地质灾害数据仓库构建及应用

DIZHI ZAIHAI SHUJU CANGKU GOUJIAN JI YINGYONG

李振华 梅红波 吴湘宁 朱传华
吴润泽 李 芳 杨建英 李程俊 编著

中国地质大学出版社
ZHONGGUO DIZHI DAXUE CHUBANSHE

地质灾害数据仓库构建及应用

DIZHI ZAIHAI SHUJU CANGKU GOUJIAN JI YINGYONG

李振华　梅红波　吴湘宁　朱传华
吴润泽　李　芳　杨建英　李程俊　编著

中国地质大学出版社
ZHONGGUO DIZHI DAXUE CHUBANSHE

内容提要

本书集多年的教学、科研成果，采用“数据驱动”的系统设计方法，以区域地质灾害预测预报主题和监测预报主题为例，从需求规格说明、概念模型设计、逻辑模型设计、物理模型设计四个阶段对地质灾害数据仓库进行了设计，并对空间数据和属性数据分别进行了 ETL 的设计和实现。采用支持向量机等模型对滑坡敏感性和滑坡位移监测进行了数据挖掘应用，初步搭建了基于 Hadoop＋Kylin 的地质灾害大数据多维分析平台。

本书侧重于实践，除适合作为高校地学信息专业本科生和研究生的教材使用外，其研究思路和研究方法也可供地质灾害防治的科研及管理人员参考。

图书在版编目(CIP)数据

地质灾害数据仓库构建及应用/李振华，梅红波等编著. —武汉：中国地质大学出版社，2018.8

ISBN 978-7-5625-4391-6

Ⅰ.①地…

Ⅱ.①李…②梅…

Ⅲ.①地质灾害-灾害防治-数据库系统-研究

Ⅳ.①P694

中国版本图书馆 CIP 数据核字(2018)第 176693 号

地质灾害数据仓库构建及应用　　李振华　梅红波　**等编著**

责任编辑：王凤林　　责任校对：周旭

出版发行：中国地质大学出版社(武汉市洪山区鲁磨路 388 号)　　邮　编：430074

电　话：(027)67883511　　传　真：(027)67883580　　E-mail：cbb@cug.edu.cn

经　销：全国新华书店　　http://cugp.cug.edu.cn

开本：787 毫米×1092 毫米　1/16　　字数：275 千字　印张：10.75

版次：2018 年 8 月第 1 版　　印次：2018 年 8 月第 1 次印刷

印刷：武汉市籍缘印刷厂　　印数：1—500 册

ISBN 978-7-5625-4391-6　　定价：68.00 元

前　言

当今世界已进入大数据时代，国家之间综合国力的竞争在很大程度上是信息的竞争，信息竞争不只表现为拥有了多少信息，更重要的是在于信息的利用度。同样地，从目前的地质工作来说，最大的问题不是数据太少，而是数据太多，以至于没有一个很好的管理和利用方式，更难以获取综合的和深层次的信息。

地学数据仓库为地学海量数据的集成管理提供了一个途径，为更好地开发现有数据打下了基础。从20世纪90年代以来，中国地质大学(武汉)胡光道教授就组建了相关团队，开展了地学数据仓库的理论分析和系统开发的研究。当时的着眼点只是为了满足“十五”期间国土资源大调查的数据管理和后续的其他数据库建设的需要，但随着各项地学数据集成应用的持续需求和应用的不断深入，该团队历经20余年，开发了多个数据仓库系统，特别是三峡地灾数据仓库系统已开发的比较完善，代表了该团队多年的研究水平，现总结成书，既是对过去工作的阶段性总结，也是对未来地学大数据时代的工作展望。

本书在理论上吸取了商业数据仓库的数据集成思想，以空间模型替代现有数据仓库基于时间的组织模型，设计出基于空间的地学数据仓库模型。此模型是商业数据仓库、地理数据仓库和现有地学数据库三者之上的进一步发展，并涵盖了时间、属性、空间等多类型数据；在实践上，研发并建设了三峡库区地质灾害数据仓库，它将“三峡库区地质灾害预警指挥系统”中各系统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，实现了直接为预警指挥系统预测预报及决策分析服务的目的。

应当说明的是，在本书涉及的理论分析和实践研究过程中，得到了湖北省地质局谭照华教授级高级工程师自始至终的指导，他丰富的实践经验和高超的理论水平给予了我们团队关键的技术支持；中国地质环境监测院副院长黄学斌教授级高级工程师也一直把关我们的研发工作，并将我们的研发成果在全国地质环境监测单位进行推广；特别是近5年我们又有幸得到了中国地质环境监测院喻孟良教授级高级工程师和三峡地质灾害研究所所长程温鸣教授级高级工程师的指导，使得我们的系统最终成为一个较为完善和适用的系统。

