

统计数据与大数据源的融合应用*

周静¹, 余浩然¹, 米清奎¹, 黄忠¹, 龚伟²

(1. 重庆市统计局 信息化建设和管理处, 重庆 401147;

2. 中共重庆市委网络安全和信息化委员会办公室 网络数据与技术处, 重庆 401121)

摘要:【目的】通过统计数据和大数据融合来扩大统计基础数据库数据资源的广度和深度。【方法】首先,将数据库分成多源异构资源装载的数据湖、大数据资源管理、为多元数据服务的大数据应用3部分,将多元异构原始数据导出成csv数据文件,根据文件中每列的内容进入元数据编辑工具录入DataFiles对象信息,使DataFiles中的variable内容和顺序与csv文件的列的内容和顺序对应,达到使用DDI元数据描述统计数据的作用。其次,将数据描述信息导出生成DDI元数据文件,将两两配对的csv数据文件和DDI元数据文件一起提交给数据平台,经过解析识别完成数据入库。再次,用大数据开发方法,引入HBase和Hive技术,结合通用统计业务模型(Generic statistical business process model, GSBPM),将数据库中的数据利用ETL工具进行重新组合或融合形成资源库和应用库,并根据不同的业务场景要求,利用MPP数据库技术工具,满足在线海量数据的实时计算以及离线数据批量计算。最后,根据统计分析需求,利用地理信息系统GIS结合指标体系形成直观、立体、多元化的数据展现及分析场景,形成“统计一张图”。【结果】完成了统计数据和大数据源融合的统计大数据平台架构。【结论】建设统计大数据平台就是建设行政区域内统一的经济、社会和人口宏观,微观数据中台,实现对经济运行数据、社会管理数据和人口分布数据的智能化关联匹配,并能全景式一体化呈现全市经济社会发展状况,具有较强的实用价值和社会效益。创新点为建设统一的数据融合标准体系,并为即将建设的重庆市统计大数据平台提供了理论研究基础。

关键词:统计数据;大数据源;融合;应用研究

中图分类号:TP315

文献标志码:A

文章编号:1672-6693(2021)02-0077-07

随着计算机技术、互联网应用的深度发展,越来越多来源广泛、运算快速、信息海量的数据汇集形成了大数据,催生了一个崭新的技术新时代^[1]。数据通常可分为传统数据和非传统数据。传统数据的官方发布主要指省(直辖市、自治区)统计局的综合数据资源。非传统数据根据国家统计局2018年制定《非传统数据统计应用指导意见》则定义为包括大数据但不限于大数据。大数据是非传统数据的主体,在很多情形下可以代指非传统数据^[2]。非传统数据是政府统计基础数据的重要来源,它拓展了统计范围、丰富了统计内容、提高了统计效能。特别是近些年来,大数据在经济社会管理中发挥着越来越重要的作用,大数据技术给传统统计数据采集、管理、分析、应用、发布等各方面工作带来前所未有的强有力冲击^[3-4]。政府部门应该思考如何克服各种数据资源差异性,让统计数据和大数据融合发展,建立以标准化数据库为核心的、强大的统计数据库信息智库系统,推进数据资源整合和开放共享^[5]。

1 大数据与政府工作融合发展的现状

1.1 大数据与政府工作融合的背景

大数据是信息化发展的新阶段。随着信息技术和人类生产生活的交汇融合、互联网的快速普及,全球数据已呈现爆发增长、海量集聚的特点,这对全球经济发展、社会治理、国家管理、人民生活都产生了重大影响。多个国家先后制订了智慧政府建设规划,以信息技术为支撑,充分开发利用各种信息资源,逐渐将居民和商业部门的大数据实践经验应用于政府部门,进一步推动和促进公共管理、公共服务的方式创新^[3-4]。例如,日本在2016年

* 收稿日期:2021-01-04 修回日期:2021-03-04 网络出版时间:2021-04-21 11:41

资助项目:全国统计科学研究重点项目(No. 2019LZ24)

