

基于数据仓库的高校数据统计服务平台研究

龙新征, 李丽, 彭一明, 王倩宜, 欧阳荣彬, 刘云峰

(北京大学 计算中心, 北京 100871)

摘要: 为了适应新形势下高校对大数据统计、分析的需求, 设计实现了一套基于数据仓库的高校数据统计服务平台的整体方案。对高校数据仓库的整体架构、数据仓库的建模进行了研究, 并设计了一种基于数据仓库星型结构的统计数据模型; 进一步探讨了统计服务平台在实现过程中 ETL(extraction transformation loading)以及界面设计等方面的关键技术; 为高校下一步搭建管理与决策支持平台提供了重要的实践基础和理论支撑。

关键词: 商业智能; 数据统计; 数据仓库; 星型模型; 统计数据模型

中图分类号: TP315

文献标识码: A

文章编号: 1000-436X(2013)Z2-0163-07

Research of university data statistical platform based on data warehouse

LONG Xin-zheng, LI Li, PENG Yi-ming, WANG Qian-yi, OUYANG Rong-bin, LIU Yun-feng

(Computer Center, Peking University, Beijing 100871, China)

Abstract: In order to adapt to the demand of statisting and analyzing huge-data in universities under current situation, an overall programme was designed and implemented to construct data statistical service platform based on data warehouse. The overall framework of university data warehouse and the modeling method of data warehouse were researched, and a statistical data model based on data star-schema model was designed. Furthermore, the key technology in the process of implementing statistical service platform like ETL(Extraction-Transformation-Loading) and UI interface design were explored. It provides practical basis and theoretical support for the next step to build the management and decision support platform of universities.

Key words: business intelligence; data statistics; data warehouse; star-schema model; statistics data model

1 引言

在多年电子校务建设的进程中, 各高校建成的信息管理系统越来越多, 对业务运转起到了很好的支撑作用, 在多年运行中产生和积累了海量业务数据。现在, 用户意识到海量数据背后隐藏着许多重要信息, 是学校正常运转的核心资源, 以灵活便捷的方式对数据进行统计、分析, 进而为高校管理与决策提供支持的需求日益强烈。

在商业领域, 这是一个热门研究课题——商业智能(BI, business intelligence), BI的定义最早由 Gartner Group 的 Howard Dresner 于 1996 年提出, 他将商业智能定义为一类由数据仓库(或数据集市)、查询报表、数据分析、数据挖掘、数据备

份和恢复等部分组成, 以帮助企业决策为目的的技术及其应用^[1]。BI 的体系在国内外的的发展各不相同: 微软的 BI 核心是根据数据库中的数据实现各种应用, 体现的是 BI 的一般思路; IBM 在此基础上增加了管控和安全; 美国数据仓库研究院则侧重 BI 的结果对策略和行动的关联和影响; 国内有些研究者认为 BI 体系可以分为 4 层: 最底层是信息系统层面, 其上数据分析层面, 再上是知识发现层面, 顶层是战略层面^[2]; 然而, 在不同体系的指导下建立 BI 系统的基本步骤却没有太多争议, 包括: 1) 确认和解读数据源; 2) 进行数据采集和存储管理; 3) 构建模型并在此基础上分析数据^[3]。

商业智能的基础是数据仓库, 数据仓库是一个

面向主题的、集成的、相对稳定的、反映历史变化的数据集^[4]。它有效地为 BI 系统提供了全局一致的数据环境，也为历史数据综合数据的处理提出了一种行之有效的解决方法。

目前国外大部分企业的 BI 应用已经进入了数据分析阶段，有些已经积累了高端的数据挖掘经验；国内 BI 的应用则还停留在数据整合的初级阶段，应用的主要领域集中在电信、保险、销售等行业，高校的情况更不容乐观：1) 数据分散在不同的源系统中，数据的规范性和共享性还存在很大问题，虽然有一些高校为了不同系统之间的数据共享建立了数据中心^[5]，但只是对数据进行了简单的处理，并没有更进一步考虑不同系统数据之间的关联性，没有像数据仓库那样面向不同的主题来管理数据；也没有考虑对历史数据如何进行管理；2) 数据统计和分析基本上是基于单个系统中的操作型数据进行的，既不能反映出不同系统之间的数据关联，又缺乏对数据的全局把握；还会因为操作型数据的动态性和分散性影响统计结果的准确性，也无法对历史数据进行统计和分析。