本书是中国地质大学(武汉)胡光道教授数据仓库团队近20年来在地学数据仓库理论分析和实践探索方面的一次系统总结。该团队前后参与的老师和学生有：胡光道、李振华、王淑华、梅红波、吴湘宁、李程俊、李芳、朱传华、肖敦辉、于炳飞、张寒、张磊、李浩、徐龙飞、秦鑫、胡炫、张俊媛、何彪、李旸、刘志欢、李冀骅、洪丽、郑二佳、马晓刚、任晓杰、李远远、赵琪等。

本书共分10章。第1章为绪论，介绍了数据仓库的起因及其应用到地学领域的发展历程，重点分析和总结了地学数据仓库与商业数据仓库和地理数据仓库的不同之处。第2章为地学数据仓库理论模型，分析了地学数据仓库的数据特点，设计了地学数据仓库的数据组织形式，并提出了基于空间控制点的地学数据仓库理论模型。第3章为地质灾害大数据应用现状，分析了地质灾害数据特点、地质灾害数据模型的应用现状，重点介绍了地质灾害易发性、危险性、易损性、风险性评价研究现状。第4章为地质灾害数据仓库设计，采用“数据

驱动”的系统设计方法，以区域地质灾害预测预报主题和监测预报主题为例，从需求规格说明、概念模型设计、逻辑模型设计、物理模型设计四个阶段对地质灾害数据仓库进行了设计。第5章为数据仓库元数据，介绍了数据仓库的技术元数据和业务元数据的管理。第6章为ETL，介绍了空间数据和属性数据抽取、转换、上载规则，并分别进行了设计和实现。第7章为数据仓库管理，介绍了在OWB中对数据仓库相关文件和资料进行备份，以及对数据仓库立方的维护和增量更新策略。第8章为联机分析处理，介绍了OLAP的相关技术，从应用的角度对OLAP进行了详细设计，并给出了一个地质灾害监测立方的联机分析示例。第9章为数据挖掘，介绍了用于数据仓库数据挖掘的常用模型方法，采用支持向量机等模型对滑坡敏感性和滑坡位移监测进行了数据挖掘应用，并对不同模型的应用效果进行了分析比较。第10章为基于大数据的数据仓库，利用开源大数据平台Hadoop中Hive搭建数据仓库，并利用开源Kylin搭建大数据联机分析处理平台，实现Hadoop下的OLAP联机分析处理，从而满足大数据背景下地质灾害信息化的迫切需求，并给出了一个地质灾害威胁立方的应用分析实例。

本书由李振华、梅红波负责编著，具体分工如下：前言、第1章、第2章由李振华执笔；第3章由李芳、张寒执笔；第4章由朱传华、李程俊执笔；第5章由朱传华、杨建英执笔；第6章、第7章由梅红波执笔；第8章由吴湘宁、梅红波执笔；第9章由朱传华、吴润泽执笔；第10章由吴湘宁、黄成执笔。全书由胡光道审定。

由于本书涉及领域较广，数据仓库的发展跨度较长，在参考文献著录时，仅列出一些学术性的文献，对于一般常识性的文献不再列入，在此向作者表示歉意。此外，因编著者水平有限，疏漏和不足之处在所难免，敬请专家和读者指正。

编著者

2018年5月

目　　录

1 绪 论

§1.1 数据仓库的由来

1.1.1 数据仓库的起因

传统的数据技术处理的是列表式的二维数据资源，即以数据库为中心，主要进行事务处理方面的数据处理工作，但随着数据采集技术的进步，数据的积累速度越来越快，特别是数据积累到一定程度，数据的分析工作会显得日益重要。然而，天生适合于事务处理的二维数据组织方式，满足不了多样化的数据分析处理的要求，特别是对历史数据进行分析的要求。同时应用到了一定阶段，数据处理也从单纯的事务性处理走向事务处理和分析处理并存，与之对应的数据组织形式也从二维走向了多维。

操作型处理也叫事务处理，是指对数据库联机的日常操作，通常是对一个或一组记录的查询和修改，主要是为企业的特定应用服务的，人们关心的是响应时间、数据的安全性和完整性。分析型处理则用于管理人员的决策分析，例如，DSS、EIS 和多维分析等，经常要访问大量的历史数据。两者之间的巨大差异使得操作型处理和分析型处理的分离成为必然。这种分离，划清了数据处理的分析型环境与操作型环境之间的界限，也直接导致了数据仓库的产生。

1.1.2 数据仓库的相关概念

目前，数据仓库的定义一般采用美国著名信息工程学家 William Inmon 博士(1992，1993)在 20 世纪 90 年代初的表述。他认为："一个数据仓库(Data Warehouse，DW)通常是一个面向主题的、集成的、随时间变化的，但信息本身相对稳定的数据集合，它用于对管理决策过程的支持。"