第一作者简介:周静,女,高级经济师、高级调查分析师,研究方向为计算机信息网络管理和大数据应用,E-mail:645132067@qq.com

网络出版地址:https://kns.cnki.net/kcms/detail/50.1165.N.20210421.0938.002.html

成立了大数据专家组,把大数据政府应用作为日本政府面向未来的关键使命。荷兰统计局于 2018 年开始全面进行服装零售商网站信息抓取服装类价格数据,并把结果纳入到荷兰 CPI 统计之中,全面替代传统的商店调查方式。中国近些年来也加强了大数据与政府工作的融合开发力度,例如:用大数据预测雾霾;杭州市政府联合阿里云计划建立杭州城市数据大脑,用大数据助力杭州“治堵”;重庆市政府的重庆政务掌上办“渝快办”,逐步实现交通、医疗、卫生、就业、社保、文化、教育、科技等数据集开放,带动大数据增殖性、公益性开发和创新应用,充分释放数据红利。

1.2 大数据与政府工作融合存在的问题

通过前期调研发现,大数据虽然能大大提升政府部门统计工作的信息化技术水平,但是要实现大数据和政府部门统计数据的融合还存在一些难点。

1.2.1 元数据标准不统一 由于目前缺乏对数据交换标准和管理标准的体系研究,导致多源异构数据整合方法的设计成为大数据与政府工作融合的最大难点。在以往数据库系统建设中,不同组织、不同系统建立了种类繁多的元数据体系,统计调查元数据的术语不统一,建模方法不统一,导致在数据生产的各个流程中元数据可比性差。因此,如何建立一套科学有效,稳定易用,便于交换的元数据标准是目前数据融合亟待解决的最大问题。

1.2.2 大数据源里存在一些“脏数据”、“废数据” 大数据里的数海包罗万象,数据的时间空间跨度大,传输期间可能被多次转摘、甚至篡改等,在统计数据与大数据融合过程中,若在入库前未对大数据来源把好关,则会出现一些不完整、不合格的“脏数据”“废数据”,或者出现大量重复的数据信息等,就会造成数据融合中筛选识别困难,甚至数据采集缺项等后果,严重降低数据价值^[6]。所以,如何有效甄别入库前大数据源里的有效数据也是统计数据与大数据源融合所面临的问题之一。

1.2.3 数据入库后的管理难题 在统一标准的情况下,数据入库后如何选用适合的大数据方法应用在统计数据与大数据融合过程中,以及融合后的海量统计大数据的管理和可视化应用也是数据融合应用的一个难点。

1.2.4 数据隐私存在安全隐患 数据共享和数据隐私存在一定矛盾,数据共享开放的需求越迫切,数据隐私安全问题就越突出。为了从不同角度全方位观察、认知事物,最好的途径就是对海量、高质量数据资源进行分析和挖掘,而共享开放和数据跨域流通为信息建立了完整数据集。可是,如果大数据是无序流通与共享,就可能存在隐私保护和数据安全方面的重大风险。

2 大数据与政府工作融合的设想

为解决上述大数据与政府工作融合中存在的问题,本研究在对北京、上海等地统计数据资源体系建设的调研基础上,参考 DDI 和 SDMX 等先进元数据标准,对统计微观数据库和宏观数据库建设方面进行了探索,主要从以下几个方面入手。

2.1 建立统计数据与大数据统一的标准体系

为打破传统数据库无法包容多源数据的瓶颈,本文将数据库分成多源异构资源装载的数据湖、大数据资源管理及大数据应用 3 部分,以完成多源、多结构数据采集交换平台的设计。具体来说,就是按照通用统计业务模型(Generic statistical business process, model, GSBPM),以国家统计局调查元数据标准为基础,使用 DDI+SDMX 标准,在描述统计需求确认、设计、开发及任务部署、采集、审核上报和数据分析及汇总过程的同时,加大描述统计数据分析及汇总、数据发布和数据归档过程,最后使用 DDI 标准来描述规范统计数据存档评估,利用 SDMX 标准指导大数据源数据识别,一一对应将统计数据与大数据源数据融合,用 DDI+SDMX 标准最终实现统计数据 and 大数据源数据识别和交换标准的统一(图 1)。具体实施步骤如下。

第 1 步,按 GSBPM,遵循统计业务流程设计、开发、采集、处理、分析、存档数据后,将多元异构原始数据导出成 csv 数据文件,根据文件中每列内容进入元数据编辑工具录入 DataFiles 对象信息,使 DataFiles 中 variable 内容和顺序与 csv 文件的列内容和顺序对应,达到使用 DDI 元数据描述统计数据的作用。

