人员难以找到并利用这些数据,更难以用其形成完整的业务视图。关系数据库管理系统较适合简单列表记录级的查询修改、删除。传统的基于关系数据库系统的 OLTP 是面向应用、事务驱动的,其数据受到业务系统和表、帐的双重封闭,因此要用支持 OLTP 的关系数据库进行分析是十分困难的。没有集成起来,分散在不同的系统环境中、格式不统一的数据不能很好地支持 DSS。

(3)缺乏必要的技术环境。DSS 需要对大量的数据库进行分析 and 综合,这些数据往往分散在各自的应用系统、各自的计算机系统或各自的网络系统中。要灵活地调用这些数据库并对大量的数据库做高密度的计算需要有高性能的计算机硬件、高性能的网络系统、高性能的数据库管理系统,需要有成熟的客户/服务器技术、并行数据库技术和数据库互操作技术。80年代这些技术尚不成熟。

(4)缺乏成熟的商品化的软件工具。即使具备了必要的数据库,有了适当的数据库组织方式和必要的技术环境还需要有成熟、高效的软件工具才能有效地利用数据库。有商品化的软件工具才能使 DSS 从理论研究,概念、方法的探讨走向实际应用。

90年代以来计算机技术、网络技术和数据库技术的迅速发展为 DSS 提供了必要的技术环境,由于 OLTP 和办公自动化普遍应用积累的大量数据库为 DSS 提供了必

要的数据基础,日趋激烈的市场竞争促进了各级管理和决策人员对DSS的实际需求,自从1991年W. H. Inmon提出数据仓库的概念和1993年E. F. Codd提出OLAP概念以来已有许多商品化的数据仓库管理系统和联机分析处理(OLAP, On-Line Analysis Processing)工具软件面市。以上诸因素的共同作用促成许多公司、机构纷纷为提高自己的竞争能力建立数据仓库系统以进行决策支持。

2. 数据仓库系统的构成

传统的数据库系统由数据库、数据库管理系统和数据库工具三部分组成。与其类似,数据仓库系统也由三部分组成:数据仓库、数据仓库管理系统和数据仓库工具,见图1。在数据仓库系统中存放数据的数据仓库居于核心地位,它是进行数据分析和数据挖掘的基础。数据仓库管理系统完成对数据仓库中数据的各种操作,负责管理整个系统的运转,在数据仓库系统中起承上启下的作用。数据仓库工具是数据仓库系统发挥作用的关键,虽然数据仓库提供了形成完整的业务视图的可能性,但还需要有成熟、高效的软件工具才能有效地利用数据仓库中大量的数据。数据仓库工具有查询工具、分析工具和挖掘工具三类。数据仓库的查询工具主要实现对分析结果的查询,如发展趋势或运行模式,而不是对记录级数据的查询,这类查询在数据仓库中是比较少的。数据仓库的查询工具主要是为各类用户提供可视化工具,以便他们更方便、更清晰地了解综合、分析和挖掘的结果。数据仓库的分析工具主要实现对数据仓库中的数据进行分析和综合,是目前数据仓库系统中应用最多的。数据仓库的挖掘工具主要用于从数据仓库的数据中发现并提取隐藏在其中的信息或知识。

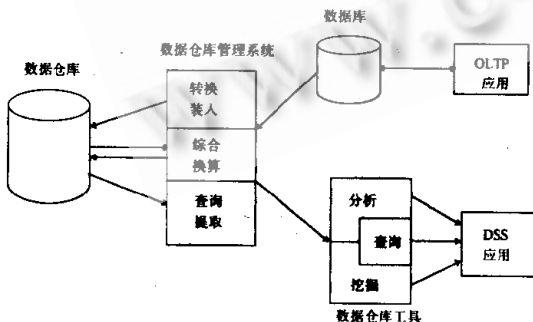


图 1

二、中国石油化工总公司 机关数据仓库系统建设

1. 中国石油化工总公司机关信息系统的特点

中国石油化工总公司机关作为跨地区、跨行业的特大型企业集团的总部对信息系统的要求既不同于一般的大型、特大型生产企业,也不同于只有行业管理功能的部委机关。它对信息的综合程度的要求比企业要高,比部委要详尽。总公司机关的内部信息有三类:机关事务信息、机关业务信息、企业信息,见图2。

总公司机关用来进行决策的内部信息主要来自企业,即其内部信息源主要在企业,所采集、处理数据以综合数据为主。各部门、处室与上下级对口单位之间的纵向联系较多。目前,通过各部门的报表系统收集的数据被封闭在各业务系统和业务系统的各种报表、台帐中,难以用于决策支持。为了进一步提高总公司机关的工作效率和决策支持能力,按照信息中心领导的要求,自94年起为建立总公司机关的DSS,陆续作了以下一些基础准备工作和尝试性的工作:建立灵活高效的数据采集系统;以程序生成器的方式建立通用业务处理系统;建立电子化信息传输渠道;建立财务历史数据库,摸索建设数据仓库的经验;引进必要的软件和工具,包括关系数据库管理系统、数据仓库查询工具、数据仓库分析工具、决策支持软件办公自动化群件。