这里的主题是指用户使用数据仓库进行决策时所关心的重点方面，如销售情况、人事情况、整个企业的利润状况等。面向主题指的是数据仓库内的信息是按主题进行组织的，并为主题进行决策的过程提供信息。集成是指数据仓库中的信息不是从各个业务处理系统中简单抽取出来的，是经过系统加工、汇总和整理，保证数据仓库内的信息是关于整个企业的一致的全局信息。随时间变化是指数据仓库内的信息并不只是关于企业当时或某一时点的信息，而是系统记录了企业从过去某一时点(如开始应用数据仓库的时点)到目前的各个阶段的信息，通过这些信息可以对企业的发展历程及未来趋势做出定量分析和预测。信息本身相对稳定，是指一旦某个数据进入数据仓库以后，一般情况下将被长期保留，也就是数据仓

库中一般有大量的插入和查询操作，但修改和删除操作很少。

它与数据库的主要区别如表 1-1 所示。

表 1-1 数据库与数据仓库的主要区别

	数据库	数据仓库
系统目的	面向事务性操作	面向分析性操作
存储单元	表	立方体
数据维度	二维	多维
使用人员	录入员、数据库专家	管理人员、分析专家
数据内容	当前数据	历史数据、派生数据
数据特点	细节的	综合的或提炼的
数据组织	面向应用	面向主题
操作类型	添加、修改、查询、删除	下钻、上钻、旋转、切片

数据仓库还涉及以下概念。

(1)度量(Measures)。度量通常是一个数值指标，用来描述实体的某个数值属性。例如：气象数据中的降雨量、滑坡数据中的位移量、地下水质数据中的矿物质含量等。度量通常有一定的取值范围。

(2)维(Dimensions)。维也称作维度，是人们分析和观察数据的特定角度。例如：地质灾害分析人员常常关心滑坡位移随着时间推移的变化情况，即从时间的角度来观察滑坡位移的情况，因此时间就构成了一个“时间维”。地下水分析人员常常要关心在城区、流域等特定区域地下水中金属离子的浓度情况，这就是从地理分布的角度来观察金属离子的浓度，因此，地理位置可以作为一个“地理位置维”。维的定义通常与具体的分析对象是相关的。

(3)维的层次(Dimension Levels)。维的层次是指描述维的取值时的细化程度。维的取值存在着从粗粒度到细粒度的多个层次。例如：在使用时间维时，时间的取值可能是粗粒度的年份，或是依次细化的季度、月份、天等粒度，因此可以将时间维划分成“年”“季度”“月”“日”几个层次。同样，在进行“地理位置维”角度的分析时，从粗粒度到细粒度可依次划分为“国家”“省、自治区、直辖市”“地/市”“区/县”“乡/镇”“村”等层次。通常，维度的层次划分会依照从宏观到微观、自上而下、从粗到细的自然粒度进行划分。

(4)维的层级关系(Dimensions Hierarchies)。维的层级关系是指层次的某种特定的组合。因为在进行分析时不一定会用到所有的层次，因此在分析时可以选择一些代表某些特定粒度的层次出来，这些被选择出来的层次构成一种层级关系。例如：省级以上的气象分析人员在分析降水量时往往只关心年度、月度降水量，并不关心季度、日降水量。因此，可以设置一个“年月”层级关系，它包含“年份”“月”两个层次。而对于县级的气象分析人员就必须关心所有粒度的降雨数据，此时可以设置“全部时间”层级关系，它包括“年”“季度”“月”“日”所有层次。因此，层级关系可以看作是分析人员对层级结构的定制。每个维都会选择一个主层次关系(Primary Hierarchy)作为其缺省的层次关系(Default Hierarchy)。

(5)维的成员(Dimension Members)。即维的一个具体的取值，这个取值应能够最直接

地描述维成员之间的不同以及维成员所处的层次。由于维是具有多个层次的，因此，维的成员按照粒度的大小也可以是不同层次上的取值组合。例如："地理位置维"粗粒度的取值是某个国家，如"中国""法国""意大利"等，都是粗粒度的取值，再将粒度细化下去，则在"省"一级的取值应该包括上一级"国家"的取值。例如："中国湖北省"就是"省"一级的取值。依此类推，所有下级的取值都必须包含上一级的取值。例如："中国湖北省孝感市云梦县城关镇黄湖村"就是"地理位置维"一个细粒度的取值，包含了所有上层级别的取值。而上层的取值实际上涵盖了所有下面级别的取值，在维度上其实表示的不是一个点，而是一个区域。例如："中国河南省"其实包括了河南省所有下辖的行政区划。