因此，要建立高校的 BI 应用，满足为高校管理与决策提供支持的需求，首先必须打破不同应用系统之间的“藩篱”，建立全局一致的数据仓库，将操作型数据转换为静态的、稳定的、规范化的、能够反映历史的分析型数据，然后在此基础上搭建统一的数据统计服务平台。

本文提出了基于数据仓库构建高校数据统计服务平台的完整解决方案。不仅给出了全面的系统架构方案，还探讨了数据模型及关键技术，通过实际应用验证了系统具备良好的可行性、可用性和可靠性，为高校下一步搭建管理与决策支持平台提供了重要的实践基础和理论支撑。

2 需求分析

通过对北京大学等高校的实际情况进行调研发现，数据统计需求按用途不同分为 2 类：一是以年报表或者季度报表的形式上报上级部门的统计数据，有着固定的报表格式、复杂的报表内容、专门的统计口径，有的报表甚至还有严格的填报流程（如：中组部的年度党员统计，涉及到二级单位填报和全校核查汇总）；二是部门日常所需要统计数据，往往和某种类型具体业务相关（如：月工资报表、人员分布情况等），和第一类需求相比，统计数据的格式和内容比较简单，没有复杂的填报流程，但时间粒度要求更细，要以月报表、周报表乃至日报表的形式提供统计结果，并且要求提供数据钻取的功能。

这些需要统计的数据涉及高校人事、学生、科研、教学等各个领域，每个领域下面又细分为了很多不同的细类，具体如图 1 所示。

通过购买昂贵的商用报表软件的方式可以实现一些简单的统计数据的生成。但也往往存在以下缺点：1) 报表的生成过于复杂，需要专业人员对报表软件有相当的熟悉程度；而在现实环境中，普通用户一般不具备计算机方面的相关知识，因此让普通用户来生成报表几乎没有可能；2) 对报表软件依赖性强，一旦统计表格的格式发生些许变化或者临时需要新的统计数据，都需要由程序开发人员重新进行开发，而不是通过简单的配置就能达到效果，因此几乎没有扩展性可言。

在高校数据统计服务平台的建设过程中既需要考虑功能性需求，实现各项业务功能；又要充分考虑最终用户的需求，使它们能通过一种最简单快捷的方式获取到最想要的的数据；更需要关注非功能性需求，保证系统的高可用性、可靠性及可扩展性。

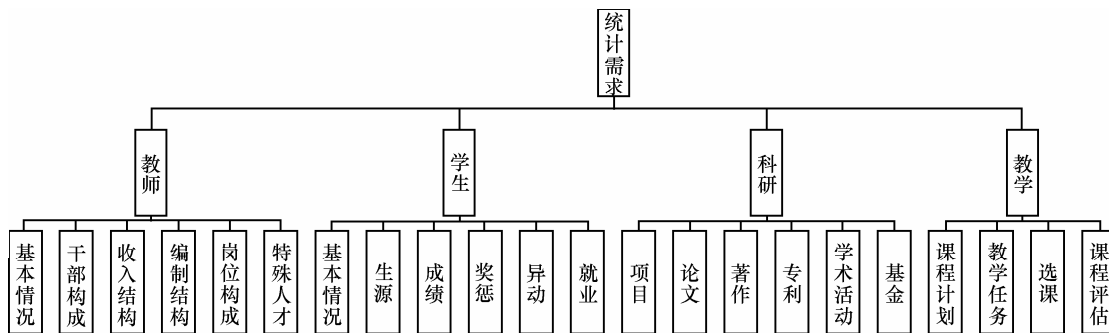


图 1 北京大学统计具体需求

3 平台设计

3.1 总体架构

图 2 是高校数据统计服务平台的总体架构。高校数据统计服务平台由源系统、数据仓库和统计平台三部分组成。

源系统：高校业务涉及的数据源比较广泛，主要有人事系统、教务系统、研究生系统、科研系统等，再加上校园网之外的一些其他外部数据源，构成了数据统计服务平台的数据基础，由于系统业务职能和具体需求不同，在实现时会选用不同的数据库，数据结构也可能存在较大差异，从而导致数据间有较大的异构性和不一致性。

