文章编号:1000－2367(2015)03－0099－07 DOI:10.16366/j.cnki.1000－2367.2015.03.020

环境影响评价数据仓库模型的构建研究

车 蕾[1,2],丁 峰[3]

(1.北京信息科技大学 信息管理学院,北京 100192;2.中国人民大学,信息学院 北京 100872;
3.环境保护部环境工程评估中心 数值模拟重点实验室,北京 100012)

摘 要:随着环境影响评价基础数据库数据中心数据量的逐年指数增长,数据库服务器对客户端的响应时间也随之延长,对数据高效管理的要求也日益明显.结合环评基础数据库项目建设过程中数据分析和挖掘预测的需求,提出基于 ODS 的环评数据仓库的总体架构,构建基于 OLAP 技术的面向多主题的数据仓库模型,研究遗漏值、数据异常等不同问题的数据清理和填充方案,构建基于 DS－ODS－DW 的数据加载模型.研究实现基于基础数据库的统计分析和数据挖掘功能,为数据管理者提供数据预测与决策支持功能.通过模型在数据分析和数据挖掘方面的应用案例,验证了该模型和方法的合理性和有效性.

关键词:环境影响评价;数据仓库;操作数据存储;文件组;分区

中图分类号:X828;TP311.1 **文献标志码**:A

自 2010 年,环境保护部开始环境影响评价基础数据库项目(简称"环评基础数据库")建设,目的在于整合全国各级纵向和横向环评部门之间的业务数据和业务数据服务功能,实现统一数据服务平台的访问、调度功能.其中,数据中心存储的结构化数据已达百万条,预计每年将增加数十万条与项目及污染物排放相关的结构化与非结构化信息[1].如何有效地集成异构数据源,构建环境影响评价数据仓库,使环境影响评价数据仓库为数据分析和挖掘预测提供有效的数据支撑,是一项亟需解决的问题.

数据仓库就是一个面向主题的、集成的、不可更新的、随时间不断变化的数据集合,它用以支持企业或组织的决策分析处理.数据仓库具有 4 个基本特征:数据仓库的数据是面向主题的,数据仓库的数据是集成的,数据仓库是不可更新的,数据仓库是随时间变化的[2].支持数据仓库的商务智能产品众多,根据规模、市场关注度和发展势头来看,SAP,Oracle,IBM,Microsoft,SAS 这 5 大跨国型商务智能供应商所提供的产品,占据了大部分的市场,也是商务智能的领跑者.其中微软已经在数据库产品 SQL Server 当中提供了一体化的商业智能套件,包括相关的数据仓库、数据分析、数据整合、报表、数据挖掘的全系列的设计、开发和管理工具.其中主要的 BI 架构工具为:Sql Server Integration Service,Sql Server Analysis Service,Sql Server Reporting Service.环境影响评价数据仓库是基于 SQL Server 2012 平台构建的.

1 环评数据仓库总体架构

环评基础数据库数据中心收集整理了近 10 年的国家审批的环评报告书、评估意见、审批批文等上万份环评文件,整理了十多个重点行业、数千个建设项目环评指标,同时收集了国家级自然保护区等敏感点信息、地形数据、水文资料、气象数据等环评所必需的大量基础数据.目前,环评数据源主要包括如下几类数据:存储在 Oracle 数据库中的环境影响评价项目库、存储在 SQL Server 中的环境敏感区库、存储在 Excel 中的全国污染源清单库和存储在文本中的建设项目环境影响评价指标库[3].

收稿日期:2015-02-15;**修回日期**:2015-05-10.

基金项目:国家自然科学基金(61272513);北京高等学校青年英才计划(YETP1503);北京市教育委员会科技计划(KM201511232016);环境保护部财政预算(1441100039);国家科技支撑计划(2012BAH08B02).

作者简介(通信作者):车 蕾(1979－),女,河南洛阳人,北京信息科技大学讲师,博士研究生,研究方向为数据仓库与数据挖掘、环境信息化建设等,E-mail:chelei@bistu.edu.cn.