(6)维的属性(Dimension Attributes)。维的属性是指和维的具体成员相关，但是却和OLAP过程没有直接关系的辅助信息，这些信息可以作为分析过程的辅助信息。例如：在地理信息维的"区/县"这一级的成员，有对应的属性"面积""平均海拔""地质构造类型""气候类型"等，这些属性有助于在分析过程中了解某个区/县的细节信息，如在分析滑坡位移时，可结合区/县的地理、地质等维度属性来分析滑坡的影响因素。维的一些属性是所有系统通用的，如长描述(Long Description)和短描述(Short Description)，这些属性系统会缺省生成，但是大多数的维属性还是自定义创建的。图1-1是维度、层级关系、层次的关系示意图，图中时间维有两种层级关系。

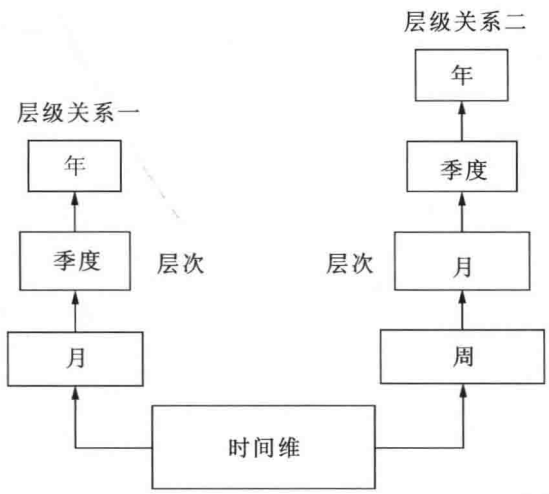

图1-1 维、层级关系、层次的关系示意图

(7)数据立方体(Cubes)。也称为多维数据立方体，其实就是将各个维度作为坐标轴构成一个坐标系，而将度量值放置在坐标系上各个不同的定位点上所构成的一个多维空间信息体。一个多维立方体可以包含多个维，也可以包含多个度量。因此，一个多维数据立方体可以用多维数组(维度1，维度2，维度3，…，维度m，度量1，度量2，…，度量n)的模型来表示。例如：(时间维，测量方法维，地理位置维，降雨量)就是一个具有三个维度和一个度量的数据立方体模型。

(8)数据单元(Cells)和事实(Facts)。多维立方体的维度可以看作是坐标系，当每个维度上给出一个最细粒度的取值时，也就确定了多维空间上的一个坐标点。这个坐标点就称

为一个“数据单元”，里面所存放的度量值被称为“事实”。如果将数据立方体看作是一个多维数组，那么数据单元可以看作是多维数据的具体取值。例如：(“20030809”“雨量器”“中国湖北省孝感市云梦县城关镇黄湖村”“12.5mm”)就是降雨量立方体里面的一个数据单元，这里用时间维、测量方法维、地理位置维三个维度上各自取的维度值“20030809”“雨量器”“中国湖北省孝感市云梦县城关镇黄湖村”，这三个值如同坐标一般唯一确定了一个数据单元，而数据单元里面所存放的度量“降雨量”的值“12.5mm”就是一个事实。图 1-2 是降雨量数据立方、数据单元及相关维度的示意图。

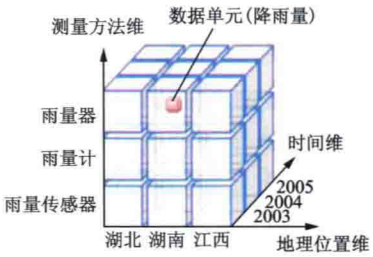

图 1-2 降雨量数据立方、数据单元及相关维度示意图

§1.2 数据仓库的国内外研究进展

由于数据仓库所固有的面向分析的特点和海量数据的集成能力，它一经出现即引起了地学数据研究者的兴趣，然而，由于地学数据与商业数据的时空差异，这一技术应用到地学领域并不顺利，时至今日，尚不能像商业数据仓库那样成熟地应用。其在地学领域的发展大致可以分为如下三个阶段。

第一阶段(1993—1996 年)：商业数据仓库(CDW：Commercial Data Warehouse)的产生。

1993 年 William Inmon(1992,1993)提出了数据仓库的概念，这一概念与传统的数据库有很大的不同，它改变了库中数据的组织形式，通过按主题组织数据、按时间划分数据的粒度，来达到海量集成和分析商业数据的目的。