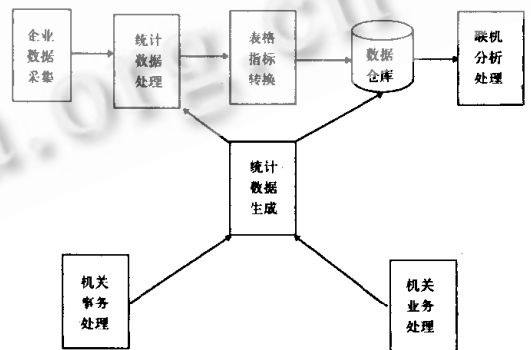


图 2 机关内部信息结构

2. 数据采集系统和财务历史数据库

为了建立灵活高效的数据采集系统,采用软件集成的方法,用Micro Soft公司的Excel VBA(Visual Basic for Application)开发了通用统计报表系统。这个系统也是财务历史数据库的数据采集系统,它已在总公司系统应用。通用统计报表系统可以通过电子邮件向财务历史数据库

传送数据。

中国石化总公司财务部的财务历史数据库是一个部门级的小规模数据仓库,以后可以作为总公司数据仓库群中数据超市的一个成员。小规模数据仓库存储的数据量少,不需要昂贵的软硬件即可有较短的响应时间。通过财务历史数据库的建设可以为总公司的数据库群建设积累经验。

财务历史数据库是基于关系数据库以优化模型结构建立的多维数据库。它是基于数据项而不是基于报表的。在财务历史数据库中,把数据采集系统采集的数据经过数据转换接口转换,以指标的形式存放,以使对数据库内部的数据结构不了解的管理人员可以方便地用其中的数据进行分析。各种统计报表中的数据项经转换,成为财务历史数据库中的指标。指标以数值说明某单位、某时间的某种情况,如资产、负债、利润等。时间、单位、指标名是指标的三个主要属性,也就是财务历史数据库中的三个主维。确定时间、单位和指标名就可以唯一确定一个指标数值。

财务历史数据库的主要功能有:指标分类管理;单位分类管理;用户权限管理;数据采集接口管理;查询模型定义;查询管理;指标数据维护等。

指标名可以按树形结构做两种分类,即两种类型的子维定义:单指标和组合指标。单指标即按一维树型分类;组合指标按行指标和列指标做二维树形分类。每个指标都可以属于多个维。用单指标和组合指标可以灵活方便地由指标生成各种报表。只保存算法,不保存计算值的指标叫做派生指标。指标可以定义换算算法,以生成新的指标数据。

单位可以做多种树型分类,即可以定义多个子维。

指标数值可以在时间、单位、指标名三个维上进行多层次的综合。

通过查询生成的各种查询表有与 Excel 的接口,可以把查询表转换成 Excel 的表单。

以上两方面的工作可由专门介绍通用统计报表系统文章和有关文档得到更详细的了解。

3. 总公司数据库建设展望

总公司的数据库和数据仓库正在抓紧建设。中国石化总公司作为跨地区、跨行业的特大型企业集团,在总公司机关将建立两级数据仓库:总公司领导决策数据仓库

和机关部门数据仓库。各大型和特大型直属企业可以按类似的模式建立自己的两级数据仓库:企业领导决策数据仓库和企业部门数据仓库。企业的部门数据仓库向总公司的部门数据仓库和企业领导决策数据仓库提供经过加工的综合数据。总公司部门数据仓库为总公司领导决策数据仓库提供经过加工的综合数据。这种结构还可以向下延伸。

建立总公司的数据仓库,从数据仓库的连接技术上分两阶段:第一阶段,在总公司机关和直属企业分别建立自己的两级数据仓库。总公司机关的部门数据仓库和企业部门的数据仓库利用现有的数据采集系统或用过渡数据仓库文件实现非直接互访连接。总公司机关和企业的领导决策数据仓库和部门数据仓库间根据环境条件可以采用非直接互访式连接或分布式连接。第二阶段,当总公司机关与各直属企业间的广域网连通后,总公司机关的部门数据仓库和企业的部门数据仓库可逐步过渡到分布式连接。总公司领导在决策过程中可以向下调用企业数据仓库中为总公司机关准备的数据。从数据仓库连接技术上分两阶段建立总公司的数据仓库可以充分利用现有的软件、硬件、网络和信息资源,有利于广域网的网络建设和网络应用建设同步进行。在数据仓库建设中还要注意做好以下工作:

(1)完善数据采集系统。建立表格、指标两种方式统一的,可以实现表格数据和指标数据相互转换的数据采集体系。

(2)完善通用的、有灵活高效的综合数据生成功能的业务处理系统。

(3)在数据采集系统、业务处理系统和数据仓库之间建立畅通的数据接口。

(4)加快总公司广域网建设,建立以 E-MAIL 和 EDI 为主的高效畅通的信息传输渠道,进而实现总公司机关和直属企业网间互访。

(5)制定总公司数据库和数据仓库建设规范,在统一规划指导下加快各级数据库和数据仓库建设。各级数据仓库和数据库间要有统一的接口。

(6)加强有关决策支持知识培训,培养有决策支持能力的高水平复合人才。

(来稿时间:1997年9月)