(1)环境影响评价项目库是整个基础数据库的核心数据,存储的是环境影响评价项目报告书的信息,对应前台共享平台中的建设项目查询系统.其中的关键指标,包括项目基本信息、所在位置、项目投资,资源需求、污染物排放量以及项目行业特征指标等.

2) 全国重点污染源清单库是基于环保部政务专网,以 Web 服务接口方式,从环境监测部门获取全国重点污染源(大气、水)在线监测排放数据以及已建成的重点企业的日常污染排放信息数据,等.包括污染源的基本信息、排放大气和水污染物的类型、排放参数、排放浓度和排放量等信息.

3) 建设项目环境影响评价指标库存储建设项目环境影响评价指标信息,包括 14 个行业指标表,其中的关键指标包括:生产产品与工艺、污染治理措施、单位产品的 SO_2 排放总量,NO_x 排放总量,烟尘排放总量等指标信息.

4) 环境敏感区库主要用于存储着国家和各省市自然保护区、生态敏感区、重要水源地以及重要人口聚集地等自然和社会环境信息,其中的关键指标包括各敏感区名称、具体位置、保护对象等数据信息.

图 1 显示了基于环境影响评价基础数据库设计的数据仓库的总体架构.由于环评数据库中的环境影响评价项目库、指标库中的数据主要来源于日常环评评估和在线监测的业务流程中采集和自动生成,为了弥补业务系统和数据仓库之间的数据同步差距,系统采用了 DB-ODS-DW 架构,即在数据源与数据仓库之间加入了操作型数据存储区(ODS,Operational Data Store).在建立了 ODS 数据平台的基础上,可以根据环评业务的分析需求建立各类主题分析模型,数据仓库主要存储和管理从业务系统经一定的业务规则转换并集中存储的业务数据,包含了详细的业务数据与各层次的汇总数据.数据集市是一种小型的数据仓库,主要面向部门级业务,并且只面向某个特定的主题,是为满足特定用户的需求而建立的一种分析环境.这里根据不同的行业划分大的分析主题,构建不同的数据集市[4].

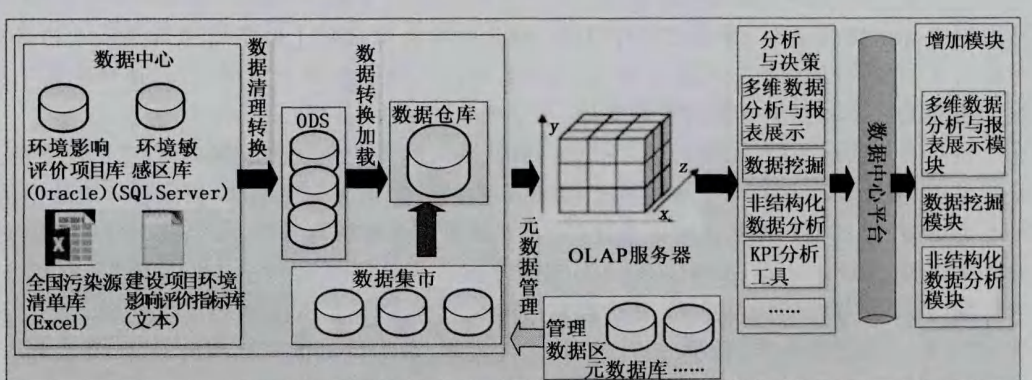

图1 数据仓库的总体架构

2 环评数据仓库建模与实现

数据仓库建模是基于 OLAP 技术的,包括确定主题、确定度量、确定维度、确定事实表和多维数据模型[5].

确定主题即确定数据分析或前端展示的主题.主题要体现出某一方面的各个分析角度(维度)和统计数值型数据(量度)之间的关系,确定主题时要综合考虑.根据环评数据中心的结构和组织特点以及它的业务需求,确定需要重点研究与分析的主题包括:"项目分布"、"项目投资"、"产品与工艺"、"资源需求"、"污染物排放"等.以"污染物排放"主题为例,就要求我们可以通过时间维度、地区维度、工厂维度和项目维度等维度来分析污染物的排放情况[6].

