文章编号：1007－5399（2017）04－0009－04

邮政省级数据平台系统规划与建设

唐　华，何太能，傅国庆

（中国邮政集团公司四川省信息技术局，四川　成都　610012）

摘　要：文章以四川邮政为例，介绍了邮政省级数据平台系统的建设背景和系统规划，阐述了省级数据平台系统建设中关键节点的实施方案，提出了省级数据平台系统的建设目标，助力邮政数据生产力和数据传播力的提升。

关键词：数据；平台；市州分库；自有库；报表；算法；安全

中图分类号：F61　　**文献标识码**：A

按照中国邮政集团公司的安排，省级邮政部署有金融数据下载平台和邮务类数据下载平台。下载平台存放了邮政储蓄逻辑集中系统等多个生产系统的交易数据和客户数据，为省级数据分析人员面向基层经营部门寻找客户，面向风险管理部门暴露风险，面向管理决策者优化作业流程、探索业务趋势及相关关系提供了直接的数据来源。这些数据至关重要，随着数据的进一步应用，出现了一些局限性，亟需建设完善的邮政省级数据平台，进一步提升数据生产力和数据传播力。

1　省级数据平台的建设背景

四川邮政省级数据员工通常的工作流程是：融入业务、与省级业务人员交流数据需求，业务人员提出数据申请，数据分析人员探讨需求是否有数据支持，对下载平台的数据进行整合、提炼，得到一定的统计结论或者筛选出必要的明细数据，交给省级业务部门下发使用。

若数据在业务应用中产生了明显的效益，则要总结经验，形成案例，联合业务部门形成数据分析课题，把课题成果形成论文。这些案例和论文通过省数据中心主办的纸质数据运营期刊进行发表和传播。

在这种工作模式下，省数据中心人员工作充实饱满、数据分析论文也写得有声有色。但数据的来源单一，数据的加工人手少、力量薄、数据业技交流深度和广度受限，数据传播时限长、范围窄，难以实时向网点传播数据。依赖于金融个人客户营销系统时，归属银行的客户在邮政代理网点的行为资料无法下发到代理金融网点、资产万元以下潜在客户无法通过金融个人营销系统下发。依赖于办公网下发时，数据文件需按下级机构清分后分别发送，手续繁琐。

2　省级数据平台的系统规划

邮政省级数据平台系统规划，即基于省级已有的数据资源建立数据平台——扩大数据生产（创立市州分库、配套数据申请流程），拓展数据来源（划出名址自有库区域），固化分析成果（给出报表工厂），开辟传播渠道（创建数据分析论文应用案例及明细数据的双向传输通道）。实践中邮政省级数据平台可采用以下系统结构，如图 1 所示。

2.1　基础平台夯实

存储规划为 40TB，促成省公司提出的“六化”目标，即数据源头全面化、客户身份唯一化、数据申请流程化、数据提取自动化、数据分析智能化、数据查询可视化。

下载平台数据及时自动流向数据平台、数据仓库，及时自动清分到各市州分库让市州数据队伍有数据可供分析。建成报表工厂并形成工厂运维机制，建成统计及明细报表分层级、分专业传送通道。实现案例自助上传及授权下载。提供自有库创建、数据导入、数据导出工具。系统对复杂需求的数据申请实现流程化。

2.2　数据平台运转

省级层面：2017 年报表课题发展到 100 个，并实现数据自动更新；以后每年增加 30 ～ 50 个报表课题。报表课题面向全省 21 个市州、约 180 个县、2500 个网点和邮政金融、电子商务、寄递、文化传媒等专业分别展示。

