



电信企业通用数据服务平台的设计与实现^{*}

潘 俊¹, 程建和²

(1. 温州大学信息安全研究所 温州 325035; 2. 中国电信股份有限公司温州分公司 温州 325001)

摘 要: 基于.NET框架设计并实现了电信企业通用的数据服务平台,通过ETL过程对业务数据进行清洗和抽取,建立数据集市;采用SOA架构开发数据服务中间件,作为任务调度和数据集查询的代理;通过可快速配置的自定义智能查询方案,自动生成向导式的用户界面。该平台具有配置快速、安全稳定等特点,可同时为一线人员和经营决策人员提供支撑,已成功投入运行。

关键词: 数据服务;抽取—转换—装载;数据集市;联机分析处理

doi: 10.3969/j.issn.1000-0801.2013.02.021

Design and Implementation of General-Purpose Data Service Platform for Telecom Company

Pan Jun¹, Cheng Jianhe²

(1. Institute of Information Security, Wenzhou University, Wenzhou 325035, China;

2. Wenzhou Branch of China Telecom Co., Ltd., Wenzhou 325001, China)

Abstract: A general-purpose data service platform for telecom company based on .NET framework was designed and implemented. The platform carries out the business data cleansing and extraction process and then generates the data mart via ETL tools. System users interact with the data mart via SOA-based data service middleware, which acts as a job scheduler and cube query executor. Particularly, the platform generates a wizard style interface as defined in XML configuration files. Characterized by quick configuration, safeness and flexibility, the system can serve the frontline staffs and senior analytical managers at the same time, and has been successfully applied in a telecom company.

Key words: data service, ETL, data mart, OLAP

1 引言

随着电信企业各业务支撑系统建设的不断深入,企业的业务数据已初现规模,积累速度也不断提升。如何有效整理各个业务管理系统中积累的数据,为各级人员提供统一的信息视图,实现真正的信息共享,从而提高企业经营分析的决策水平,已成为电信企业的迫切需求,商业智能

(business intelligence, BI)^[1,2]技术正是能满足此特定需求的有效手段。

商业智能技术综合运用数据仓库 (data warehouse, DW)、数据集市 (data mart, DM)、联机分析处理 (online analytical processing, OLAP) 等技术来收集、管理和分析结构化和非结构化的业务数据信息。其中,数据仓库是面向主题的、集成的、相对稳定而又反映历史变化的数据集,

^{*} 浙江省科技计划基金资助项目 (No.2012C33086), 浙江省自然科学基金资助项目 (No. LQ12F02008), 温州市科技计划基金资助项目 (No.G20100196)

它集成了多个异构数据源并按照主题进行重组,主要面向数据分析任务^[3]。数据集市中的数据具有数据仓库中数据的特点,可以看作数据仓库的子集,一般面向部门、业务单元或特定应用,因而规模较小,便于快速实现^[4]。联机分析处理是数据仓库系统的主要应用,支持复杂的分析操作,侧重决策支持,并且提供直观易懂的查询结果,其目标是满足决策支持或满足在多维环境下特定的查询和报表需求^[5]。近年来,借助于商业智能技术,设计和开发人员在海量数据处理、业务分析和信息展现等多个方面做了大量的工作^[1,2,4-7],但是所开发的数据应用分析平台仍存在可扩展性差、后期维护困难、操作复杂等问题^[7]。

本文在充分调研电信企业需求的基础上,针对电信企业数据服务应用系统开发存在的问题,设计并实现一种支持快速开发的基于数据集市的通用数据服务系统,可作为企业进行商业智能分析的良好依托平台。系统使用数据清洗技术对海量数据进行采集、校验、装载和存储,建立起面向主题的数据集市;采用 SOA 架构提供后台调度、数据集查询、数据下载等关键服务,便于功能扩展;针对本地业务频繁变化的现实需求,提出基于 XML 的 OLAP 配置和动态报表配置方案,实现配置文件和数据集市的关系映射,可快速定制业务模型以响应需求变化。为了简化操作,系统采用向导式界面,引导用户完成智能查询的定制过程,并采取分级数据权限访问和记录系统日志的方式来满足安全性的需求,本系统的实施可解决非技术人员定制业务模型的难题。