常用的数据仓库模型由星型模型、雪花模型、事实群模型 3 种.这里采用的是雪花模型.图 2、图 3 分别展示了"污染物排放"和"资源需求"两种主题下的多维模型.其中"污染物排放"主题下的多维数据集模型如下:

污染物排放情况(维表；事实表)

维表

={DimAirDrain，DimIndustry，DimArea，DimChimney，DimFuel，DimDevelopmentOrganization，DimPublicInvolvment，DimpublicInvovlvment，DimAttachment，DimPollute<DimPolluteInformation，DimProject，DimAssessmentOrganization}；

说明："<"表示维的层次关系

事实表={FactPollute}.

"资源需求"主题下的多维数据集模型如下：

污染物排放情况(维表；事实表)

维表

={DimAssessmentOrgaization，DimDevelopmentOrganization，DimPublicInvolvment，DimAttachment，DimIndustry，DimChimney，DimArea，DimTime，DimFuel，DimProject}；

DimTime={Time<Day<Month<Quarter<Year}

事实表={FactPollute}.

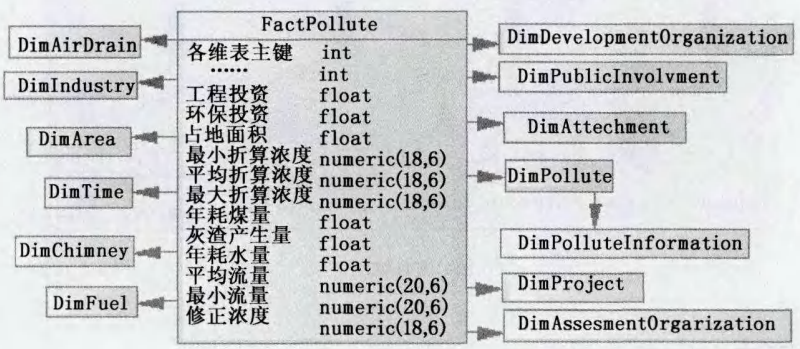

图2 污染物排放主题域多维数据模型

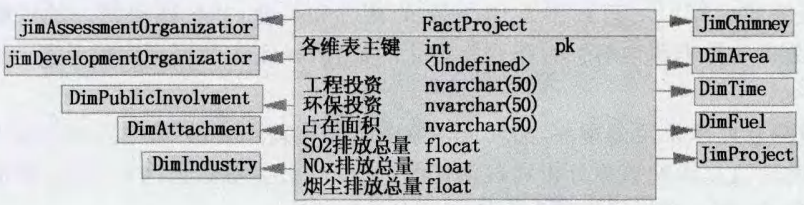

图3 资源主题域多维数据模型

2.1 物理存储设计

2.1.1 文件组

面对大数据，为了提高系统存储和响应效率，这里通过文件组实现分类存储. 主文件组和用户定义文件组两大类. 主文件组包含主数据文件和任何没有明确分配给其他文件组的其他文件. 系统表的所有信息存储在主文件组中. 用户定义文件组与环评数据仓库的分析主题对应，将同一主题涉及的数据表存放在一个文件组中，如图 4 所示.

