

doi: 10.12052/gdutxb.170015

主数据管理技术在企业信息集成中的应用研究

林 穗, 李煜臻, 孙为军

(广东工业大学 计算机学院, 广东 广州 510006)

摘要: 随着大数据的广泛应用, 政府对企业信息的精准管理亟待加强. 大数据背景下, 针对企业海量数据多源异构的问题, 基于主数据关键性、独一性以及长期有效性的特征, 提出了构建主数据平台是实现企业信息集成的最佳方式. 通过主数据“多数一源, 一源多分”的原则, 构建完整一致、集中统一的企业信息集成机制, 建立符合企业信息规范的数据管理体系, 实现企业基础信息的全面共享和统一数据的分发, 从而使政府决策部门全面、动态、准确地了解和掌握企业登记注册和生产经营等方面的情况.

关键词: 多源异构; 主数据; 主数据管理; 企业信息集成

中图分类号: TP311 **文献标志码:** A **文章编号:** 1007-7162(2017)03-0067-05

A Research on the Application of Master Data Management Technology in Enterprise Information Integration

Lin Sui, Li Yu-zhen, Sun Wei-jun

(School of Computers, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: With the wide application of big data, accurate management of enterprise information needs to be strengthened. Under the background of big data, aiming at the problem of multi-source heterogeneous data, based on the criticality, uniqueness and long-term validity of master data, it is the best way to realize the enterprise information integration by constructing the master data platform. Through the principle of “multiple data one source, one source multiple distributions”, a complete, unified, centralized and unified enterprise information integration mechanism can be built, and a data management system established conforming to enterprise information norms, and realizing the comprehensive sharing of basic information of enterprises and the distribution of unified data, for government decision-making departments to understand and master comprehensively, dynamically and accurately the business registration and production and management.

Key words: multi-source heterogeneous; master data; master data management; enterprise information integration

随着企业的逐年增多, 同一家企业在政府各个部门如财政局、税务局、统计局的信息存在冗余以及不一致的现象^[1]. 首先, 异构数据源多且存放分散. 其次, 数据侧重于业务处理, 集成性差、可用性低. 为了协助政府决策部门进行宏观决策分析以及企业综合情况的分析, 实现系统集成、资源整合和信息共享, 需要一个综合平台来集成这些信息. 信息集成技术旨在实现异构数据系统的数据共享, 将异构数据源统一封装, 屏蔽数据源之间的差异性, 并提供统一的重构表达、数据管理以及存储分发^[2]. 现有的信息集成方法包括联邦数据库 (Federated Database System)、数据仓库 (Data Warehouse) 和基于中间

件模型的方法. 集协作与独立于一身的联邦数据库^[3], 各数据源则通过相互提供的接口进行访问, 但编写大量的接口程序会造成繁重的工作量, 因此只有在数据源数量较少时才适用. 数据仓库^[4]通常是对原有数据源中的局部数据进行移植, 虽然处理简单但数据量庞大, 也面临着数据更新周期长、数据重复存储的问题^[5].

而隶属于模式映射方法的中间件模式, 多为支持MQ和Web Service两种主流的适配器^[4], 由于该模式是对各数据源的数据进行提取、识别以及匹配整合, 避免了大量数据的存储, 因此适用于集成数据量大、更新速度快的数据源^[6]. 具有较为突出特点的中

间件模式,常常被学者将其与主数据^[7-8]相结合,由此出现了主数据管理^[9]这个新型数据管理方式,并应用到了生活中的各个领域.有学者基于开源主数据管理工具Talend实现了对高校主数据的管理,有效解决学校多个业务系统异构数据的集成问题^[10];也有学者利用主数据的特殊性质将其利用于物料管理、电网调度、轨道交通调度等多个方面,实现信息的高效整合与分发,提高数据的准确性以及利用率.对于企业信息中的主数据而言,它能够生成反映企业信息的关键部分,即便所需集成的各数据源非常庞大且都为异构时,也能保证高质量的企业数据,同时也借助于中间件来提高数据分发和接收过程中的安全性和稳定性^[11].因此将主数据管理引入到企业信息集成当中,相比于此前的信息集成技术,将会在很大程度上缓解企业信息冗余、不一致等问题.