第 2 步,将数据描述信息导出生成 DDI 元数据文件,将两两配对的 csv 数据文件和 DDI 元数据文件一起提交给数据库,按照 SDMX 标准解析识别完成数据入库。以后增加的来自外部组织的数据,只要符合 DDI 标准或 SDMX 标准也可以直接被解析识别,并将信息存储到数据仓库中。数据库接收到 csv 数据文件和 DDI 元数据文件,将其中的数据 and 元数据解析并存储。例如,图 2 为国内生产总值分行业增加值的综合报表。该报表用 SDMX 标准识别了其中各元数据对象,拆分出来的元数据对象指标名、计量单位、目录、报告期等就可以作为生成该报表的数据描述解析的基本内容。

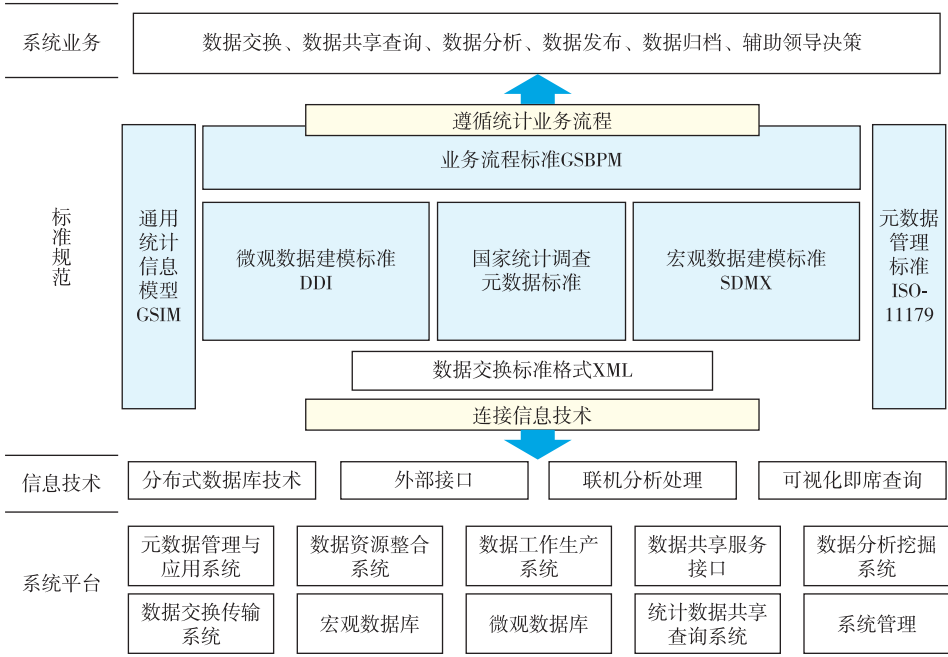


图 1 数据融合标准体系模型图
Fig. 1 Model diagram of data fusion standard system

		指标名						
		3-8 分行业增加值						
计量单位		Value-added by Sector						
单位: 亿元				(100 million yuan)				
目 录		行 业		2013	2014	2015	2016	报告期
指标名		地区生产总值		3 177.56	3 500.7	3 702.8	4 044.5	变量值
		农、林、牧、渔业		756.30	832.64	880.52	977.63	
		工业		472.38	514.4	485.85	479.22	
目 录 项		建筑业		326.11	362.77	390.41	424.5	
		批发和零售业		361.10	418.09	440.75	467.8	
		交通运输、仓储和邮政业		148.35	185.15	187.8	199.79	
		食宿和餐饮业		139.03	154.66	174.14	191.96	
		信息传输、软件和信息技术服务业		60.92	65.24	79.28	96.23	
		金融业		187.14	210.63	242.82	281.9	
		房地产业		301.69	295.56	299.69	249.95	
		租赁和商务服务业		34.82	36.42	43.75	49.77	
		科学研究和技术服务业		29.70	32.96	38.13	44.29	
		水利、环境和公共设施管理业		26.57	29.91	35.41	41.5	
		居民服务、修理和其他服务业		38.49	42.17	45.27	26.53	
		教育		105.80	114.22	125.54	131.39	
		卫生和社会工作		47.44	53.16	61.14	72.46	
		文化、体育和娱乐业		23.80	25.39	29.17	34.57	
		公共管理、社会保障和社会组织行业		117.92	127.35	143.09	145.02	
分 组		第一产业		736.03	809.52	854.72	948.35	
		第二产业		797.39	875.97	875.82	902.68	
		第三产业		1 644.14	1 815.2	1 972.2	2 193.5	
表 注		注: 本表按当年价格计算。						
		Note: Total data this table are calculated atcurrent prices.						