2.1.2 分区

环评数据中心存储的数据中，对于每个项目，大气污染源按小时排放的折算浓度和排放量的数据就有上万条，每月就有上千万条数据了. 这时就可以按季度划分数据子集，各季度的数据集中存储在一起，特别是需要将这些数据从 OLTP(联机事务处理)加载到 OLAP(联机分析处理)系统之类的操作可能仅需要几秒钟就能完成，而不采用分区可能需要几分钟、甚至几个小时才能完成. 对当前季度的数据主要执行 insert、update 和 delete 操作，而对于以前季度的数据则主要执行 select 查询，这样按季度对表进行分区的优点尤为明显. 具体算法如下：

```
CREATE PARTITION FUNCTION myDateFQ1(datetime)——建立分区函数,按季度进行分区
AS RANGE RIGHT FOR VALUES('20120401','20120701','20121001')
CREATE PARTITION SCHEME myDateFA1——根据分区函数创建分区方案
AS PARTITON myDateFQ1
TO(EnviromentalCenterDW_FG1, EnviromentalCenterDW_FG2, EnviromentalCenterDW_FG3, EnviromentalCenterDW_FG4)
CREATE TABLE PartTable(col1 datetime,col2 char(10))——按分区方案建立表
ON myDateFQ1(col1)
```

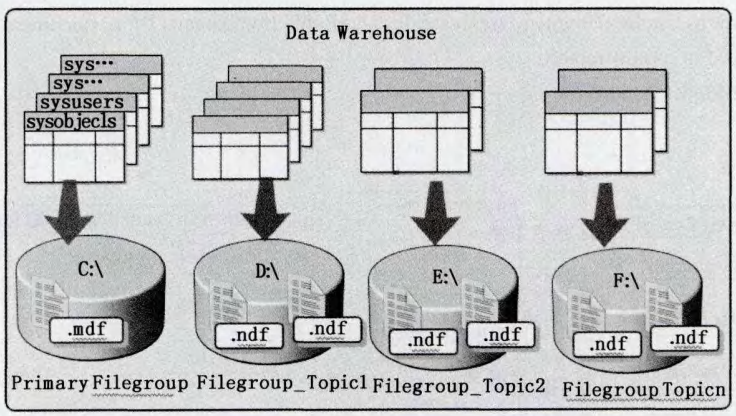

图4 文件组

3 数据仓库的填充

数据仓库的填充包括数据抽取、数据转换和数据加载过程,即 ETL (Extract-Transform-Load 的缩写,即数据抽取、转换、装载的过程)[7]. ETL 作为 BI/DW 的核心和灵魂,能够按照统一的规则集成并提高数据的价值,是负责完成数据从数据源向目标数据仓库转化的过程,是实施数据仓库的重要步骤.

3.1 数据清理

数据质量对于数据仓库来说至关重要. 然而大多数数据是参差不齐的,概念层次不清,数量级不同,甚至数据缺失、异常等问题会直接影响数据分析和数据挖掘,所以有必要对数据进行清理. 数据清理包括遗漏值处理、不一致数据处理等. 表 1 列出了环评数据清理的关键方案.

表 1 环评数据清理关键方案

	问 题	处理方案
遗漏值处理	项目位置、项目投资、建设单位、评价单位等基础信息缺失	参考并调用相关联数据表中的同名字段数据进行匹配和填充; 用条件性拆分控件筛选出项目地址为空的数据行,用派生列控件根据项目名称来填充
	项目占地面积、工程投资等信息缺失	同规模同类型的项目数据类比填充
	污染源排放参数(烟囱高度、排放温度、烟气流速等)缺失	分析同类型、同规模项目已有字段数据均值进行匹配和填充
	报告书受理、评估、存档等时间缺失	参考项目业务受理与审批流程表中的时间信息进行填充
数据异常或不一致	同一项目不同数据表基本信息数据(内容或格式)不一致	应用数据校验规则、格式变换函数、派生列等解决方法
	不同数据表中存储相同项目的基本信息表达字段不一致	对不同数据表的字段进行清理、合并、规范
	污染源排放参数、项目投资、资源消耗量、污染物排放量等字段出现异常大值或负值	应用数据校验与计算规则,或采用统计函数进行分析处理

3.2　数据加载

系统的数据加载包括两部分：(1)通过 SSIS 工具从数据源采集数据，再经过清洗转换，把数据从异构数据源(存储在 Oracle 数据库中的全国污染源清单库；存储在 Excel 中的环境影响评价项目库以及存储在文本中的建设项目环境影响评价指标库)加载到 ODS 系统中.(2)根据数据组织特点逻辑结构和业务需求，再把数据从 ODS 加载到 DW.