第二阶段(1996—1999 年)：地学数据仓库(GDW：Geological Data Warehouse)萌芽。

1996 年，美国联邦地理空间数据委员会(简称 FGDC)颁布了国家空间数据基础设施计划(简称 NSDI)(FGDC,1996)，有关地理空间数据仓库(GSDW：Geo-Spatial Data Warehouse started)的研究开始展开，如 1997 年美国国家技术情报局(NTIS)展示了一个地理数据仓库的原型系统；1999 年 Sylvia(1997)探讨了空间数据仓库建设的标准问题及相应开发技术；同年 Shekhar 等(1999)在总结近 20 年以来空间数据库的发展之后，将数据仓库作为空间数据库以后发展的一个重要方向。我国也开始了这方面的工作，1999 年李德仁说明了数据仓库技术在地球空间数据框架中所起的重要作用；杨群等(1999)、杜明义等(1999)对地理信息数据仓库所涉及的技术和模型进行了描述。

在此阶段，国际上实用的地学数据仓库建设也相继开始，比较有名的数据仓库有 2 个：①加拿大水文地理局的海洋深度数据仓库(Forbes 等,1999)，用 Oracle 关系数据库设计和开发而成，存储和管理上千兆的深度数据，同时采用数字地形模型(DTM)对水平和垂直方向的两种数据格式进行了转换；②美国国家水质评价(NAWQA)(Cohen,1999；Bell,2000)数据仓库，它实际上是一个联机数据库，包括 650 万条记录，涉及 46 个州的 2800 条河和 5000 个钻井，但是这些数据分别保存在 EXCEL 或 ASCII 文件中，通过 USGS 主页访问数据仓库。

通过上述文献可知，这一时期地质学家或地理学家已意识到数据仓库的重要性，开始了理论方面的探索，但实用的数据仓库还是没有脱离数据库的框架，只是海量数据在数据库中的简单堆积，不能算是严格意义上的数据仓库。

第三阶段(2000 年至今)：地学数据仓库发展(以地理数据仓库的发展为主)。

自 2000 年开始，地理数据仓库体系结构的研究兴起，如 Keighan Edric 等(1999)探讨了空间数据仓库的体系结构，同时也对数据仓库的广延性(多数据类型)、可扩展性(TB 级存储量)和多分辨率进行了展望；赵霈生等(2000)、周炎坤等(2000)探讨了空间数据仓库的体系结构和关键技术。同时，数据仓库的重要性也更为人们所认识，如美国地调所 Charles (2000)认为：数据仓库是实现地理数据共享的关键技术，他展示了数据仓库的结构和有关的数据标准；2001 年在加拿大召开了第 2 届国际数字地球研讨会，其中“数据仓库与数据挖掘”成为会议的一个主题(Nickerson，2001；Donnelly，2001；Curkendall，2001)，会上 Zhi Li 等(2001)认为集成和共享的数据仓库应是未来数字地球的一个重要组成部分；另外，地理领域之外的讨论也开始进行，如张夏林等(2001)展望了数据仓库技术在国土资源信息系统中的应用。与此同时，实用性研究也开始展开，如 Barclay 等(2000)建立了一个单主题的数据仓库——大型地图库，存储来源于 USGS 和 SPIN-2 的影像；Jermaine(2001)对空间数据仓库的索引问题进行了研究，设计出 T2SM 的高性能空间索引结构。

2002 年地理数据仓库取得了一个重要进展，即在原有数据仓库的基础上强调了数据的空间特性，如 Papadias 等(2002)提出了时空数据仓库的思想，认为可将空间维与时间维合并成一个混合维；尹章才等(2002)对时空数据仓库进行了探讨，认为时空数据仓库是时态地理信息系统和数据仓库相结合的产物。与 GIS 结合的研究也引起了关注，如陈琳等(2002)结合 GIS 技术，研究了地理信息数据仓库的体系结构和关键技术；邹逸江(2002)重点描述了空间数据仓库与测绘数据库和应用系统的区别与联系，以及相应的空间数据仓库的体系结构。

2003 年仍然以地理数据仓库的发展为主，令人欣喜的是地理数据仓库已有背离商业数据仓库设计初衷的迹象，越来越面向存储和检索，而不是像商业数据仓库那样纯粹是为了面向分析。Savary 等(2003)提出一个基于 GML(Geography Markup Language)和 XML (Extensible Markup Language)的异质 GIS 数据仓库的设计，其中 GML 表示空间数据，XML 表示属性数据，强调的是存储；Li 等(2003)针对商业数据仓库中空间数据的联机分析(Online Analytical Processing，OLAP)的检索操作进行了分析，Choi 等(2003)对时空数据仓库也作了类似的分析，强调的都是检索。同时，国内的 Qian 等(2003)和 Zhang 等(2003)也对一种空间数据仓库检索的方法进行了分析，Qian 设计的是一种对大型数据库和数据仓库均普适的算法，Zhang 则定位在商业数据仓库中空间数据的检索；Carr 等(2003)对数据仓库技术在 EOSDIS 数据池中的存储和检索功能进行了研究，对存储和检索都进行了强调。