根据关于主数据管理的前人研究工作成果,本文的核心研究目标如下:

- (1) 实现数据统一化:数据建模,信息集成;
- (2) 实现数据规范化:企业编码,唯一校验;
- (3) 实现数据接口分发:单点登录,统一门户.

1 企业信息集成主数据管理平台

1.1 平台框架

本企业信息集成主数据管理平台作为一个统一门户,在构建统一标准的规范体系的同时也采用了访问安全控制策略.各异构数据源中待集成的数据通过ESB总线传输至数据交换中间件,经过中间件集成后的数据将具有标准、一致的格式,该全局模式下的规范数据可通过总线提供给各大平台及工具,以实现企业集成信息的大数据分析、可视化以及企业画像等.图1是企业信息集成的主数据管理平台框架.



图1 企业信息集成的主数据管理平台框架

Fig.1 Framework of master data management platform for enterprise information integration

1.2 主数据及主数据管理

对于同一家企业而言,诸如企业名称、组织机构代码等信息都是唯一的,而由于采集数据缺失、部门间数据不对称等问题,会导致这些信息重复甚至错误存储在不同部门的数据库系统中^[12-13].为了解决上述问题,对企业的关键信息进行统一存储和管理,以保证数据的同一性和完备性,主数据随之而来^[14].

1.2.1 主数据

主数据(Master Data, MD)是企业信息中相对静态不变的数据,是以其独一性区别不同企业或区别企业其他业务数据的关键部分^[15].由于主数据是全局模式下的规范数据,因此它能够被各个职能部门反复使用和高度共享.企业主数据通常包括企业名称、组织机构代码、法定代表人等^[16],这些都属于企业信息的关键部分,在不同职能部门的数据库系统中都是必不可少且唯一的.

作为描述企业关键信息的主数据,应该具备的特性包括关键性、独一性以及长期有效性^[17].为了保证主数据的3种特性,使得企业的关键数据在不同的数据库系统中保有一致可靠的版本,就需要进行主数据管理.

1.2.2 主数据管理

主数据管理(Master Data Management, MDM)能够创建和维护权威的、一致的、安全的企业关键信息视图,并对其进行统一的存储管理,保持某种数据对外的单一视图.该视图能够反映企业关键信息的一个一致可靠的版本,形成共享的信息提供给需要此信息的其他职能部门所使用,使得企业的关键信息保持一致.

主数据管理应考察各个数据来源中的数据,通过制定严格规范的数据标准使主数据始终保持独一、有效.在主数据管理中,各异构源中的重复数据在此也需要完成识别和匹配整合,形成准确且持续有效的主数据.主数据管理主要包括数据模型管理、数据管理以及数据交换等.

首先,主数据管理对企业信息主数据进行统一存储和编码,不改变原有数据源系统的编码规则,最大限度地保护了数据源;其次,通过约束规则,除去了部分重复冗余现象,一定程度上确保了数据统一准确;并且,将数据通过统一数据接口分发,各部门所获取的企业关键信息一致,形成紧密的信息数据网络.

2 主数据模型

主数据模型即主数据管理的标准规范框架,是整个管理平台建立的中心,先建立好模型,再将数据按照标准规则导入模型中,这样才能更加规范地管

理数据.

2.1 模型创建

在主数据模型创建之前,首先根据实际政务需

要,将不同来源的多元异构数据建立统一标准体系规范,将名称、指标、计量单位等设置规范,并进行数据的初步处理.如表1所示.

表 1 标准规范指标示例
Tab.1 Example of standard specification index

表名	中文解释	字段名	字段中文解释	数据类型	指标规范名称	处理方式	对应一级指标名称
1	省属企业基本情况	F001_1	组织机构代码	VARCHAR2(9)	组织机构代码	名录比对	
		F001_2	企业名称	VARCHAR2(100)	企业详细名称	名录比对	
		F001_3	法定代表人	VARCHAR2(30)	法定代表人（国资）	并列使用	法定代表人
		F001_4	企业注册资金	NUMBER(22)	注册资本（国资）	并列使用	注册资本
		F001_5	企业资产总量	NUMBER(22)	资产总量（国资）	单独使用	
		F001_6	现有职工人数	NUMBER(22)	职工人数（国资）	单独使用	
		F001_7	企业经营范围	VARCHAR2(600)	经营范围（国资）	并列使用	经营范围
		F001_8	企业主业范围	VARCHAR2(200)	主业范围（国资）	单独使用	
		F001_9			是否为省属企业	新增	取值范围：是/否