市州层面：实现自有库自建、自有库与市州分库数据整合分析。

2.3　数据深度应用

省级层面：提供案例文档在线展示、自有库数据网点采集、目标客户分派、营销活动发起、营销效果指标自动取数度量等功能。

市州层面：尝试发布自建报表，支持网点一线人员营销。

2.4　数据进化增殖

建立省市级数据应用暨采集系统与数据平台仓库双向传送数据的接口。编码逐步完成各应用系统与数据平台间数据的实时自动融合。

在本平台嵌入 R 语言、Python、Hadoop 等工具，尝试

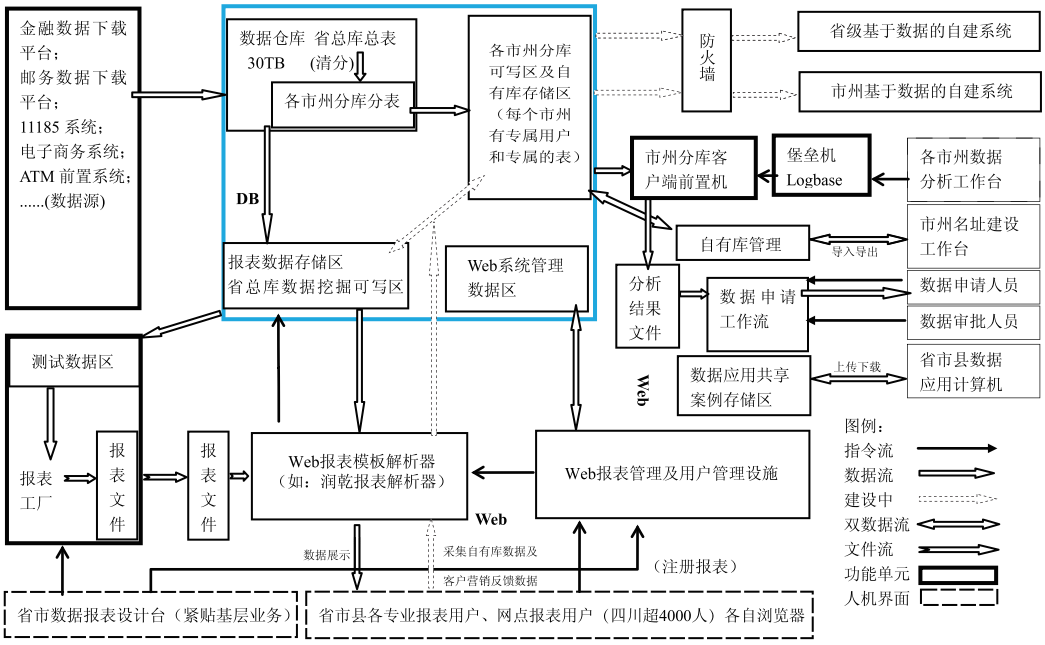

图 1　邮政省级数据平台系统结构图

挖掘数据相关性。

3　省级数据平台的建设实践

3.1　建立市州分库，扩大数据生产

数据从下载平台向省总库的全量传递和增量更新可用数据库链接实现，也可在 shell 引导下用数据泵抽取。为了使用便捷，依据下载平台数据表的实体逻辑，在数据平台对数据表建好索引和分区设施至关重要。为了大量数据全量传递更加快捷，可先 unusable/drop 索引，抽取数据完成后，再 rebuild/create 索引。

3.1.1　清分算法

为了把省总库的数据清分给各个市州，让市州紧贴业务自主分析，可按照直接或间接（通过账号开户网点等关系）得到数据的市州机构，执行 insert into ...select...from...where inst=’CityNo’即可。按照这种办法，如果某省有 21 个市州，这种写法会全表扫描 21 次，完成数据市州清分所需的时间相当长。也可对同一类数据表全省的总表启用游标，每扫描一条记录便将其移至所属市州，继续处理下一条记录。这样，只需扫描一次数据总表即可完成全省各市州的清分，所需理论时间是前一种办法的 1/21。如果在系统 I/O 上仍出现瓶颈，可以为每个市州重建内存数据表，先将市州的数据写入各自内存表，再从内存表一次性插入市州分库的对应表。为节省内存开销，可采取批量更新、分段提交的办法，也可减轻回滚段的压力。此外，对于一次性全部更新的市州数据，清空分库原表时用 truncate table 可以避免出现使用“delete all；commit；”时虚高数据表水位、空耗表空间现象。

3.1.2　权限管控

市州分库的使用者可能无意间发出指令，对下载平台同步的数据进行写操作，在市州技术欠发达地区，这种情况出现的概率较高。这些操作一旦执行，就破坏了数据的原貌，影响了数据的真实性。这就需要在表空间的安排上未雨绸缪，将下载平台的总表清分及市州的数据与总表安排在同一用户下、同一片表空间里，将表名加上市州标识的前缀。这样，总表往市州清分无需额外授权。另外，只为每个市州建立一个用户、一片本市州专属的独立表空间。用逐表 grant 授权 select 的方式，编写脚本批量授权市州只读前缀为自身市州名的那些表。这样即可解决市州只读原始数据和可新建临时表的矛盾，杜绝了市州数据被其他市州交叉访问，触碰数据安全管控规则。此外，还需要把无法清分的公共数据表授权给各个市州的用户 select。