数据仓库：数据仓库全面接收源系统数据，ETL 进程对数据进行规范化、验证、清洗，并最终装载进入数据集市，通过数据集市支持系统进行数据查询、分析；整个数据仓库包含四大层次^[6]。

1) 复制层(SSA, system-of-records-staging-area)：SSA 直接复制源系统的数据，尽量保持业务数据的原貌；与源系统数据唯一不同的是，SSA 中的数据在源系统数据的基础上加入了时间戳的信息，形成了多个版本的历史数据信息。

2) 原子层(SOR, system-of-record)：SOR 是基于模型开发的一套符合 3NF 范式规则的表结构，它存储了数据仓库内最细层次的数据，并按照不同的主题域对数据分类存储；高校数据统计服务平台根据目前部分需求将全校数据在 SOR 层中按人事、学生、教学、科研四大主题存储；SOR 是整个数据仓库的核心和基础，在设计过程中应具有足够的灵活性，以能应对添加更多的数据源、支持更多的分析需求，同时能够支持进一步的升级和更新。

3) 汇总层(SMA, summary-area)：SMA 是 SOR 和 DM 的中间过渡，由于 SOR 是高度规范化数据，因此要完成一个查询需要大量的关联工作，同时 DM 中的数据粒度往往要比 SOR 高很多，对要生成 DM 中的汇总数据需要进行大量的汇总工作，因此，SMA 根据需求把 SOR 数据进行适度的反范（例如，设计宽表结构将人员信息、干部信息等多个表的数据合并起来）和汇总（例如，一些常用的人头汇总、机构汇总等）；从而提高数据仓库查询的性能。

4) 集市层(DM, data mart)：DM 保存的数据是供用户直接访问的；可以将 DM 理解成最终用户直接最终想要看的数据；DM 主要是各类粒度的事实数据，通过提供不同粒度的数据，适应不同的数据访问需求；高校数据统计服务平台 DM 中的数据以 2 种不同类型存储：一类以星型模型建设，便于部门日常的灵活查询和统计，另一类按宽表以及重新组织的适应固定报表的表结构存储，便于高校的年统和季度统计工作。

统计平台：高校数据统计服务平台采用 B/S 架构的 3 层体系结构，即：数据操作层、逻辑层、表示层。

1) 数据操作层：充分考虑系统的高可用性，数据统计服务平台与数据仓库所使用的数据库互相独立，由此保证数据统计服务平台对数据进行加工处理时不会影响数据仓库中的数据；数据存取模块实现对数据统计服务平台数据的访问。

2) 逻辑层：分为报表预定义、报表查询、报表生成、报表填报、报表审核及报表汇总等模块；每个模块分别实现不同的功能；在统计平台中，不同身份的用户其功能权限和数据权限是不一样的：报