接着根据规范指标创建模型,首先输入模型的基本信息,如模型代码以及模型名称等,然后进行属性、元属性等配置,具体流程如图2所示.属性是指模型的具体字段,如上述表1所示的组织机构代码、企业名称等.元属性则是属性的细分化,如企业产品类型

属性的具体类型分类.属性与元属性除了基本的名称代码,还需要配置类型、长度、取值方式、计量单位等.在创建好模型之后,进行模型的审核,检验模型是否成功准确建立.在审核之后,可以对模型进行查询以及变更.

2.2 编码规则

主数据作为整个省份所有企业情况信息的支撑数据,在整个管理平台中,必须保持数据的准确性以及一致性.在信息集成中,每条数据实体通常都用唯一一个特定编码来表示,从而来保证数据唯一以及格式统一.因此编码规则的制定是否合理科学,对于主数据管理平台的建立尤为重要.

鉴于政府数据敏感性以及安全性的考虑,采用了自动编码的形式.自动编码是指流水号形式的编码,编码本身不赋予任何意义,数据通过审核之后由系统自动生成.配置规则时,还需要配置编码类型、编码长度、编码步长等几项,这里设定编码类型为定长,编码长度为9位,编码步长为1,编码从000000001至999999999.配置好编码规则,数据导入模型经过之后,会形成统一格式的企业编码,示例如表2所示.

表 2 企业编码示例
Tab.2 The example of enterprise encoding

企业详细名称	企业编码	证券代码	证券简称
广州好莱客创意家居股份有限公司	000000001	603898.SH	好莱客
广州万孚生物技术股份有限公司	000000002	300482.SZ	万孚生物
胜宏科技（惠州）股份有限公司	000000003	300476.SZ	胜宏科技
深圳四方精创资讯股份有限公司	000000004	300468.SZ	四方精创

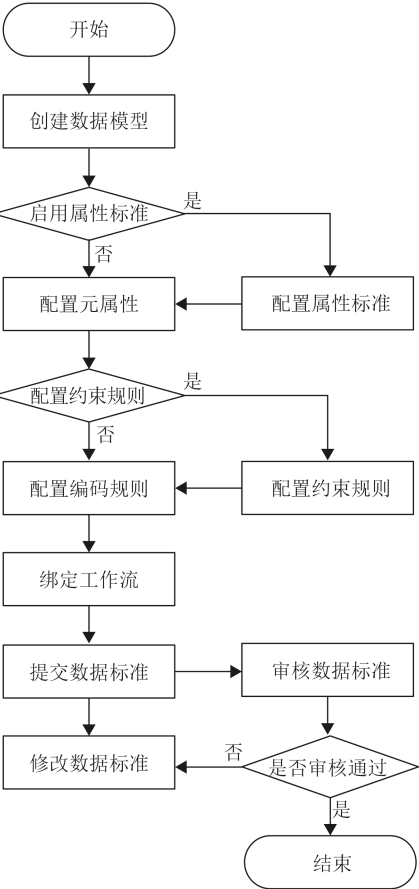


图 2 模型创建流程图

Fig.2 The flow chart of model creation

2.3 约束规则

为了保障数据的唯一性、准确性以及格式统一性,通过配置约束规则,对数据进行初步校验清洗.约束规则主要约束主数据模型中的元属性,分为唯一性校验与关联性校验两种.

2.3.1 唯一性约束

唯一性约束是用来保证数据的唯一性.需要配置以下几项.

(1) 条件表达式:对当前唯一性约束进行条件判断,配置约束的生效条件及作用范围.如果满足条件表达式则进行校验,否则不进行校验.

(2) 值类型:配置进行唯一约束的属性值类型,配置元属性只显示对应类型的.