2 系统设计

2.1 系统设计原则

本系统的设计目标是提供一套通用的数据分析和业务建模应用方案,支持底层清单级别和高层统计级别需求,满足多层次数据使用人员的个性化需求。在设计中主要遵循如下几个原则。

(1) 即插即用

由于本地业务需求频繁多变,本系统应具有快速部署的能力,业务人员能够利用现有的开发知识快速定制业务模型。为此,系统采用快速配置方案实现对业务需求改变的快速响应。系统使用 XML 作为数据集配置和维度配置的元数据描述语言,可灵活使用数据集原表。同时,通过向导式界面将配置自动转化为智能取数界面,方便非技术人员进行业务定制。

(2) 统一数据访问

采用基于 SOA 架构的数据访问中间件技术,定义数据集市查询语言 (DMSQL), 并实现 ETL (extract, transform, load, 抽取、转换和装载) 调度服务和查询调度服务这两个核心服务,提供一个可扩展的框架,便于日后的任务类型扩展。

(3) 清单级别数据支撑

通过自定义数据集、维度及查询条件提供清单级别数据,解决以往高层经营分析决策人员面对的汇总视图缺乏清单级别数据支撑的问题,增强一线数据支撑人员对经营决策的信任,并使得数据整体上保持一致。

(4) 统计级别的快速分析和多维展示

对底层清单级别数据进行汇总分析后,提供图表性、前瞻性的高层次视图,使得数据分析人员能获得即时有效的决策数据。

2.2 系统整体架构

本系统采用基于 Microsoft .NET 平台的 B/S 分层架构,其中表示层主要负责与用户的交互及数据分析结果展现;业务层主要负责配置文件和数据集市的映射转换;数据存取层主要负责数据查询和分析。系统的设计思路是:首先制定 ETL 方案,使用 ETL 工具对各业务数据源产生的海量数据进行清洗和抽取,建立面向特定主题的数据集市;接着设计实现数据服务中间件,提供任务调度、数据集查询等关键服务。最后,构建数据服务平台,通过可快速配置的用户自定义智能查询方案,自动生成向导式的智能取数平台,并提供清单视图、决策统计视图、数据多维展示等商业智能分析功能。基于以上分析,系统整体架构主要由数据源、数据集市、数据服务应用系统 3 部分组成,3 者之间分别通过 ETL 和数据服务中间件交互,系统整体架构如图 1 所示。

从图 1 中可以看出,数据集市的数据来源于多个系统,包括数据仓库、CRM 系统、BSS、计费系统和其他业务系统,其中数据仓库是主要数据来源。系统采用自行开发的 ETL 系统定时从源数据库中抽取数据,并归纳整理为维表和若干基础表;同时根据调度方案,依赖 SQLPlus 脚本自动将基础数据汇总整理为相应的宽表和事实表以供查询。经过维度化处理的数据包括若干维表和事实表,可转化为多维数据集,通过 PowerPlay 多维展示平台满足经营分析决策人员的快速分析需求。