3.1.3　用户环境

要让市州进入市州分库分析数据，除了 CS 模式下的服务端环境，还需要客户端环境。在数据安全可控的前提下，为了把 logbase 堡垒机扩展到市州使用以记录操作日志，需深挖堡垒机潜能。将用户与邮政专网综合网 IP 绑定；预置各市州的 Windows 用户，不允许手动输入其他市州的 Windows 用户，防止用户密码泄露后被其他市州冒用；禁用 RDP 粘贴板和磁盘映射。扩充一台 Windows 2008 前置机，需在远程会话主机配置中将连接的加密方式置为“客户端兼容”，安全层选择 RDP 安全层，取消勾选“仅允许运行使用网络级别身份验证的远程桌面的计算机连接”，去掉“登录设置”中的“始终提示密码”，以此减轻堡垒机自带前置机的负担。在前置机上，为各市州创建用户；限定市州间独立的磁盘空间；限定每个市州 RDP 会话数为 1；设定会话在停止人机交互 10 分钟内自动注销以减少服务器资源的无谓占用；安装 oracle 各市州独立的客户端（如 PL/SQL

developer）环境；PL/SQL 的密码预置在前置机中，市州用户使用时密码自动填充，无需告诉市州分库的密码，减少密码泄漏风险。

3.1.4　数据导出

上文解决了市州分库的内容问题和数据操作安全问题，在可用性方面，还有数据安全下载的问题尚未解决。此类数据在下载前需要经省数据中心审核是否含有敏感数据，根据数据提取代码论证技术路线的正确性。因此，需要引入 Web 工作流，将市州分库数据申请、审核、下载融入工作流中，对数据审批后下发。准备下载环境时，直接让 Web 服务器 ftp 市州分库前置机各市州特定目录。如果市州分库前置机与下载平台部署在金融网段，不能被部署在综合网段的 Web 服务器主动访问，还需要提供工具，让市州分库的分析人员及时将结果文件推送到 Web 服务器。若市州共用一个 ftp 账号，平台的建设者可用 Windows VC++ 等工具录入一个 ftp 的客户端隐藏密码及各市州 ftp 主目录，若用 .bat、资源管理器里面 URL 送入 ftp://user:password@webserverURL 等手段，需在 Web 服务器上为每个市州创建不同的 ftp 账号，指向不同的 ftp 主目录，存放本市州的数据分析结果。

3.2　整合自有数据，丰富数据来源

省、市、县多年的名址建设积累了较多客户资源，只是这些资源大多分散、孤立、封闭地存在。若能与业务生产数据整合，将会产生意想不到的效果。如自有库中高收入会员数据、车主数据反映了个人的收入水平、消费能力，与个人邮政金融资产比较，可以发掘出金融产品的潜在需求。市州分库建成后，生产数据已清分到各个市州的虚拟环境里，自有数据若能传入，就为数据整合提供了可能。

3.2.1　数据上传

流行的 ftp 上传工具，例如 Serv-U，可以实现只允许上传不允许下载，也能为各市州的 ftp 用户指定各自的主目录，这似乎保证了数据可以上传不能随意下载。但是上传数据也不可掉以轻心，市州可能有程序员写一段代码，生成可执行文件上传后破解 oracle 的密码、破坏 Windows 里面的安全设置。在 ftp 服务器里，或者在堡垒机的监控策略中，精准识别上传的文件是普通数据文件还是可执行的代码并非查看文件名的扩展名那么简单，写这样一段代码也并不容易。因此，还需另辟蹊径，为自有库建库创建 WEB upload 工具，其中 Java web struts 2 可以实现上传数据拒绝可执行文件。在数据库层面，预先将各市州自有库和市州分库的可写区设置在同一片数据表空间。

3.2.2　数据整合

在数据库层面已预先将各市州自有库和市州分库的可写区设置在同一片数据表空间及同一个用户下，SQL 中带模式名的表名可访问下载平台同步过来的数据表，不带模式名的表名可访问自有库中的表。接下来则全部依靠 SQL。数据库市州用户太多，若操作者对 SQL 等数据库技能掌握不够熟练，可能会无意中发一些关联匹配大表的操作，将全省共享的数据库拖慢。可以提醒市州自建 oracle 学习环境供相关人员操练。