图 2 SDMX 描述综合报表实例
Fig. 2 Example of SDMX description comprehensive report

第 3 步,使用通用统计信息模型(Generic statistical information model, GSIM)打通 DDI 和 SDMX 两类标准,使信息之间可以进行数据的流转、形态的变化;然后再使用 ISO-11179 元数据管理标准进行两类元数据的管理。

第 4 步,通过 DDI+SDMX 把入库后的 csv、excel 等数据统一为 XML 载体,因为 XML 具有良好的易读性、可扩展性及平台无关性,还有标签注解等功能,特别适合做数据研究载体^[7]。这些原始数据再按照需要形成 Cube 用于形成主题数据集市。

通过建立这套完整的标准规范,向上可对接系统业务各个阶段,如数据交换、数据共享查询、数据分析、数据发布、数据归档和辅助决策等;向下可对接信息技术和系统平台,成为数据融合技术之间的传输纽带和翻译器,实现了统计数据与大数据标准的统一。

2.2 质量优先、数据量并重,科学把握数据源头

为了更有效清理大数据源里的“脏、废”数据,首先要按照“质量优先、数据量并重”的原则,审慎面对海量数据。第二,进一步完善现代统计调查体系,在强化科技支撑、推进大数据等技术在统计中应用的同时,研究大数据校正机制和规范方案,加快建立大数据安全评估制度。第三,结合统计工作方法,科学分析使用大数据,充分应用抽样调查分析和重点调查分析等方法来拟合验证数据来源的真实性,取精华、去糟粕,严把数据源质量关,提高大数据安全管理能力。第四,加强算法应用前的安全评估,确保算法执行效果安全可控。

2.3 建设政府统计数据与大数据融合智库平台

虽然建立大数据与政府部门统计数据的标准体系解决了多元异构数据融合的难题,但融合后数据的存储和管理则需要通过搭建智库完成。为了更快地完成大量数据的离线计算、实时查询分析、高时效性大规模并行计算的场景,研究组采用大数据架构建立总数据库,里面包括贴源层数据库、标准层数据库、资源库和大数据应用库,用于整合、存储统计数据和大数据源数据。

首先,按照 DDI+SDMX 标准将多源异构数据统一标准后,通过 ETL 工具或流式采集技术将政府部门数据和大数据等多源异构数据整合到贴源层数据库中,再利用 SDMX 和 DDI 中的元数据标准,统一数据库表结构命令规范,形成全量数据字典目录,将贴源层数据进行标准化,再将标准化后的数据存储于标准层数据库中,完成政府部门统计数据和大数据融合后的数据入库。

其次,按照 GSBPM,利用 ETL 工具 Kettle 将标准层数据库中的数据进行重新组合,融合形成资源库。资源库分为基础库、专题库和主题库,资源库服务于现有业务系统。然后,根据数据应用要求,利用 ETL 工具 Kettle 将标准层数据库或资源库中的数据进行重新组合,融合形成应用库。应用库服务于决策支持、应用展示以及信息资源服务。资源库和应用库可利用 MPP 数据库 Clickhouse、Hive 技术满足在线数据实时计算以及离线数据批量计算等不同业务场景要求,通过 HBase+Hive+Clickhouse 方式,其中:HBase 负责对制度采集、网络获取或部门交换数据进行原始数据的存储和主题库资源的存储;Hive 负责将需要进行运算的数据从 HBase 中抽取后根据模型进行离线运算;Clickhouse 将需要进行查询分析展现的数据进行抽取与存储,为数据分析与决策提供支撑,最后成功构建政府部门统计数据和大数据融合统一的大数据架构,结构如图 3 所示。