数据服务应用系统主要实现了不同粒度的数据提取、统计和展现功能。粒度反映了数据集市按照不同的层次组

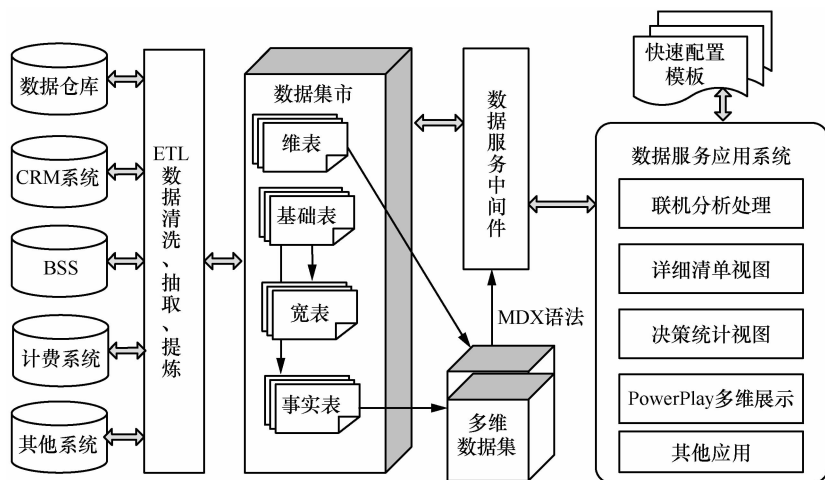


图1 数据服务平台整体架构

织数据,根据不同的查询需要,存储不同细节的数据。粒度越小,数据越细,查询范围就越广泛。相反,粒度级别越高,表示细节程度越低,查询范围越小。详细清单视图面向明细粒度的数据查询,主要满足一线数据支撑人员对明细数据的需求,统计数据面向汇总级粒度的数据查询,主要满足高层人员对汇总的高层数据图表的需求。对于详细清单查询,从数据集市中的数据集(cube)原表中直接提取数据。这种方法可灵活使用数据集原表,性能比较容易控制,能利用现有开发知识迅速启动,但需要使用大量的代码定义数据集、维度以及数据集查询,本系统通过XML配置文件较好地实现了该方案。对于统计数据查询,使用钻取(drill)技术从多维数据集中提取数据,钻取包含向下钻(drill-down)和向上钻(roll-up)操作,下钻指从概括性的数据出发获得相应的更详细的数据,上钻则相反。这种方法要求熟悉MDX语法,优点是操作规范,能充分利用分析服

务(analysis service, AS)的功能,并可以使用ADOMD.NET提供的完整的服务来获取和操作多维数据。

3 系统实现方案

3.1 ETL的实现

ETL过程在数据集市的建设与维护中占据了相当重要的位置,负责抽取、转换数据源的数据并装载入数据集中。降低ETL过程设计、维护代价以及提高ETL过程的执行效率,是ETL方案成功的关键。本系统提出了支持多种脚本的可扩展的数据抽取方案,该方案由客户端程序、Web服务层、调度数据库、服务端程序组成,支持Perl脚本、SQL脚本和CMD批处理命令,其架构如图2所示。

在此架构下,首先根据业务需求开发数据抽取脚本,目前支持Perl脚本、标准SQL脚本和CMD批处理命令,并可根据需要扩展。然后通过ETL配置工具定义作业信

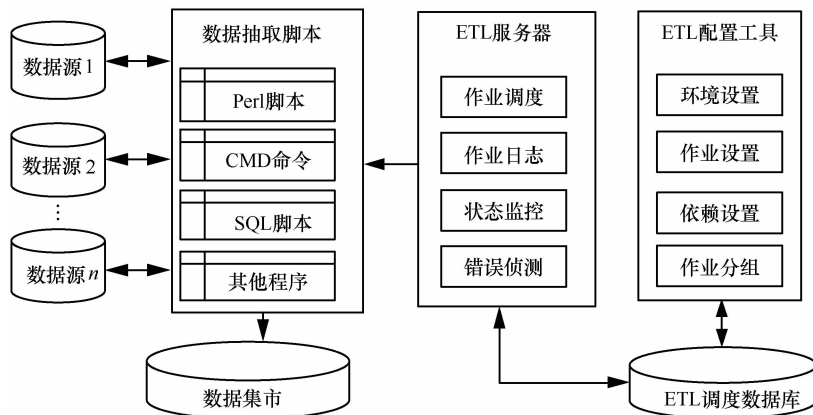


图2 ETL架构

息、作业分组、作业依赖等配置,最后 ETL 服务器程序将按照客户端的配置定时调度相应的脚本,生成增量数据存储在数据集中。ETL 系统的业务代码以存储过程的形式封装在调度数据库中,提高了系统的数据响应速度和数据安全性,有利于后期的系统维护工作,同时,在中间层中用 Web Service 构建业务逻辑层,为 ETL 配置工具和 ETL 服务器程序及其他应用程序提供数据接口,提高了系统的可扩展性。

3.2 数据服务的实现

系统采用 Microsoft 的 Web Service 作为 SOA 架构的实现方案,设计了后台调度、信息通知、数据下载和数据集查询 4 种关键服务,图 3 给出了后台调度查询服务的时序。