在地球科学领域，于焕菊等(2006)分析了我国华北地区地震空间数据仓库的结构；王永志等(2008)分析了地学空间数据仓库的构建技术；鲍玉斌等(2009)描述了海洋环境数据仓库多维建模技术；黄解军等(2009)说明了面向数字矿山的数据仓库构建及其应用技术；廖晓玉等(2009)设计了松花江流域水资源空间数据仓库；陈红顺等(2009)设计了广东省韶关市环境污染数据仓库；魏红雨(2014)提出了基于 4G 地学空间数据集成模型并构建了相应的数据仓库：针对地质学(Geology)、地理学(Geography)、地球化学(Geochemistry)、地球物理

学(Geophysics)数据具有多学科综合特点和多重异构问题,进行数据分析、处理、融合等集成操作,以构建地学数据模型,并采用地学空间数据仓库和数据质量评价控制等关键技术,为数据集成和信息共享提供了尝试性的方法,并为后续的数据挖掘和数据融合处理奠定了基础。

胡光道团队自 20 世纪 90 年代开始就一直从事地学数据仓库的理论探索与应用开发工作。胡光道等(1998)将数据仓库技术应用到矿产资源评价领域,用以提高金属矿产资源勘查、分析评价的能力;李振华(1999,2002)和王淑华(2004)构建了矿产资源管理数据仓库,将不同比例尺粒度级矿产资源空间数据进行分级存储和数据综合,通过空间控制点将不同类型的数据叠加,实现多源地学数据的整合。胡光道等(2011)通过集成融合三峡库区不同时空范围的各类数据资料,按五个主题对数据进行分类,并建设了三峡库区地质灾害数据仓库;蔡胤等(2010)对三峡库区地质灾害数据仓库的 ETL 过程开展了研究;朱传华(2010)采用"数据驱动"的系统设计方法,并以区域地质灾害预测预报主题和滑坡监测预报主题为例,构建了三峡库区地质灾害数据仓库,采用支持向量机等方法,进行了数据仓库挖掘的实例应用。梅红波(2010)采用维度建模的方法,建立了三峡库区单体滑坡灾害数据仓库的总线结构,实现了数据仓库的并行、增量构造。张鸣之等(2014)以构建国家级地质环境数据中心,实现地质环境信息大综合、大集成为目标,将各类操作型数据面向业务分析整合,实现了不同粒度、不同维、不同侧面查询及观察数据的功能,为业务分析和决策支持提供了数据保障。吴湘宁(2014)构建了一个地质环境数据仓库,并实现联机分析处理和数据挖掘功能的完整体系,由此形成了一套地质环境数据集成、分析、挖掘、展示的完整框架。

§1.3 从数据仓库到地学数据仓库

通过对上述文献分析,可以认为:现有地学数据仓库在商业数据仓库基础上有了一定的发展,在地学数据仓库中考虑到了空间维的问题,也考虑了与 GIS 的结合问题,甚至在数据存储和数据检索方面做了很多工作,但从框架上还是没有摆脱现有商业数据仓库基于时间的数据组织形式,以下几个问题仍难以解决。

(1)不同类别数据集成。商业数据可以按时间集成,而地学数据按时间集成就比较困难,特别是不同类别不同格式的数据集成更为困难。

(2)分级存储。海量数据往往需要根据数据粒度的大小来进行分级和分布式存储,如果地学数据以时间来组织,就难以确定数据的粒度,也就没有分级存储的依据。

(3)数据的综合。对地学数据的分析,主要侧重于空间方面的综合性分析,而以时间为主的数据组织形式显然与此不太协调。

因此,地学数据仓库的设计必须考虑地学数据以空间为主的特点,区别于原有商业数据仓库基于时间的特点,重新设计基于空间的地学数据仓库模型。商业数据仓库、地理空间数据仓库和地学数据仓库三者的区别如表 1-2 所示。

表 1-2 商业数据仓库、地理空间数据仓库和地学数据仓库的区别

	商业数据仓库	地理空间数据仓库	地学数据仓库
数据组织	时间	以时间为主,以空间为辅	以空间为主,以时间为辅
数据粒度	时间粒度(如年月日)	时间粒度	空间粒度(比例尺)与时间粒度共存
数据类型	属性数据	属性数据和 GIS 数据	所有类型的数据
数据立方	需要	需要	仅用于属性数据
建库目的	数据分析	数据分析	数据集成与数据分析并重