2.4 平台分级管理,加强数据安全保护

为了加强数据隐私保护,必须体系化、一致性全面考虑信息网络安全管理工作^[8],除了严格遵守国家关于数据安全、个人信息保护相关的法律法规,还应加强隐私保护、数据安全和数据流动利用的新技术研发。本研究所设计的系统在数据融合过程中,注意运用数据脱敏和数据发布匿名保护技术,并在平台应用中进行分级管理,把融合数据的应用汇集,然后对识别认证后的访问者身份进行分级设置管理,应用权限分高级、中级和普通级,为不同级别的用户开放不同级别的数据,既保证了数据共享,又兼顾数据安全。

3 案例验证

2017 年本研究组对重庆 28 个市级部门进行统计工作调研,重点了解各部门现行统计报表制度、数据平台使用情况以及对共建数据采集管理平台需求,为大数据和统计数据融合、构建统计大数据平台奠定了基础。最终,本文基于上述统计数据和大数据源融合的思路和调研结果,设计了重庆统计大数据平台(图 4)。主要包括数据生产、数据管理和数据应用 3 部分。

数据生产主要包括数据采集、存储过程,汇聚重庆市统计局综合数据资源(包括一套表、各种普查、电子台帐、部门综合数据资源、统计局监测体系数据资源和其他各种临时零散统计数据等)、国内、国际统计数据资源和

互联网大数据资源等。数据管理主要把汇集的多元异构数据标准化并融合管理,形成大数据资源体系,并分类管理形成若干专题库。数据应用采用分层管理方式,即超级应用、行业应用和通用应用共3级,超级应用是最高级别,有最大应用权限;行业应用其次,主要为统计专业人士服务;通用应用主要面向大众,并根据各层次特别分门别类设置栏目,有条件地共享开放数据。