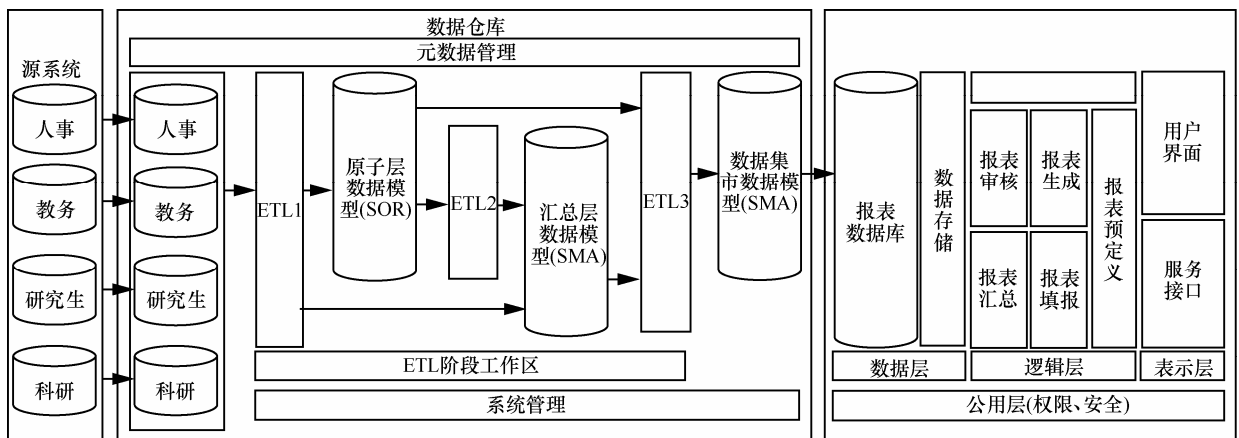


图 2 数据统计服务平台逻辑架构

表预定义是给系统管理员用的；报表生成、报表填报是给院系管理人员使用的，只能查看和操作本院系的数据；报表审核、报表汇总是给学校相关部门的管理人员用的，可以操作全校数据；功能权限和数据权限通过公用层与北京大学身份认证服务平台对接，统一进行管理。

3) 表示层：提供交互界面给用户使用，此外还提供一些服务接口供其他系统调用。

3.2 数据仓库建模

目前较为流行的数据仓库的建模方法较多，常用的有 Inmon 所提倡的范式建模法和 Kimball 所提倡的维度建模法^[7]。维度建模法针对各个维做了大量的预处理，通过这些预处理能够极大地提升数据仓库的处理能力，相对于范式建模法来说，在性能上占据了明显的优势；同时维度建模非常直观，紧紧围绕着业务模型，可以直观地反映出业务模型中的业务问题。不需要经过特别的抽象处理即可以完成维度建模。因此高校数据统计服务平台的数据仓库采取维度建模的方式构建。维度建模法采用事实表——维表的方式来构建数据仓库，数据集市、事实表存储实际的数据，维表存储事实表中对象的属性，事实表和维表的关联关系有“星型模型”、“雪花模型”和“混合模型”3种，常用的是“星型模型”。数据建模的步骤如下所示。

1) 结合具体需求确定分析主题，结合高校主要业务定义了一个公共维度主题和人事、学生、教学、科研4个业务主题：公共维度包含时间维、地理维、国标及校标，时间维和地理维在不同的应用场景可以使用视图形式转换为具体的分析维度，国标和校标主要用来解决在数据集成过程中的一致性问题；人事主题核心内容是教师的基本情况，具体分析主体有收入、岗位、职称以及杰出人才等；学生主题核心内容是在校生基本情况，具体分析主题有招生、成绩、奖惩、异动、就业等；科研主题主要分析全校师生科研成果完成情况，根据实际业务可以纳入所有科研成果，如项目、论文、著作、专利、学术活动等；教学主题以教学活动相关内容为主，如课程计划、教学任务、选课、教学工作量等。

2) 确定分析粒度，通俗地说就是分析对象的详细程度。为了满足分析的可扩展性及需求的多样性，以最小粒度来设计数据模型总是能达到最好的

分析效果，如：记录每个学生的明细情况、记录每项科研成果的详细情况。

3) 设计维表，维度是统计和分析数据的角度，与统计查询的参数相对应。在选取维度时应该将实体作为一个对象，把与该对象相关的所有重要属性都提取出来作为独立维度。

4) 设计事实表，为了跟踪具有生命周期的活动数据的变化过程以保留历史信息，设计事实表时使用缓慢变化维的方法以捕获变化数据^[8]。事实表中的版本、开始时间和结束时间3个字段是实现缓慢变化的核心。版本表示同一事物历史状态的顺序，开始时间和结束时间表示在该段时间内该事物处于某一状态，每一条数据的结束时间等于新数据的开始时间，这样该事物不同时间段的状态就分布在一条时间轴上，从而可以得到任一时间点该事物的状态信息。

以教师基本情况为例，其星型模型如图3所示。

3.3 统计数据模型

数据统计服务平台数据统计的结果展示给用户的最主要形式是报表。对于纷繁复杂的数据统计需求，对报表进行层层解析，明确报表含义、加工规划及数据来源，最终完成数据统计。根据北京大学的实际情况，设计了数据统计的处理流程，分为以下四步。