基于以上认识,我们近年来进行了地学数据仓库的初步研究(Li 等,2003;胡光道等,1998,1999,2002;李振华等,1999,2002;王淑华,2002),认为在地学数据仓库的设计中,应当采取基于空间的数据组织形式,用比例尺作为粒度级别代替商业数据库中的时间粒度级别,以实现海量数据的分级存储和不同粒度的数据综合;以空间控制点代替时间控制点(日、月、年),通过空间控制点将不同类别的数据叠加,以实现多源地学数据的集成;以此模型为基础,可实现任意类型、任意比例尺、任意区块的数据输出,在考虑时间维的情况下,此模型还能实现任意时间段的数据输出。

有意思的是,我们在 2004 年提出的上述基于空间控制点的地学数据集成模型,2005 年 Google Maps 在集成遥感数据和地理数据时,也采用了同样的模型。

2 地学数据仓库理论模型

§2.1 地学数据仓库的数据特点

基于地质数据的特殊性，相对一般意义上的数据仓库，地学数据仓库的数据特点如下。

(1)数据的性质不同。一般数据仓库中的数据表现为时间属性，而地质数据表现为空间属性(有的变动较快的数据还具有时空四维的特征)。地质数据多是描述性数据，一般不随时间而变化(至少在研究期内，可以认为是静止不变的)。因此，在建立地学数据仓库时，要着重考虑其空间方面的特点。

(2)数据的更新属性不同。一般数据仓库中的数据是不可更新的，而地质数据是可更新的。在地学数据仓库中，数据是必须更新的，在某一地区如果有新的数据出现，必须立即覆盖旧的数据，以保证数据的准确性。因此，地质数据的生命期的概念与一般数据仓库也不相同，如果没有新数据，老数据继续存在。当然，地质数据的刷新频率是相当慢的(几年甚至几十年一次)，这也符合分析型数据的特点。

(3)数据类型复杂程度不同。一般数据的数据类型比较简单，地质数据的数据类型比较复杂。常用的整型、实型、字符型等简单数据类型满足不了地质数据描述的需要，具有空间特点的地质数据应用在关系数据模型中还需要作技术上的处理(当然，地质数据应用在数据库系统中也有同样的问题，但在数据仓库系统中，必须转换成结构性的数据以便于海量组织和分析，使得这个问题变得更为突出)。

以上是几个主要的不同点，当然，在其他方面如集成性等与一般数据仓库是相同的。

王珊(1998)将数据仓库定义为：是一个用以更好地支持企业或组织的决策分析处理的、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。鉴于地质数据的特点和目前数据仓库的实现平台(还是传统的关系数据库)，地学数据仓库可以定义为：是一个用以更好地支持地学决策分析处理的、面向主题的、集成的、不常更新的、能存储空间数据的、随时间和空间不断变化的地学数据集合。

§2.2 地学数据仓库中的数据组织

由于地质数据有着自已的特点，这导致了地学数据仓库在数据组织上与一般数据仓库有所不同。一般数据仓库中按时间进行组织数据，而地学数据仓库中则按空间进行组织。本书选取比例尺为空间的度量参数(图 2-1)。

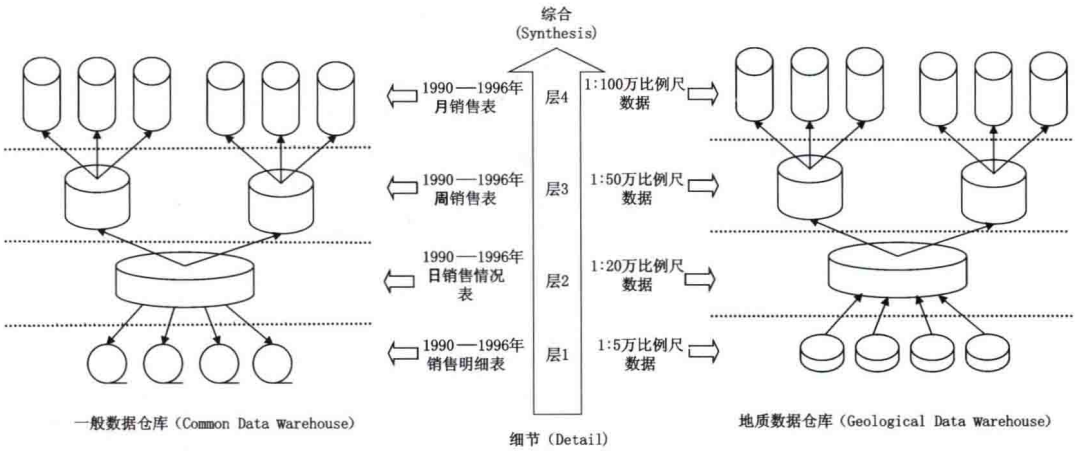

图 2-1 地学数据仓库与一般数据仓库在组织结构上的比较

图 2-1 所示左边为一般数据仓库组织结构(在不失原意的情况下略有改动),层 1、层 2、层 3、层 4 分别指早期细节级、当前细节级、轻度综合级、高度综合级(廖晓玉等,2009);右边地质数据仓库是基于一般数据仓库,地质数据仓库在结构上与一般数据仓库是基本一致的,但也有几点如下的区别。