3.2.3　数据导出

如何导出数据的问题上文已经说明。但是平台设计者或需留意：自有库本身应提供下载功能，但是这个功能可能给市州分库的数据下载提供了一种可能。在 Web 开发中，需要封堵这个漏洞。可采取对自有库上传渠道进入的数据自动生成校验码，市州只能从表中下载自己上传的数据，不包括市州分库平移过来的数据。

3.2.4　额外收获

预先将各市州自有库和市州分库的可写区设置在同一片数据空间、同一个用户下，还有一个好处：技术人员可跟进自有库建设，从底层把握自有库的结构，为技术能手运用原生的 oracle sql、pl/sql 等技术建设维护自有库提供了游刃有余的环境。

3.3　工具量产报表，固化分析成果

报表分系统面临两大任务：一是为数据生产者、省市数据分析人员提供快捷制作报表的工厂；二是为数据消费者、省、市、县网点各专业业务人员分类、分权展示报表数据。

3.3.1　报表工厂

为了实现第一项任务，需采用中国邮政集团公司为电子商务平台采购的润乾报表工具。通过参数传递这一通信方式，实现 Web 程序与报表展现的分离。使用报表工具制作报表就像使用 Excel 一样简便。报表工具封装了数据库操作的细节，新制作的报表文件加入系统可以立即呈现报表，即使 weblogic 处在“生产模式”下，也无需重启 Weblogic 服务。

搭建了报表基础结构 Web、报表应用 Web、数据库环境，布放了设计器，开辟了设计环境与生产环境间报表自动传送的通道。生产报表的工厂就此建好。

报表的生产过程中，对于超千万级含有网点的明细数据，可以再增加所属市、所属县机构号字段，适当冗余，以空间换时间，改善市、县级获取明细或统计数据的查询响应体验。

3.3.2　分类展示

为了实现第二项任务，需为报表设定“所属专业、省市县网点层级、具体所属市县”三维属性，为平台用户也设定“所属专业、省市县网点层级、所属市县”三维属性。平台按照属性关联，为登录人员展示属性相符的报表。

机构也有“专业和上级”二维属性，可参照集团公司 ERP 项目的机构表，结合 Web 方式向基层调查采集的机构数据，逐步清理出网点级机构在金融、邮务各专业生产系统中的机构号及上级机构对照表，为数据统计、一次登录多专业报表展示提供基础条件。

3.3.3　数据更新

下载平台海量数据环境下，报表的数据获取自然不能单独依赖通往下载平台原始表的数据库链接去完成。在自建的数据平台中，邮政单独开辟了一片表空间存放报表展示所需的数据表，这些表是数据分析人员加工后的结果表。所有的数据自动更新都在这类表上体现。自动更新对编写数据生成代码提出了效率的要求，对系统管理人员统筹安排提出了调度的要求。对此，在管理措施上，邮政拟定了数据分析人员

报表数据生成代码编写规范，明确每份代码服务的课题、涉及的原始表、中间表、结果表，给出单次运行所需的时间和更新时点及频次建议，供系统管理员参考。调度方面，操作系统的crontab、shell、c，数据库层面的oracle job与pl/sql procedure提供了基本手段。

3.3.4　工具效果

截至2016年12月，四川邮政省级数据平台共创建了60张图文并茂的数据报表。报表包括金融风险数据、潜在客户数据、业务到期客户、业务变动趋势图等多种类型，按金融、电子商务、寄递、文化传媒、金融风险作一级分类，再按统计、明细作二级分类。每一张报表都是一个数据分析课题，融入了四川邮政数据队伍近几年来聚焦业务、坚持不懈、求新沉淀的数据创新成果。

3.3.5　万能工具

关于报表工具，润乾、帆软（FineReport）都是可以考虑的选项。不过报表工具也非万能。仍需预留编程接口，允许数据分析人员或开发人员在平台加入jsp页面，嵌入可视交互echarts或电子地图等组件，在运行效率、展示特效方面解决报表工具不便实现的问题。