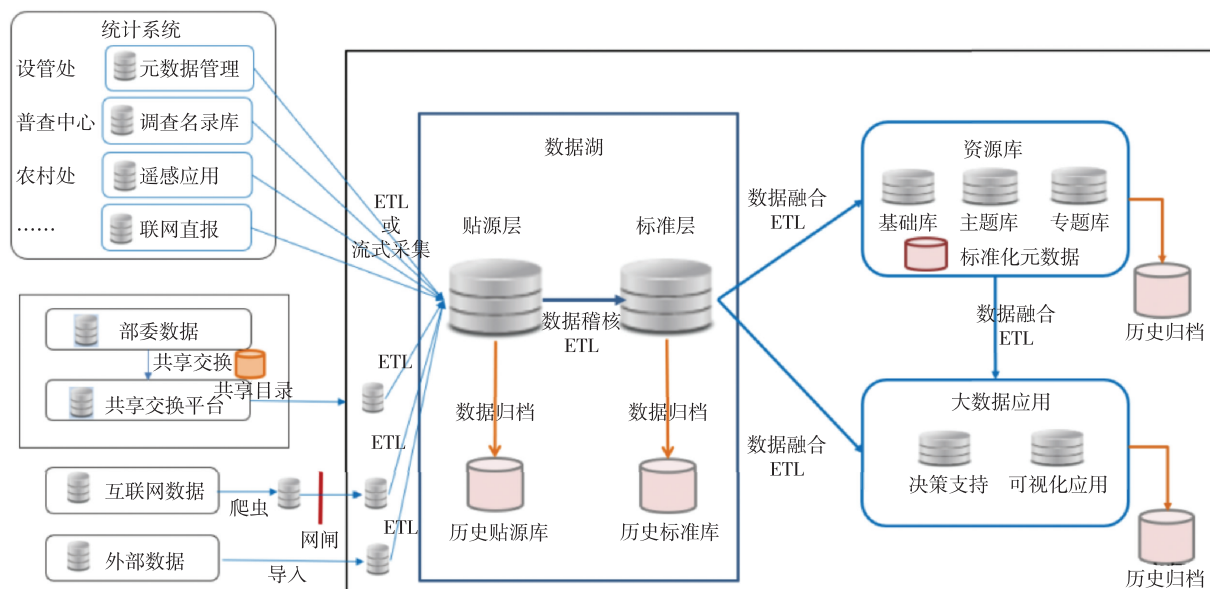


图3 数据融合模拟图

Fig. 3 Data fusion simulation chart

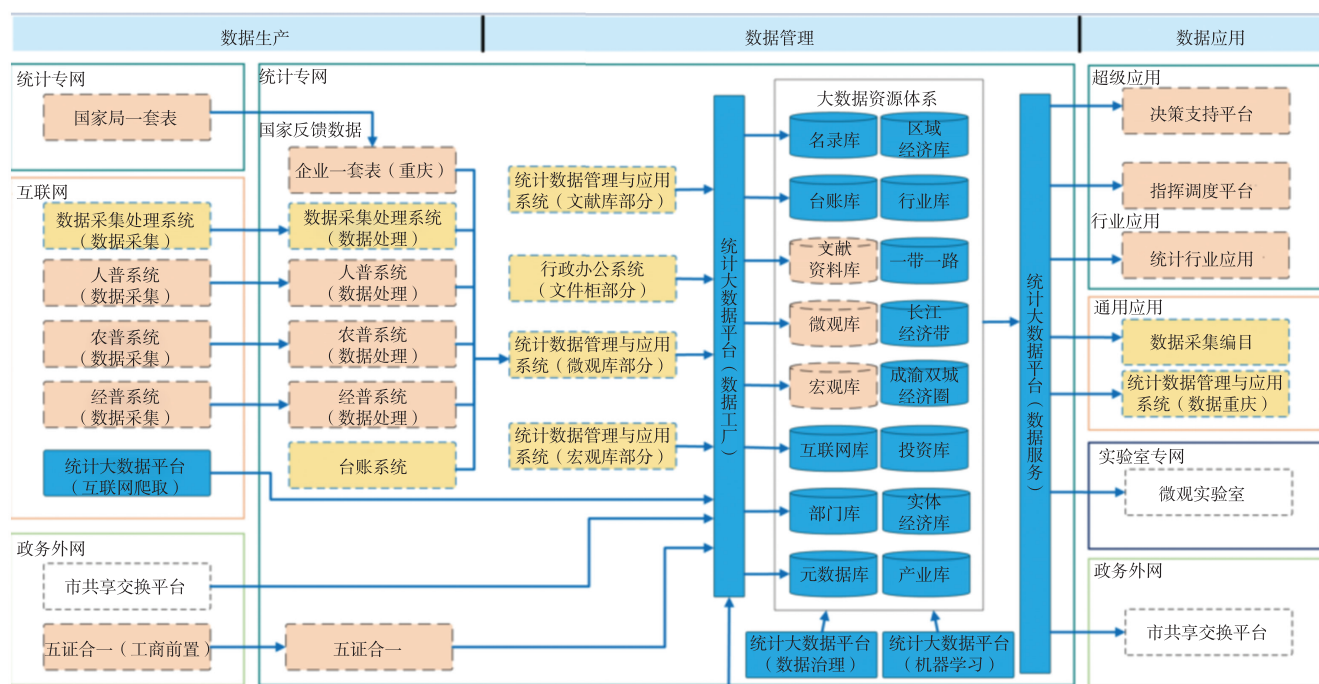


图4 统计大数据平台模拟图

Fig. 4 Simulation chart of statistical big data platform

图5是统计大数据平台应用分级模型,把融合数据的应用汇集,然后对识别认证后的访问者身份进行分级设置管理,应用权限分高级、中级和普通级,即超级应用、行业应用和通用应用,为不同级别的用户开放不同级别的数据,既保证了数据共享,又兼顾数据安全。

统计大数据平台建设涵盖了统计数据与大数据的采集、元数据的标准化、数据存储融合以及最终形成统一的可视化成果的完整过程,即通过一套表、各种普查、部门综合数据资源、五证合一、国内、国际统计数据资源和

互联网大数据资源等信息同步,完成对统计数据实时数据采集,再依据数据标准化系统,通过采用同样的标准以及描述工具统一元数据,对数据进行清洗、转换和存储,实现对数据的分析成果。此外,利用大数据等先进技术对平台数据的采集、存储、融合、管理和应用等进行全流程监控,建立统计“一张图”可视化展示,确保统计数据生成各个环节有标准、留痕迹。建设统计大数据平台其实就是建设全市统一的经济、社会和人口宏观、微观数据中台,实现对经济运行数据、社会管理数据和人口分布数据的智能化关联匹配,以全景式一体化呈现全市经济社会发展状况。

