城市大数据的生态模型及应用

邓晖
中兴软创科技股份有限公司，江苏 南京 211153

摘要
从提出一个生态模型开始，阐述了建立一个可持续的城市大数据生态所需要的关键角色以及地方政府在演进这些角色中所能发挥的作用。接着，给出了一个实际案例作为这个模型的参考实现，并分享了案例中企业在配合政府建立大数据生态过程中所开展的一系列工作以及工作中总结的经验和教训，验证这个模型在实践中的可行性。最后，给出了一个具体的大数据应用案例：通过大数据手段来帮助政府优化行政审批流程，使得优化后的流程对市民更有利，从中一窥未来政府通过大数据进一步精细化社会管理的潜力。

关键词
大数据；产业模型；社会治理；社会服务；可信分析

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2016020

Big data ecosystem model and application in city

DENG Hui
ZTE Soft Technology Co., Ltd., Nanjing 211153, China

Abstract
With an abstracted model of big data ecosystem, the key roles which are necessary for setting up a sustainable big data ecosystem in city level market, were introduced. The local government's role on the evolution of roles in big data ecosystem were also discussed. An implementation reference of this model was demonstrated, with sharing a series of works in the implementation, as well as lessons and learned during work, which undertaken by the company, and the local government who cooperated with ZTE soft in the reference, to witness the feasibility of this model of big data ecosystem. Finally, an application case of big data technology was introduced to illustrate the potential capability of local government when moving forward to delicacy management of society.

Key words
big data, business model, social governance, social service, convincible analysis

1 引言

中国经过改革开放以来30多年的发展，城市化步伐不断加快，每年有1 500万人口进入城市，如**图1**所示。到2025年，中国将会有近三分之二的人口居住在城市，已经进入了一个城市社会。与此同时，城市人口的增加带来的交通拥堵、环境污染、资源过度消耗、各类突发事件增加等社会矛盾日益突出，各种“城市通病”与日俱增，城市管理难度加大，这对城市管理者的管理能力和服务水平提出了考验。城市要保持可持续发展越来越受到各种因素的制约，需要转变方式、调整结构、适应日益增长的人民生活方式、不断解决突发性事件等问题。人们在探索中意识到，智慧城市是医治“城市病”的最佳良药。

为了规范和推动智慧城市的健康发展，住房和城乡建设部于2012年12月5日正式发布了“关于开展国家智慧城市试点工作的通知”，并印发了《国家智慧城市试点暂行管理办法》和《国家智慧城市（区、镇）试点指标体系（试行）》两个文件，正式启动了全国智慧城市建设高潮。到2015年公布的第三批试点名单，共计289个大小城市加入了试点城市范围[①]，住房和城乡建设部智慧城市试点城市分布情况见**表1**。

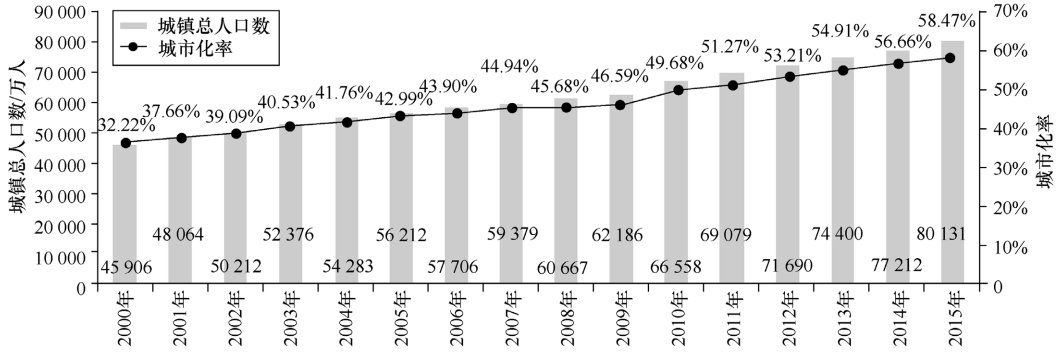
在一轮接一轮的智慧城市建设过程中，大数据技术在城市建设的应用逐渐成为智慧城市建设的热点之一。2015年中兴通讯股份有限公司（以下简称中兴通讯）把“以大数据应用为中心”的智慧城市建设称为“智慧城市2.0”，从而与之前“以建设业务系统为中心”的智慧城市建设区分开[②]。

2 城市大数据生态模型

2.1 城市大数据

在长期的城市建设与运营过程中，政府积累了大量的数据，如经济、民生、交通、旅游、医疗、安全等各行各业的数据。同时也积累了大量的业务系统。以重庆市为例，包括51个部门，平均每个部门有5~6个系统，整个政府有近300个系统在运行，如**图2**所示。

这些系统包含的数据涉及了城市的方方面面，其中蕴藏的价值亟需有效的手段进行挖掘与发现。



注：2012-2015年为测算值，未对外来人口及新增城市人口做区分

资料来源：《中国统计年鉴2011》、2011年统计公报、第六次全国人口普查

图1 2000–2015 年全国城镇人口数情况

① http://baike.baidu.com/link?url=rNZKUmzraibqD-L5Rf0u1qxYNmjEgLOo1BrxjARPZtwaKjKjuVFws7TRdLmhW2nL7o0JRy14eJAV7R3d-4uy8_