1) 确定统计相关的星型模型，即数据统计针对的是事实表中间的哪些事实，涉及到哪些统计指标，统计的粒度如何。

2) 确定报表中具体的每一行和每一列分别代表的统计指标，统计指标简单地讲即维度取值，每个统计指标对应到维表中是某个维度取某个值，也有可能是多个维度取值的累加。

3) 确定单元格的统计方法，每个单元格的统计指标应该是其对应的行、列所代表的维度取值的并集。

4) 将维度转化为可执行查询的语句，去事实表中查询出相应的统计数据 and 事实数据，为了方便，在数据统计服务平台的报表数据库中还可以将查询到的统计结果固化，以数据库表的形式存储下来。

基于此步骤设计的统计数据模型如图4所示。将1张业务报表拆分为5张配置表，它们分别是指标表、维表、维值表、行列表、指标维度组合表；最后计算得到的结果存储在指标数据表中。

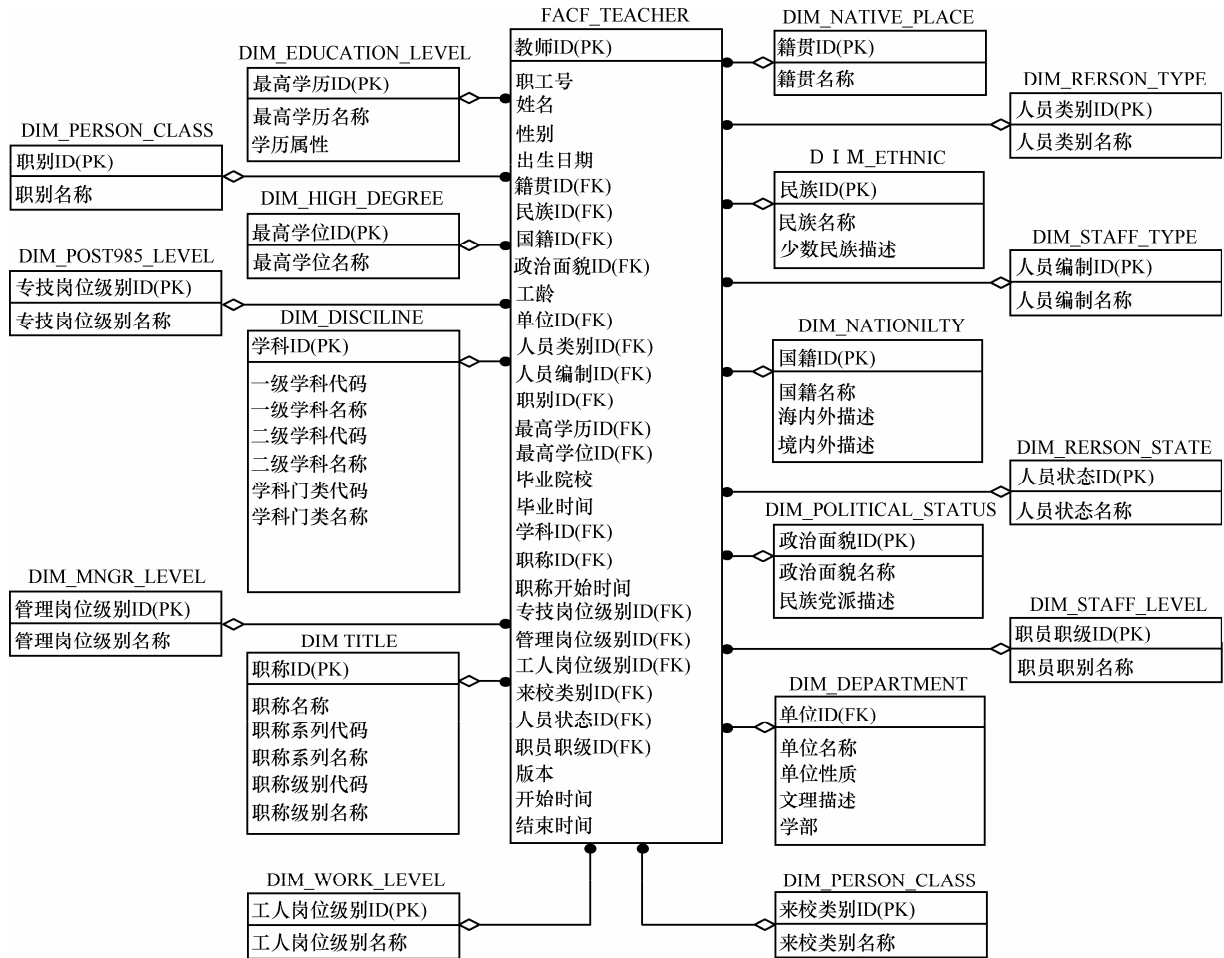


图 3 教师基本情况主题星型模型

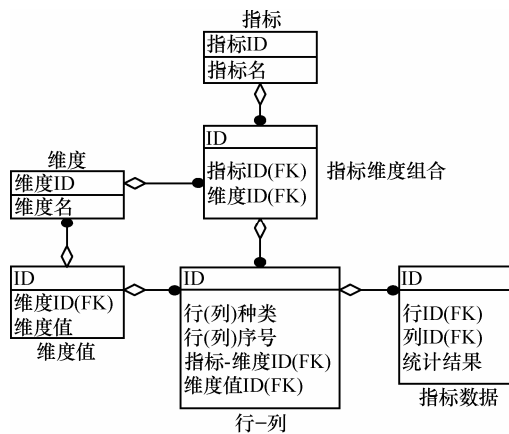


图 4 统计数据模型

4 平台实现

实现北京大学数据统计服务平台采用的是增量开发的方式，目前先基于学生和教师的数据搭建了数据仓库和统计平台，后续需要根据需求不断丰富数据仓库和统计平台的数据：一是要不断地根据

需求改进和完善现有数据在数据仓库中的组织方式，二是要不断地将科研、教学等业务系统中的数据纳入到数据仓库和统计平台中，形成新的数据集市和统计数据。

4.1 ETL 规则设计和实施

ETL(extraction transformation loading)负责将分散的、异构数据源中的数据抽取到临时中间层后进行清洗、转换、集成，最后加载到数据仓库或数据集中。ETL 是实施数据仓库的核心和灵魂，ETL 规则的设计和和实施约占整个数据仓库搭建工作量的 60%~80%^[9]。