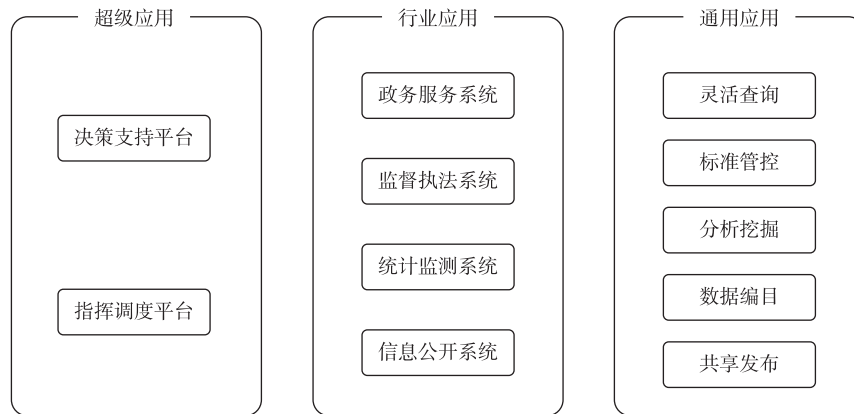


图 5 统计大数据平台应用分级图例

Fig. 5 Application classification legend of statistical big data platform

4 结束语

本研究在理论上探讨了实现大数据和政府统计数据融合起来的可能性。通过融合统计局、政府其他部门与宏观经济及社会发展有关的数据、互联网舆情数据、运营商相关数据等,研究开发大数据处理新工具,根据每个数据分析项目量体定制,以产生决策智能为目标,将多种数据源中相关数据提取、融合、梳理,整合成一个分析数据集,形成统一的标准化数据资源整合,创新跨部门、跨行业、跨类型的数据共享和利用模式,不断扩大基础数据库数据资源的广度和深度,逐步探索数据共享开放,实现有效的资源共享^[8]。初步设计了统计大数据平台为数据融合案例,涵盖了统计数据与大数据的采集、元数据的标准化、数据存储融合以及最终形成统一的可视化成果的完整过程,模拟验证了该数据融合研究的可行性。

下一步将以此研究为基础,把项目研究理论成果应用到政府统计部门工作实际中,重庆市政府统计部门准备用 3 年时间搭建重庆统计大数据平台,目前正在编制统计大数据平台初步设计及概算方案。平台坚持以“创新、融合、安全、高效”为建设原则,在汇集重庆市“四上”企业行业统计一套表、经济、农业和人口大型普查调查原始数据等微观数据库和综合汇总数据库基础上,融合政务外网、电子台帐、五证合一、互联网大数据等经济活动信息,设置文献资料库、名录库、台帐库、互联网库、区域经济库、成渝双城经济圈等专题应用库,并保证应用库可以按需实时增减修改完善,建成统计数据深度挖掘和智能分析系统,开展政府、社会和互联网大数据的智能化比对,实现宏观和微观数据的相互印证,最终形成重庆全市统一的经济、社会和人口宏微观数据中台,并通过一体化展示、移动 APP 发布等渠道,分层级向各级党政领导、政府部门、研究机构和社会公众提供“智能、精准、安全”的资源共享服务。为了进一步提升统计服务智能化水平,平台还设计了统计大数据应用研究实验室,计划借力现代信息技术,不断提升大数据在统计工作中的开发应用能力,探索区块链、物联网、新一代通信等高新技术在统计工作中的应用,推动统计工作乃至决策服务的智能化发展。

参考文献:

- [1] 徐继华,冯启娜,陈贞汝.智慧政府:大数据治国时代的来临[M].北京:中信出版社,2014.
XU J H, FENG Q N, CHEN Z R. Smart government: the advent of the era of big data governance[M]. Beijing: CITIC Press, 2014.
- [2] 余芳东.大数据在政府统计中的应用瓶颈及融合路径[J].调研世界,2018(11):3-9.
YU F D. The application bottleneck and fusion path of big data in government statistics[J]. Research World, 2018(11): 3-9.