(1)选取的度量参数不同。前者为时间,后者为空间。

(2)数据流向不同。前者只有层 2(即当前细节级)能接受外界的数据,并且其他层的数据都来源于这一层。而后者的各层都能直接接受数据。

(3)数据的可更新属性不同。图 2-1 所示地学数据仓库中的层 1 采用存储设备符号表明了数据的可更新性(目前数据仓库在实现环境上仍是传统的数据库,这使得在数据的可更新性上并不需要作特别的设计)。要注意的是,底层的数据如有变动,将会级联地改变上层的相关数据。

§2.3 基于空间控制点的地学数据仓库模型

2.3.1 设计思想

由于地学数据都具有空间特点,因此对于不同领域不同格式的数据,它们都能通过坐标控制点在空间上建立对应关系。另外,控制点的数量直接影响到数据叠合的精度和以后数据切割线的锯齿的大小,因此也需要在控制点数目和计算效率间进行平衡。

有一个例子或许可以说明基于空间控制点的地学数据集成的重要性。科索沃战争期间,以美国为首的北约“误炸”了我国驻南联盟大使馆,美方给出了所谓“旧地图”的解释。在此假设,如果美方采用了上述基于空间控制点的数据集成技术,使用的是集成后的数据,那么不管地图多么旧,但起码遥感数据能反映大使馆的存在,集成后的数据当然就能反映大使馆的位置,这样,它就连“旧地图”的托辞都找不到了。

2.3.2 基于空间控制点的地学数据仓库“金字塔”模型

数据仓库逻辑上呈金字塔结构，自底向上按比例尺从大到小分层，在每一个层内又分为多个图层，每个图层代表某一地学类型的数据，多个图层依照控制点的空间对应关系进行叠合(图 2-2)。

虽然不同类型的数据有不同的数据格式，但由于地学数据的空间特点，它们之间存在着空间坐标的对应关系，因此都可通过共同的“控制点”来进行数据层的叠加。正是由于“控制点”的存在，才得以在横向上不同格式的同一比例尺数据之间建立联系，同时，在纵向上同一格式的不同比例尺数据之间也建立了联系。

这种模型的优点是：

(1)与现有的数据集成模型相比，本模型可集成不同类型、不同比例尺、不同格式的数据。

(2)可依比例尺进行分级存储。中小比例尺数据集中在省级数据中心，大比例尺数据分散于基层单位，以减轻数据中心的数据存储量和访问量，并且这种方式也符合现有的地调系统行政管理模式，上层的中小比例尺数据供规划、预测等较宏观的工作，下层的大比例尺数据则适合基层单位的日常细节型工作。

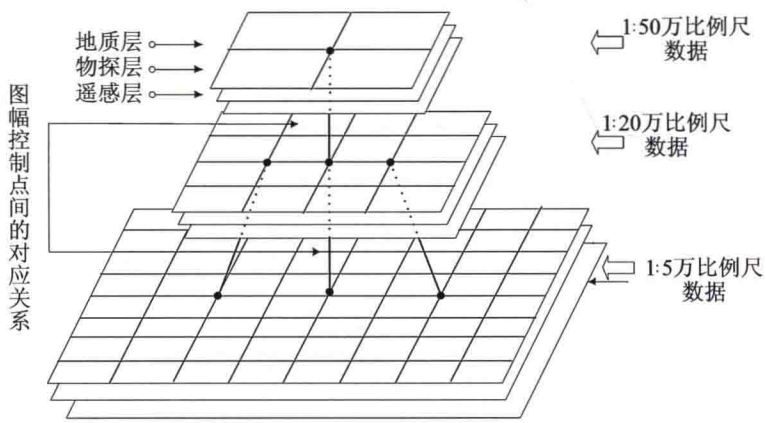

图 2-2 基于空间控制点的地学数据仓库“金字塔”模型

(3)该模型精度可变，普适性强。数据叠合的精度直接取决于控制点的多少，控制点越密，叠合的精度就越高，反之则越低。即使是在最坏的情况下，整个图幅内没有一个控制点，但由于每个图幅都有 4 个角点，不同类别的数据依然可以通过这 4 个角点实现数据叠合，数据调出(如调到工作区)时就一次按一个图幅调出，这时数据输出方式与现有地学数据库的输出方式相同。由此可看出本数据仓库的一个特点：精度可变，普适性强，最差情况下也能等同于现有数据库的输出水平。

2.3.3 不同比例尺图层间的数据处理

目前，地学数据仓库涉及的数据类型可分为两种：属性数据、空间数据。对这两种数据