3.2.1　数据从异构的 DB 加载到 ODS

图 5 描述了数据从异构的 DB 加载到 ODS 的流程. 其中，输入信息包括：存储在 Excel 中的环境影响评价项目库的 8 张表，分别是建设单位基本信息表、评估单位信息表、项目验收信息表、污染指标表－本工程污染指标表、污染指标表－总体工程污染指标表；存储在文本中的建设项目环境影响评价指标库的 2 张表，分别是火电行业指标表和石化行业指标表；存储在 Oracle 数据库中的全国污染源清单库的 3 张表，分别是污染源基本信息表、大气污染源排放口信息和大气污染源按小时排放的折算浓度和排放量. 输出信息包括：SQL Server 数据库——环评操作数据存储 ODS.

3.2.2　数据从 ODS 加载到 DW

图 6 描述了数据从 ODS 加载到 DW 的流程. 其中，输入信息包括：污染源基本信息表、大气污染源按小时排放的折算浓度和排放量、大气污染源排放口信息、污染指标表－总体工程污染指标表、污染指标表－现有工程污染指标表、污染指标表－本工程污染指标表、评价(估)单位信息表、建设项目基本信息表、建设单位基本信息表和项目验收信息表、石化、火电等行业指标表. 输出信息包括：环境影响评价数据仓库 DW.