数据抽取包括初始化数据装载和数据刷新：初始化数据装载主要关注的是如何建立维表、事实表，并把相应的数据放到这些数据表中，在数据仓库建模小节中已经做了详细介绍；而数据刷新关注的是当源数据发生变化时如何对数据仓库中的相应数据进行追加和更新等维护，北京大学数据统计服务平台采用了触发器方式（又称为快照式）来实

现数据刷新，具体来说就是：在 SSA 层需要抽取数据的数据表上建立插入、修改、删除 3 个触发器 (trigger)，每当源数据库中数据表的数据发生变化时，复制到 SSA 的数据也会相应发生改变，相应的触发器将变化的数据写入一个临时区 (buffer)；在数据库层定义了一系列的作业 (job) 和存储过程 (procedure)：作业规定了包括数据刷新频率和数据刷新先后次序在内的一系列任务调度策略，调用相应的存储过程从临时表中抽取需要刷新的数据，临时表中抽取过的数据被标记或删除；触发器方式的好处是：数据抽取的性能高、规则简单，对于编程人员来说易于上手，特别适合北京大学数据仓库现有规模还较小的特点，是一种简单易行的好办法；但随着以后数据仓库规模的越来越大，数据表越来越多，需要编写的触发器、存储过程和作业就越来越多，可能会不利于管理，今后是否还要采用这种方式还需要根据实际情况做进一步的研究。

数据清洗主要是针对源数据库中出现的二义性、重复、不完整、违反业务或逻辑规则等问题的数据进行统一的处理，表 1 列出了北京大学在对业务系统进行数据清洗时发现的几类最常见的问题及针对这些问题所采取的策略。

数据转换主要是为了将数据清洗后的数据转换成数据仓库所需要的数据：来源于不同源系统的同一数据字段的数据字典或者数据格式可能不一样，在数据仓库中需要给它们提供统一的数据字典和格式，对数据进行归一化；另一方面，数据仓库所需要的某些字段的内容可能是源系统所不具备的，而是需要根据源系统中多个字段的内容共同确定，例如，数据仓库中的人员类型“事业单位专业技术人员”实际上是根据人事表中“编制类型=事业单位”、“岗位级别=985”并且“人员类别=在职职工”等多个字段的内容共同得出的，像这样的