3.4　电子传递案例，拓展共享范围

依靠成熟的Web上传下载组件，可以实现成熟的数据应用案例和有价值的数据分析论文上传、修改、下载。这一措施虽然平凡，但解决了各地自建ftp服务器、http文件服务器、samba服务器，或者QQ、微信共享导致的资料零散问题，也让纸质数据运营期刊的内容传得更快、传得更远。

3.5　依靠技术市县，安全敏捷上线

3.5.1　传输安全

把报表应用和数据申请、自有库建设、案例共享功能部署在https协议下，可以增进传输安全，防范敏感数据被窃听。weblogic12.1.3以上的Web容器集成了https协议的有关组件，使得部署更加方便，可尝试选用。

3.5.2　IP绑定

IP与工号绑定使得特定的工号只能在具有注册IP的电脑上使用，IP绑定首先是一个Web开发问题，编码人员可以在开发过程中实现。IP绑定也是一个网络管理问题。在Internet环境中，由于ISP提供的公网地址可变，http上的IP绑定不具有应用价值。但在邮政专网的框架下，IP地址可以静态配置，省市间MPLS VPN +BGP模式下市州PE向BGP注入路由信息时，已被过滤了非规划网段的路由发布。市县、市到网点静态路由模式下，跨局域网的IP冒用更加不可能。这就是说，因为网络管理可信赖，跨市县、跨局域网的工号冒用可被禁止。局域网内的IP冒用防范，在技术上可考虑MAC和IP绑定策略、微小子网、专设以太线防嗅探，在管理上，敏感的机器不能让无关人员接触以免泄漏MAC或IP资料。

系统同时兼容实现了用户号与IP“1-1对应,1-n(多个离散值)对应，n-1对应、某用户不限定IP”，杜绝了敏感用户被跨局域网或者跨设备使用的情况。

3.5.3　短信验证

条件成熟时，启用短信验证也是一项有效的安全措施。时至今日，短信发送涉及的socket通信编码技术早已被各级开发人员烂熟于心，无需阐述。

3.5.4　分级维护

四川邮政省级数据平台要在全省21个市州、180个县（市区）、2 500多个网点应用，近4 000个工号，工号参数（使用者姓名、电话、IP地址、密码）可能随时变动。为此，四川邮政创建了分级配置工号参数的功能页面，采取了省预置工号角色，分级授权、分级配置工号参数的办法。在县辖网点机构复核、网点机构专业对应方面，采取了类似Web手段，依靠市、县基层的力量，确保平台在全省全面敏捷上线。

4　结语

从访问来源、访问对象、访问量分布看，截至2016年12月，经过一年的建设和推广，四川邮政省级数据平台60类报表被访问14万次，在全省2 340个网点得到应用，得到了各层级、多专业业务人员的广泛认可。下一步，四川邮政将发挥平台报表的无限叠加潜能，自动生成数据分析报告；运用报表工具的报表填报功能进行名址数据采集和客户营销反馈；引入新的算法让数据仓库数据在Web侧做分布式处理；面向省、市基于数据自建系统开放数据接口，引入统计工具尝试相关性分析。若省级数据平台定位准、姿态稳，必要时可对集团公司大数据平台发挥延伸作用。

参　考　文　献

1　谢希仁. 计算机网络（第七版）. 北京：电子工业出版社，2017

2　黄传河. 网络规划设计师教程. 北京：清华大学出版社，2009

3　Rechard Blum[美]著，武海峰译. Linux命令行与Shell脚本编程大全（第2版）. 北京：人民邮电出版社，2012

4　李刚. 轻量级Java EE企业应用实战（第4版）：Struts 2+Spring 4+Hibernate整合开发. 北京：电子工业出版社，2014

5　银行业专业人员职业资格考试办公室组编. 风险管理. 北京：中国金融出版社，2016

6　David M.Levine[美]著，黄耀锋译. 商务统计学（第5版）. 北京：中国人民大学出版社，2010

收稿日期：2017—03—30

作者简介：唐华（1976～），男，四川南充人，高级工程师，主要从事数据分析和平台建设研究；何太能（1979～），男，四川南充人，工程师，主要从事业技交叉融合和数据运营团队建设研究；傅国庆（1963～），男，四川成都人，硕士，高级工程师，主要从事数据挖掘与营销应用研究。

注：本文系中国邮政集团公司四川省分公司2016年数据平台重点项目“四川邮政数据综合应用平台系统建设”研究成果，项目编号为：川邮2016.21。