(3) 区分大小写:控制校验时,对于字母是否要区分大小写.例如,当选“是”时,A和a为两个完全不同的数据;当选“否”时,A和a则是同一个数据.

(4) 校验规则级别:包括“错误型校验规则”和“警告型校验规则”,前者是一旦校验失败不允许数据进入系统,后者是校验失败仅有提示信息,数据依然可以正常进入系统管理.

2.3.2 关联性约束

关联性约束是指通过元属性间的关联关系,校验数据是否填写规范,满足业务要求.同样是通过条件表达式进行设定,检验元属性的值是否在上下限范围或者枚举范围、是否必填项未填等.

3 数据交换

建立主数据管理平台的最终目标是以多源异构的数据为依据,建立集中统一的信息集成机制,并将完整一致的高质量数据分发至各部门各单位.数据交换接口通过ESB总线进行集成,采用WebService方式,通过SOAP数据传输协议,采取以对象类型为主的方式进行数据传输.主数据管理平台与其他业务系统信息集成的方式分为单点登录与统一门户两种.

(1) 单点登录.

用户登录其他企业业务系统门户后,可通过点击系统链接,实现自动登录主数据管理系统,同时企业门户系统可对登录的用户做相关权限及功能限制等.

(2) 统一门户.

主数据管理系统作为用户数据的源头,通过门户跳转系统用户管理模块,实现用户数据的新增、变更等操作,操作完成后,系统可将用户相关信息推送到其他业务系统门户,实现门户用户的统一管理.

4 实验测试

平台搭建完成之后,需要检测主数据平台是否如预期实现信息集成,并且与其他业务系统成功进行数据交换.测试主要采用LoadRunner工具进行实时性能测试监控.根据测试需求,分别测试数据在并发单点登录或统一门户用户数为1、5和10时的响应时间是否在可接受范围内.不同并发压力下,新增数据事务的平均响应时间、90%响应时间和标准偏差如表3所示.

表 3 并发用户测试性能表

Tab.3 The test performance table of concurrent user					
数据类型	并发用户数	平均响应时间/s	标准偏差	90%响应时间/s	事务通过率/%
统一门户	1	0.976	0.02	0.99	100
	5	1.022	0.083	1.181	100
	10	1.122	0.267	1.443	100
单点登录	1	0.408	0.007	0.416	100
	5	0.819	0.175	1.052	100
	10	1.323	0.364	1.585	100

统一门户及单点登录功能标准偏差对比如图3所示.

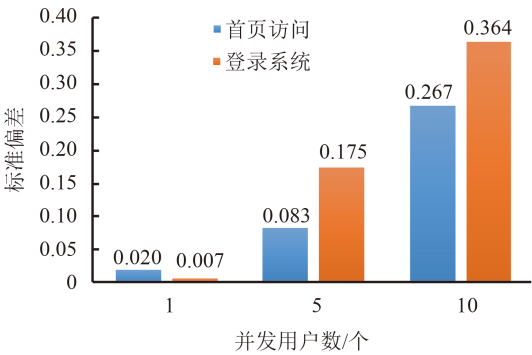


图 3 统一门户及单点登录功能标准偏差对比图
Fig.3 Standard deviation comparison chart for unified portal and single sign-on functions

在测试过程中,通过对数据性能表现进行分析,得出被测系统在要求的并发用户数下性能表现良好,平均事务响应时间符合用户体验时间模型,服务器运行稳定,主数据管理平台成功与其他业务系统数据交换.

5 结语

主数据管理平台针对政府各部门企业信息冗余不一致的现象,以企业信息情况的特定需求为准则规范,根据主数据关键性、独一性以及长期有效性的特征,构建集中统一的数据管理模型,有效地将多源

异构的企业信息集成,并分发提供给其他业务部门使用。主数据管理平台成为了一个良好有效的信息集成机制,并增强了整个数据管理体系应用的数据同一性和信息准确性。随着主数据管理平台的成功上线,成果已实际应用到各个政务部门。接下来将着手完善数据同步与更新问题,优化用户体验,以保证主数据管理平台的持续性和高效性。

参考文献:

- [1] 贺益盛. 基于主数据的信息系统集成与应用研究[D]. 广州:华南理工大学软件学院, 2014.
- [2] 郭延全. 异构数据库环境下数据集成技术的研究[D]. 大连:大连海事大学信息科学技术学院, 2007.
- [3] 陈跃国, 王京春. 数据集成综述[J]. 计算机科学, 2004, 31(5): 48-51.
CHEN Y G, WANG J C. A review of data integration [J]. Computer Science, 2004, 31(5): 48-51.
- [4] 黄锦辉. 福建烟草数据中心主数据管理系统的设计与实现[D]. 厦门:厦门大学软件学院, 2013.
- [5] 武彤, 谭光炜. 基于索引视图实现动态数据仓库的实时数据加载[J]. 计算机科学, 2016, 43(6A): 493-496.
WU T, TAN G W. Real-time data loading of dynamic data warehouse using index view set [J]. Computer Science, 2016, 43(6A): 493-496.
- [6] 钱鹏程. 基于主数据管理技术的企业信息集成方法研究[D]. 上海:上海交通大学计算机科学与工程系, 2009.
- [7] 翁盛鑫, 庄严, 陈祁. 基于数据整合应用的预约挂号平台设计与实现[J]. 中国数字医学, 2012, 7(3): 53-56.
WENG S X, ZHUANG Y, CHEN Q. Design and implementation of appointment registration platform based-on data integration [J]. China Digital Medicine, 2012, 7(3): 53-56.
- [8] 郑龙. 企业数据管理系统的研究和实现[D]. 上海:上海交通大学计算机科学与工程系, 2008.
- [9] 于玺, 张义强, 赵彦宾, 等. 基于模式匹配的主数据质量控制[J]. 计算机应用, 2013, 33(A01): 73-75.
YU X, ZHANG Y Q, ZHAO Y B. Quality control of master data based on pattern matching [J]. Journal of Computer Application, 2013, 33(A01): 73-75.
- [10] 朱默. 基于开源平台的高校主数据管理平台分析与设计[D]. 南京:东南大学计算机科学与工程学院, 2015.
- [11] 王欢. 中海油主数据模型和主数据管理系统的设计与实现[D]. 哈尔滨:哈尔滨工业大学软件学院, 2015.
- [12] 许青林, 覃国民, 姜文超, 等. 敏感数据自主可控的云存储平台元数据管理[J]. 广东工业大学学报, 2014, 31(4): 46-53.
XU Q L, QIN G M, JIANG W C. Metadata management of cloud storage platform based on sensitive information self-Management [J]. Journal of Guangdong University of Technology, 2014, 31(4): 46-53.
- [13] 李桥兴, 强保华, 杨春燕. 大数据基元的HBase数据库存储模型与实现[J]. 广东工业大学学报, 2014, 31(3): 8-13.
LI Q X, QIANG B H, YANG C Y. The storage model of big data basic-elements in HBase database and its realization [J]. Journal of Guangdong University of Technology, 2014, 31(3): 8-13.
- [14] 束进, 牛渝, 周巍伟. 企业信息化建设中的主数据管理[J]. 上海船舶运输科学研究所学报, 2016(1): 81-84.
SHU J, NIU Y, ZHOU W W. The master data management in enterprise informatization [J]. Journal of Shanghai Scientific Research Institute of Shipping, 2016(1): 81-84.
- [15] XIAOCHEN D, XUE H. Multi-decision-tree classifier in Master Data Management System [C]// Proceedings of 2011 International Conference on Business Management and Electronic Information(BMEI 2011). Guangzhou, China: IEEE Beijing Section, IEEE Wuhan Section, Guangdong University of Business Studies, Engineering Information Institute, 2011, 3: 756-759.
- [16] LAMOLLE M, MENET L, LE D C. Incremental checking of master data management model based on contextual graphs [J]. Enterprise Information Systems, 2015, 9(7): 681-708.
- [17] 李刚, 焦谱, 文福拴, 等. 基于偏序约简的智能电网大数据预处理方法[J]. 电力系统自动化, 2016, 40(7): 98-106.
LI G, JIAO P, WEN F S, *et al*. A partial order reduction based method for big data preprocessing in smart grid environment [J]. Automation of Electric Power Systems, 2016, 40(7): 98-106.