字段的形成也依赖于数据转换。

4.2 前台展示界面的实现

常见的数据仓库的前端展示工具有 BO、Cognos 等，能基于 Web 的直观界面，能提供报表、图表、仪表盘等多种展示方式。但都是商业产品，价格比较昂贵。ExtJs 是一款开源的创建前端用户界面，是一个基本与后台技术无关的前端 ajax 框架，具有功能强大、编程简单的特点，数据统计服务平台的用户界面基于 ExtJS 开发。

图 5 是北京大学数据统计服务平台统计报表展示的详细界面，该界面分成 2 个部分，上半部分是统计数据展示区，展示的是报表的详细内容，所有数据均已经自动生成；下半部分是细节数据展示区，若在报表中选择某个单元格，可以在细节数据展示区查看细粒度的详细数据，并提供将这些数据导出的功能。

5 平台部署

部署情况如图 6 所示。数据仓库的组成主要包括 ETL 服务器、数据仓库处理服务器、备份服务器与磁带库、数据管控服务器和调度服务器；高校数据统计服务平台自身包括数据库服务器、Web 应用服务器。在此架构方案中，充分考虑系统的高可用性，高校数据统计服务平台与数据仓库所使用的数据库服务器互相独立，由此保证二者在加工处理时不会相互影响。

6 结束语

本文提出了基于数据仓库技术的高校数据统计服务平台，通过合理的架构设计、科学的数据建模实现了对数据的集中存储、加工，以及统计数据生成、统计数据查询等功能。该系统能够有效地满足高校新形势下的业务发展需求，对于促进高校数

表 1 北京大学业务系统数据清洗的常见问题及策略

主要问题	表现形式	产生原因	清洗策略
完整性问题	大量空值字段的出现	源系统中对很多字段没有做非空限制	1) 交由源系统重新录入、补齐 2) 在数据仓库对应的维表中建立一个新的字段，将这些空字段的值统一赋值
	超出字典表范围	填写这些值的时候是直接让用户填写而非下拉框选择	1) 交由源系统重新录入、补齐 2) 在数据仓库对应的维表中建立一个新的字段，将这些空字段的值统一赋值
一致性问题	一个特定的字段在不同的表中内容不同	录入，同步的问题	1) 交由源系统重新录入、修改 2) 选取最可靠的表中的字段为确定值，
	应该成为主键的值不唯一	源系统中未建立有效的主键关系	消除错误、重复的主键

第一表：2013年上半年北京大学党员队伍状况统计表（一）——总人数

总计	行政管理 人员	专职 教师	女	其中								其中 35岁 (含) 以下 青年 工人	其中 35岁 (含) 以下 青年 工人	附 中、 小 教 工	高 退 休 人 员	学 生	博 士 生	硕 士 生	本 科 生	其中					非 在 编 人 员
				职称状况				其他 专业 技术 人员												本科生中					
				教授	副教授	讲师	其他	其他	其他	其他	其他									其他	其他	其他	其他	其他	
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	
总人数	264	3	36	7	5	24	5	6	1	5				2	165	60	105								53

统计数据显示区

查询结果

导出	工工号	姓名	性别	民族	出生日期	院系	所在支部	入党成熟度	党员状态	人员状态	入党时间	职称	年级	学历	人员类型	人员类别
1			女性	汉族	1989-12-07	国家发展研究院	国家发展研究院...			正常			2013	本科毕业	学生	博士生
2			女性	汉族	1990-12-17	国家发展研究院	国家发展研究院...			正常			2013	本科毕业	学生	博士生
3			女性	汉族	1991-04-20	国家发展研究院	国家发展研究院...			正常			2013	本科毕业	学生	博士生
4			男性	汉族	1990-12-12	国家发展研究院	国家发展研究院...			正常			2013	本科毕业	学生	博士生