后台调度服务是通用的定时调度任务服务,提供定时执行多种任务的能力,支持 SQL 查询、SQLPlus 脚本、PERL 脚本、存储过程等接口,并提供一个可扩展的框架支持日后可能引入的其他接口。下载服务将指定数据集转换为多种格式的文件供用户下载,目前支持 access、csv、tsv、dbf 格式;出于性能和扩展性的考虑,本系统使用 Serv-U 提供 FTP 文件下载服务,并提供 Web 下载方式作为备选方案。通知服务将指定信息以邮件、短信等方式发给订阅人。数据集查询服务是最主要的服务,它将 XML 格式的查询配置协议转换为 SQL(或 MDX 语言)后进行查询,出于移植上的考虑,该服务的接口支持传统数据库和多维数据库查询,已实现 3 种形式的查询服务。

- SelectSql:后台以调度的方式处理提交的 SQL 脚本,并生成数据表保存查询结果。
- DirectSelectSql:将查询先存储为记录而不是实时调度处理,仅当用户请求数据时才临时处理,不生成物理表,仅提供 DataSet 形式的数据返回结果。

- DataProvider:将 DataProvider 的 XML 配置保存为记录,提供 DataSet 形式的数据返回结果。

3.3 数据服务应用平台的实现

3.3.1 联机分析处理

联机分析处理的技术核心是“维(dimension)”这个概念。“维”是人们观察客观世界的角度,是一种高层次的类型划分。通过把一个实体的多项重要的属性定义为多个维,用户可以对不同维上的数据进行比较,一般一个多维数据集可以用一个立方体(cube)的方式进行描述。

本系统中的联机分析处理配置包括两个方面:维度配置和数据集配置。利用维度配置可定义数据集所使用的数据源以及各种分析所关心的维度(如日期、产品、优惠等)。数据集是若干维度字段、普通字段和量度字段的集合,在本系统对应到一个具体的数据表、视图或 AS 数据集。利用数据集配置可以定义各种主题的数据集,并可使用各种维度作为取数的过滤条件。配置完成并部署后,使用数据集应用系统的“自助查询”功能,可将该配置自动转化为基于向导的自助式的智能取数界面,取数步骤如下。

(1)根据数据集配置生成各种主题的数据集。

(2)选择所关心的数据集并根据数据集维度列设置查询条件。

(3)从配置的数据集普通列选择需要展示的列,此维度列也可作为输出列而被选择。

(4)系统根据前 3 步的操作,自动生成标准 SQL 语句(或 MDX 语句),并可以接收更为复杂的参数和调度功能设置。配置通知消息,在此步将记录通知服务的内容。