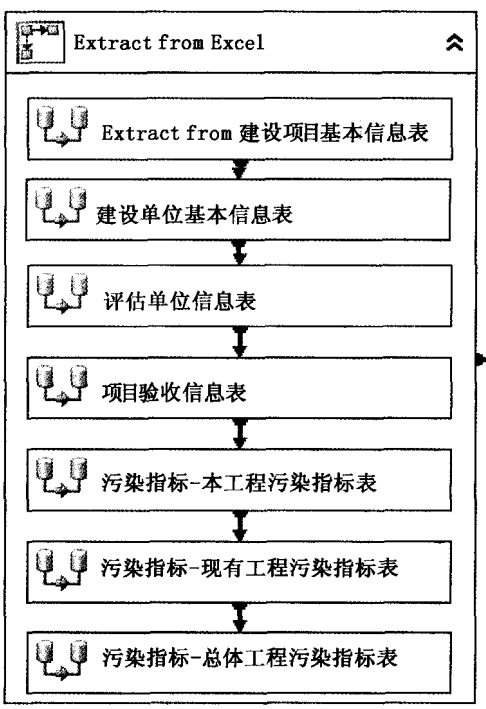

图5 数据从异构的DB加载到ODS

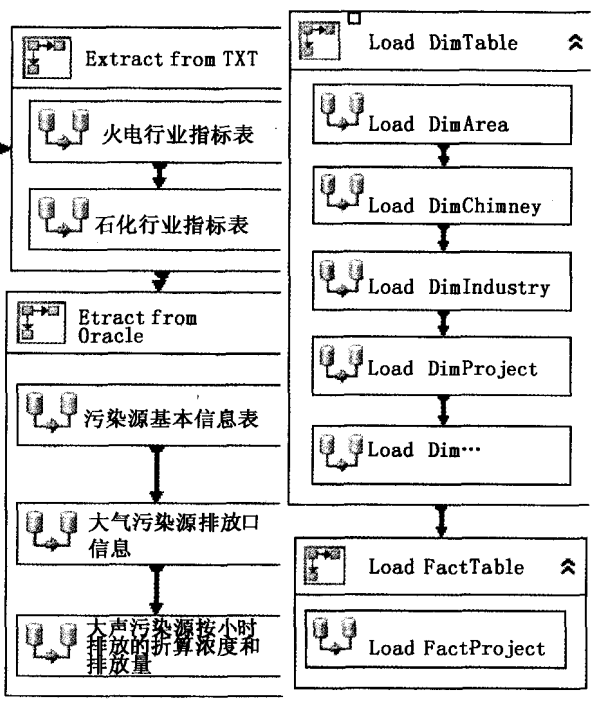

图6　数据从ODS加载到DW

4　模型应用

针对前面所述研究，下面分别给出一个多维数据分析和数据挖掘的应用案例，显示了以省市与项目为主要维度，以评估结论和年为筛选维度，以关键指标为度量值的多维数据分析结果. 多维数据分析模型为：Query=({年='2010'，评估结论='有条件可行'}，{省市，项目}，{工程投资，环保投资，环保投资比重，SO_2 排放总量，NOx 排放总量…}). 显示了基于时序挖掘模型的污染物排放量预测结果，即基于历史污染物排放

量情况(2001—2010 年),假设现有数据是截至 2010 年的污染物排放量情况(实线部分),通过多维数据分析,预测 2010—2015 年之后污染物排放情况(虚线部分),进而实现为数据应用人员及管理部门提供数据深度利用与辅助决策的功能.

年	评估结论	总计													
2010	有条件同行														
省市	项目	工程投资	环保投资	环保投资比重	SO2排放总量	N_{O_x}排放总量	烟尘排放总量	环境污染影响指数	工程投资	环保投资	环保投资比重	SO2排放总量	N_{O_x}排放总量	烟尘排放总量	环境污染影响指数
江苏省	江苏新海发电有限公司2×	1437982	160166	0.10	10058	17810	3410	661,597.62	1437982	160166	0.10	10058	17810	3410	661,597.62
	汇总	1437982	160166	0.10	10058	17810	3410	661,597.62	1437982	160166	0.10	10058	17810	3410	661,597.62
内蒙古自治区	内蒙古大唐多伦电厂一期工	943672	113171.4	0.11	6506	18698	2188	607,173.81	943672	113171.4	0.11	6506	18698	2188	607,173.81
	内蒙古上都电厂二期扩建工	485670	45737	0.09	7102	8501	992	345,063.10	485670	45737	0.09	7102	8501	992	345,063.10
	汇总	1429342	158908.4	0.10	13610	27199	3180	952,236.90	1429342	158908.4	0.10	13610	27199	3180	952,236.90
新疆维吾尔自	新疆华电库尔勒热电厂工程	116360	14417	0.11	7102	8501	1297	349,420.24	116360	14417	0.11	7102	8501	1297	349,420.24
	汇总	116360	14417	0.11	7102	8501	1297	349,420.24	116360	14417	0.11	7102	8501	1297	349,420.24
总计		2983684	333491.4	0.10	30770	53510	7887	1,963,254.76	2983684	333491.4	0.10	30770	53510	7887	1,963,254.76