统计数据显示区

图 5 北京大学数据统计服务平台统计报表展示界面

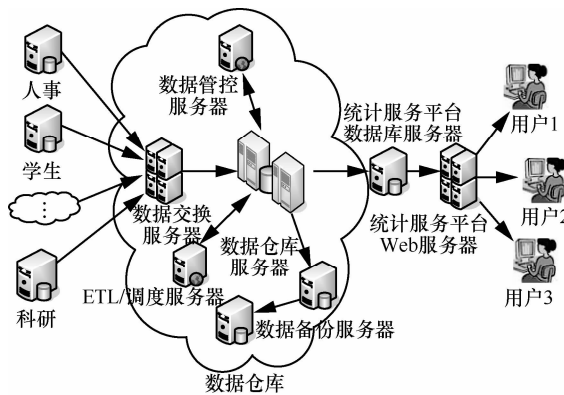


图 6 高校数据统计平台物理架构

据集约化管理水平的提升、搭建数据统筹管理和决策支持服务的长效机制框架具有十分重要的意义。

参考文献：

[1] DRESNER H. Profiles in Performance: Business Intelligence Journeys and the Roadmap for Change[M]. New York: John Wiley and Sons, 2009.

[2] 罗永辉, 陈明亮. 商业智能哲学思索-BI 框架体系的另一种视角[J]. 软件世界, 2007, 7:68-71.

LUO Y H, CHEN M L. The philosophic thinking of business intelligence-another view of the architecture of BI[J]. Software World, 2007, 7:68-71.

[3] WASSERMAN T J, MARTIN P, SKILLICORN D B, et al. Developing a characterization of business intelligence workloads for sizing new database systems[A]. Proceedings of the 7th ACM International Workshop on Data Warehousing and OLAP[C]. 2004. 7-13.

[4] INMON W H. Building the Data Warehouse[M]. New York: John Wiley and Sons, 2003.

[5] 王倩宜, 李丽, 欧阳荣彬等. 数据综合服务管理平台主体方案探讨[J]. 实验技术与管理, 2011, 28(5):4-6.

WANG Q Y, LI L, OUYANG R B, et al. Discussion on infrastructure proposal of data integration and service platform[J]. Experimental Technology and Management, 2011, 28(5):4-6.

[6] 马宁, 曹震, 李军. 基于数据仓库的银行监管报表系统研究[J]. 软件导刊, 2013, 12(4):137-141.

MA N, CAO Z, LI J. The research on the bank superintendency re-

porting system based on data warehouse[J]. Software Guide, 2013, 12(4):137-141.

[7] KIMBALL R, ROSS M. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling[M]. New York: John Wiley and Sons, 2002.

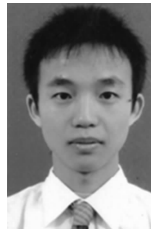
[8] 高亮. 高校数据仓库建设研究[J]. 武汉大学学报(理学版), 2012, 58(S1):179-184.

GAO L. The research on the construction of data warehouse in universities[J]. Journal of Wuhan University(Natural Science Edition), 2012, 58(S1):179-184.

[9] 王亦然. 数据仓库技术基于业务报表数据的应用研究[D]. 北京: 中国地质大学(北京), 2010.

WANG Y R. The Research of Application Business Reporting Data Based on Data Warehouse Technology[D]. Beijing: China University of Geosciences (Beijing), 2010.

作者简介：



龙新征 (1984-)，男，湖南湘阴人，北京大学工程师，主要研究方向为数据库技术、数据仓库、高校教育信息化等。

李丽 (1980-)，女，湖北黄冈人，北京大学工程师，主要研究方向为数据库应用技术、数据仓库、高校教育信息化。

彭一明 (1984-)，男，湖南澧县人，北京大学工程师，主要研究方向为数据库应用技术、高校教育信息化。

王倩宜 (1972-)，女，辽宁大连人，北京大学高级工程师，主要研究方向为数据库应用技术、高校教育信息化。

欧阳荣彬 (1979-)，男，湖南宜章人，北京大学高级工程师，主要研究方向为数据库应用技术、高校教育信息化。

刘云峰 (1973-)，女，山东莘县人，博士，北京大学高级工程师，主要研究方向为数据库应用技术、高校教育信息化。