(5)系统把配置好的查询发送给后台查询服务器,若配置了通知消息,系统会将通知消息发送给通知服务器。然后跳转到等待页面,等待后台查询服务的调度运行。

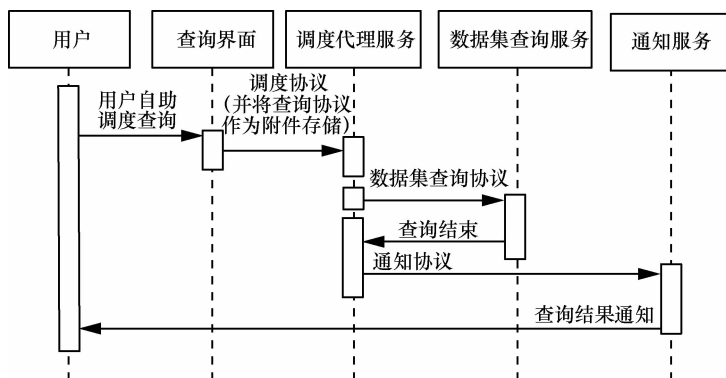


图 3 后台调度查询交互时序



首页

自助查询

我的作业

固定报表

图形报表

301.月底收入曲线

302.日收入报表

097.产品到达数曲线

098.产品受理量曲线

099.三大主量发展情况

303.补收退款统计

发展量报表

营销报表

营业报表

财务月报表

服务质量报表

清单下载

经营分析月报

定制报表

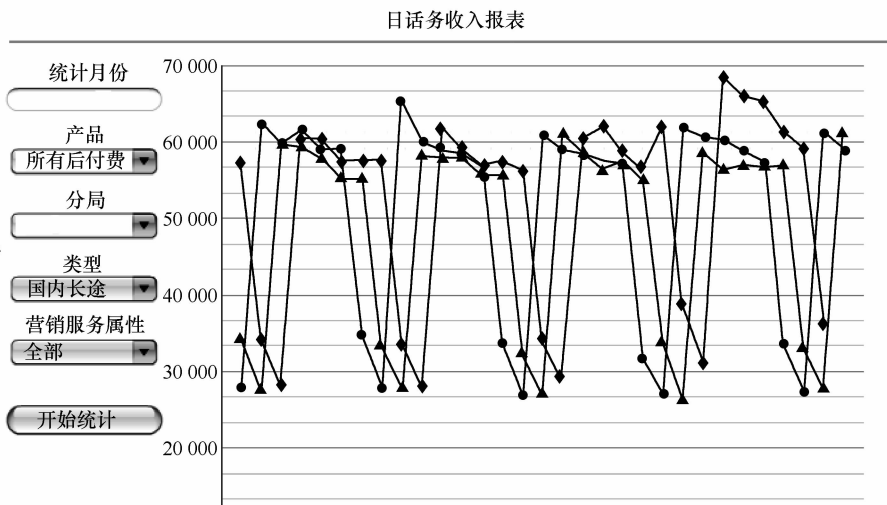


图4 报表示意

(6)查询结束后,调用下载服务,可将数据转化为指定的文件类型提供给终端用户。

可以看出,通过将配置简易性和用户良好体验有机地结合在一起,系统维护人员和用户的操作量都减少到最大限度。

3.3.2 报表配置的实现

本系统提供灵活的架构进行快速报表定制,可支撑丰富的报表类型和数据来源类型。系统提供4种报表的展现方式。

- 水晶报表,适合于做固定格式的交叉报表。
- Xcelsius 报表,提供 flash 形式的图形报表。
- GridView 报表,适合展示运算时间短,且数据量较少的二维表格数据。
- 需 UI 呈现,且运算时间长或数据量大的报表,通过委托由后台查询服务进行处理。

同时,针对数据集市的特点,可以考虑用数据集市的维表作为报表的条件列进行展现。图4给出了一个报表示意。

针对多源数据的异构问题,本系统使用 Provider 模式屏蔽底层异构数据,对上层提供统一接口。如果要增加新的数据格式,继承 Provider 基类并实现 Execute()虚方法即可,这种结构大大提高了对其他数据源的可扩展性。本系统实现了3类数据源。

- DbCommandDataProvider(通用的数据库查询数据提供者),它支持 Oracle、Sqlserver、OleDb、ODBC 等各种数据库访问类型,并支持文本和存储过程两种方式的数据获取途径。返回的数据类型可以为 IDataReader、DataSet 以及单值。

- OracleCommandDataProvider,Oracle 专门版本,为访问 Oracle 做了专门优化。
- QueryServiceDataProvider,用于获取已配置好的查询服务的查询结果。

4 结束语

本文针对电信企业数据集市应用存在的问题,设计并实现了通用独立的数据服务平台。该平台通过 ETL 调度,可对相关业务系统数据进行清洗和抽取,建立面向特定主题的数据集市;通过数据访问中间件,可提供任务调度、数据集查询等服务;通过可快速配置的用户自定义智能查询方案,可自动生成向导式的智能取数界面。基于本文工作的数据服务系统已经在中国电信股份有限公司温州分公司成功投入运行,系统解决了前端取清单困难问题,开发了完全基于配置的报表模块,大大提高报表开发效率,实现了快速支撑,成为支撑前端考核、通报、日常报表、营销分析、评估的核心系统,真正实现了基层和对前端的有效支撑。本平台可作为基于数据集市的数据服务应用实施方案的良好参考,以本平台为基础,开放相应的通信接口,还可以实现多系统报表数据整合。

参考文献

- 1 Popovic A, Hackney R, Coelho P S, et al. Towards business intelligence systems success: effects of maturity and culture on analytical decision making. Decision Support Systems, 2012, 54(1): 729~739

(下转第134页)