图7 关键指标的多维数据分析结果

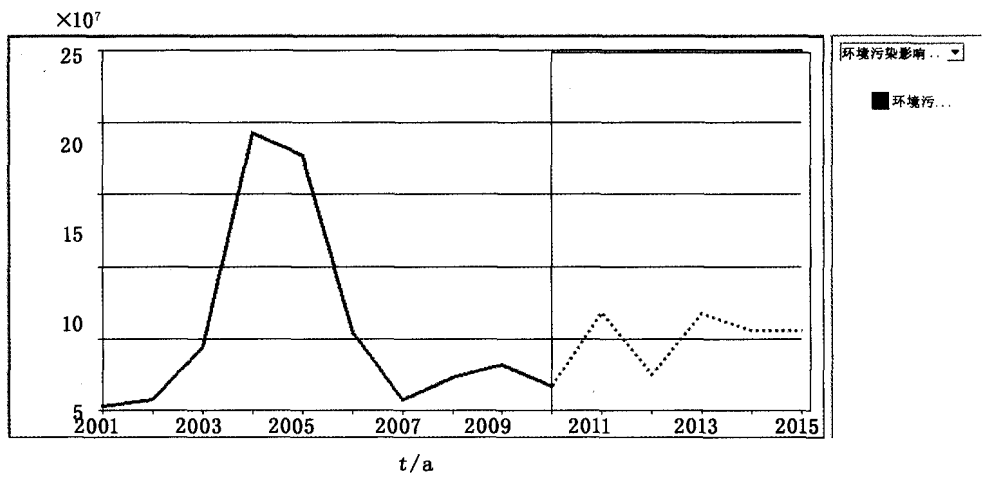

图8 基于时序挖掘模型的污染物排放量预测

5 结束语

环境影响评价数据仓库模型的建设将环境管理过程中各自为政的操作型数据进行面向分析的整合、形成一个集成的一致的数据中心,直接为环评部门进行全局范围的复杂数据分析、战略决策和趋势分析提供数据分析支持.项目实施之前,存在数据未统一管理、客户信息彼此间没有联系和信息缺乏全方位共享等问题;业务系统交互密集,导致数据量庞大且复杂度高.项目实施之后,加强了资源的有效整合,提高了部门的工作效率,有利于不同业务系统数据信息的共享.

结合环评基础数据库项目数据分析和挖掘预测的需求,提出基于 ODS 的环评数据仓库的总体架构,通过构建基于 OLAP 技术的面向多主题的数据仓库模型,横向打通了各业务系统的数据,有效地把操作型数据集成到统一的数据环境中以提供决策数据访问及分析,通过进行数据清理和数据加载,实现适应不同维、不同粒度、不同侧面查询和观察数据的需求.开发了基础数据库数据挖掘应用原型,为数据管理者提供数据挖掘与预测功能.

参 考 文 献

[1] 赵晓宏,丁 峰,李时蓓,等.环评基础数据库建设与展望[J].环境影响评价,2014(4):33-35.

[2] Han Jiawei, Micheline Kamber, Jian Pei. DATA MINING: Concepts and Techniques[M]. 3 版.北京:机械工业出版社,2012.

[3] 伯鑫,刘梦,丁峰,等.环境影响评价报告书数据计算及分析自动化系统设计[J].电力科技与环保,2011,27(6):49-50.

[4] 杨智鹏.数据仓库在电力行业的应用[J].山西财经大学学报,2010,32(1):225-226.

[5] 薛冬娟,高天一,潘 颖,等.船舶企业质量控制模型及数据仓库的构建[J].计算机工程与应用,2012,48(6):229-232.

[6] CHE Lei, DING Feng, WEI C, et al. The Application of Multidimensional Data Analysis in the EIA Database of Electric Industry[J]. Procedia Environmental Sciences, 2011, 10: 456-459.

[7] 温国锋，陈立文. 已完建筑工程数据仓库的建立与应用研究[J]. 计算机工程与应用，2011，47(4)：245-248.

Research of Construction of Environmental Impact Assessment Data Warehouse

CHE Lei[1,2], DING Feng[3]

(1. College of Information management, Beijing Information Science & Technology University, Beijing 100192, China;
2. College of Imformatian, Renmin University of China, Beijing 100872, China; 3. Key Laboratory of Numerical Simulation, Appraisal Centre for Environmental and Engineering, Ministry of Environmental Protection, Beijing 100012, China)

Abstract: With the exponential growth of data of data center of the environmental impact assessment (EIA) basic database, the database server's response time to the client becomes longer, and the demand for efficient data management becomes more obvious. Combined with the demand of data analysis and data mining during the construction process of the EIA basic database, the overall architectures of EIA data warehouse(DW) based on ODS is put forward. The Multi-theme DW based on OLAP and the physical storage model based on file group and partition are built. The different methods of Data cleaning and data filing are studied. The data loading model based on DA-ODS-DW is built. The study realize Statistical Analysis and Data Mining based on the basic database, which provide Data Forecasting and decision support for the management. The legitimacy and effectiveness of the model and methods are verified by the Applications in data analysis and data mining.

Keywords: environmental impact assessment; data warehouse; operational data store; file group; partition