- [3] 王宇灿,李一飞,袁勤俭. 国际大数据研究热点及前沿演化可视化分析[J]. 工程研究,2014(9):282-293.
WANG Y C, LI Y F, YUAN Q J. Visual analysis of international big data research hotspots and frontier evolution[J]. Engineering Research, 2014(9):282-293.
- [4] 李丹丹. 大数据在政府治理中的应用研究[J]. 科技创新与应用,2021(5):162-166.
LI D D. Research on the application of big data in government governance[J]. Science and Technology Innovation and Application 2021(5):162-166.
- [5] 杜亚彬,张芳丽,王琪. 大数据与政府统计改革[J]. 山东农业工程学院学报,2020,37(10):67-70.
DU Y B, ZHANG F L, WANG Q. Big data and government statistics reform[J]. Journal of Shandong Institute of Agricultural Engineering, 2020, 37(10):67-70.
- [6] 张雯婷,徐映梅. 大数据下政府统计调查升级[J]. 合作经济与科技,2020(16):168-171.
ZHANG W T, XU Y M. Upgrade of government statistical surveys under big data[J]. Cooperative Economy and Technology, 2020(16):168-171.
- [7] 田华明,罗紫倩,乔晨曦. 基于XML的网络制造数据交换研究与分析[J]. 大众标准化,2020(24):190-191.
TIAN H M, LUO Z Q, QIAO C X. Research and analysis of XML-based network manufacturing data exchange[J]. Public Standardization, 2020(24):190-191.
- [8] 罗丽琳,蒲清平,黄燕. 大数据提升网络主流意识形态引领力研究[J/OL]. (2020-09-09)[2021-03-04]. 重庆大学学报(社会科学版). <http://kns.cnki.net/kcms/detail/50.1023.c.20200909.1226.004.html>.
LUO L L, PU Q P, HUANG Y. Big data boosts mainstream ideology leadership research[J/OL]. (2020-09-09)[2021-03-04]. Journal of Chongqing University (Social Science). <http://kns.cnki.net/kcms/detail/50.1023.c.20200909.1226.004.html>.

Research on the Fusion Application of Statistical Data and Big Data Sources

ZHOU Jing¹, YU Haoran¹, MI Qingkui¹, HUANG Zhong¹, GONG Wei²

(1. Office of Administrative and Information Construction, Chongqing Bureau of Statistics, Chongqing 401147;

2. Department of Technology and Network Data, Chongqing Municipal Network Security and Informatization Committee Office, Chongqing 401121, China)

Abstract: [Purposes] To integrate statistical data and big data, thus expand the breadth and depth of data resources in the basic statistical database. [Methods] Divide the database into three parts: data lake loaded with multi-source heterogeneous resources, big data resource management, and big data application serving multi-data. Export the multi-heterogeneous raw data into a csv data file. The content enters the metadata editing tool to enter the DataFiles object information, so that the variable content and sequence in DataFiles correspond to the content and sequence of the columns of the csv file, achieving the effect of using DDI metadata to describe statistical data. Then export the data description information to generate a DDI metadata file, submit the paired csv data file and DDI metadata file to the data platform together, and complete the data storage after analysis and identification. Next use the big data development method, introduce HBase and Hive technology, combine the general statistical business model GSBPM, use ETL tools to recombine or merge the data in the database to form a resource library and application library, and also use MPP Database technology tools according to the requirements of different business scenarios to meet the real-time calculation of online massive data and offline data batch calculation. Finally, according to the needs of statistical analysis, use the geographic information system GIS combined with the index system to form an intuitive, three-dimensional, and diversified data display and analysis scenario to form a statistical map. [Findings] Complete the statistical big data platform architecture integrating statistical data and big data sources. [Conclusions] The construction of a statistical big data platform is to build a unified economic, social and population macro and micro data center in the city, to realize the intelligent association and matching of economic operation data, social management data and population distribution data, and to present it in a panoramic and integrated manner. The economic and social development of the city has strong practical value and social benefits. The foundation and the biggest difficulty of building a platform is to build a unified data fusion standard system, which is also a major innovation of this article. At the same time, new big data tools are used in research to produce decision-making intelligence as the goal, tailor-made data analysis projects, and extract, merge, comb and integrate relevant data from multiple data sources into analysis data sets, which provides a theoretical basis for the “landing” of the Chongqing statistics big data platform to be built.

Keywords: statistical data; big data sources; fusion; applied research

(责任编辑 黄 颖)