② http://www.cww.net.cn/news/html/2015/7/29/201577291713222299.htm

表1 住房和城乡建设部智慧城市试点城市分布情况

试点城市	第一批/个	第二批/个	第三批/个	合计/个
省会城市	5	5	0	10
地级市	30	36	37	103
县级市	18	30	33	81
区、新区	34	27	24	85
乡镇	3	5	2	10
合计	90	103	96	289

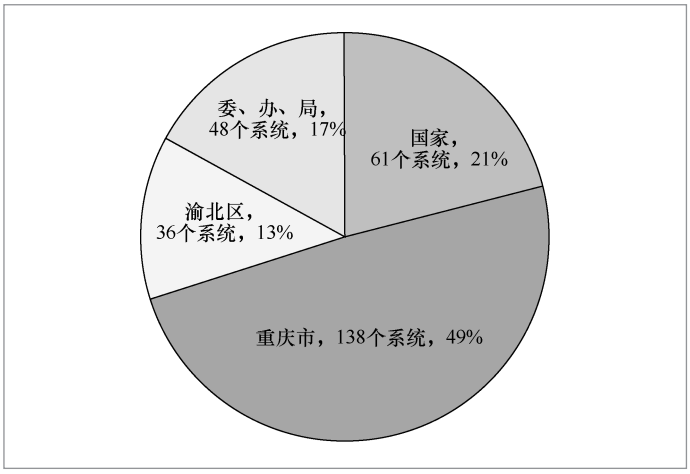


图2 重庆市应用系统按建设级别分类

与互联网公司所拥有的大数据不同，城市大数据具有自身的特点，见**表2**。

城市大数据与互联网大数据虽然各有不同，但可以互相补充，从而共同发挥更大的经济效益和社会效益。

2.2 大数据生态

大数据的潜在经济价值催生了大数据的交易。自2015年4月15日全国首家大数据交易所——贵阳大数据交易所正式挂牌交易起，先后有北京大数据交易所、上海大数据交易所、广州大数据交易所、陕西大数据交易所和长江大数据交易所等机构启动，围绕大数据交易开始形成一个生态系统，如**图3**所示。

围绕这个生态系统最外围的是工具厂商，这些厂商提供大数据采集、转换、存储、分析、可视化等技术手段。Google、Cloudera、Amazon等公司为大数据的技术推动做出了巨大贡献，同时大量的开源社区和产品逐渐成为大数据技术潮流的中坚力量。

处于生态中心位置的是大数据交易商。数据生产者为大数据交易商提供初级数据，后者通过数据标准化把初级数据转换成高级数据存储在基础设施运营商处。基础设施运营商通过提供存储服务和计算服务获得市场地位，并从中衍生出PaaS运

表2 城市大数据与互联网大数据的对比

大数据4V特征	城市大数据	互联网大数据
容量（volume）	单体数据[注1]比互联网数据小1~3个数量级	大部分单体数据记录数以十亿条、百亿条计
多样性（variety）	比互联网数据大1~2个数量级，数据涉及面极为广泛	大部分集中在某些消费领域，“单体海量”的数据充分体现了“单品海量[注2]”的互联网经济特征
快速（velocity）	城市大数据应用对速度的要求比互联网应用低1~2个数量级	互联网应用基本要求在线响应时间小于2 s，如大数据在广告、推荐、信用等领域的应用
价值（value）	城市大数据的价值比互联网大数据高1~2个数量级，可以应用于社会治理、城市管理、社会服务、经济发展等各个领域	互联网大数据主要用于企业自身经营水平的改善

注1：技术上数据常以“实体对象”的方式组织在一起，每一个实体对象存储为一条记录。单体数据意指这种单一实体对象类型所对应的整体的记录集，如电商的订单数据、银行的交易记录等。

注2：单品海量是互联网电商术语。由于互联网广泛接触的渠道，使得电商可以把一个品类的商品的销售数量做到传统的线下卖场难以想象的规模大小，电商把这种现象称为“单品海量”。

营商来减轻大数据交易商对技术的需求。

数据挖掘者通过基础设施运营商提供的计算服务，结合领域知识对高级大数据进行挖掘。获得的有价值成果通过大数据交易商的交易平台提供给数据消费者。这样吸引更多的数据消费者源源不断地加入交易平台，提出更多的问题并支付获得的满意答案，由此衍生出大数据咨询商来引导大数据消费者更好地消费。

数据消费者的支付通过大数据交易商的分成平台，使得数据生产者也获得相应的回报，这样整个生态就能够自生自长，实现良性循环。

在真实世界里，一个企业可能兼有这个生态中的一部分、一个或多个角色，或者只专注与某个细分领域，如“中关村大数据产业联盟”就活动在“咨询”这个细分领域。

从大数据交易生态系统也可以看出，有些领域如“数据标准化”，多数有识之士觉得很重要，但标准化并没有先行发展起来。究其原因就在于标准化其实是一个局部问题而非全局问题。另外，这个产业模型让数据本身不需要发生大规模移动或复制，避免了数据安全、个人因素、产权归属、初级或高级数据定价等复杂的社会问题和商业伦理问题，为持续进行交易提供理论依据。

2.3 政府与数据交易商

在大数据交易这个生态系统里形成初始的生态平衡是非常关键的。在全国此起彼伏的智慧城市建设以及交易所设立的浪潮中，政府有天然的优势来孵化数据交易商角色，具体原因如下。

（1）政府是最大的数据生成者

如前所述，城市大数据基本都在政府手中，另一小部分在党政机关或事业单位

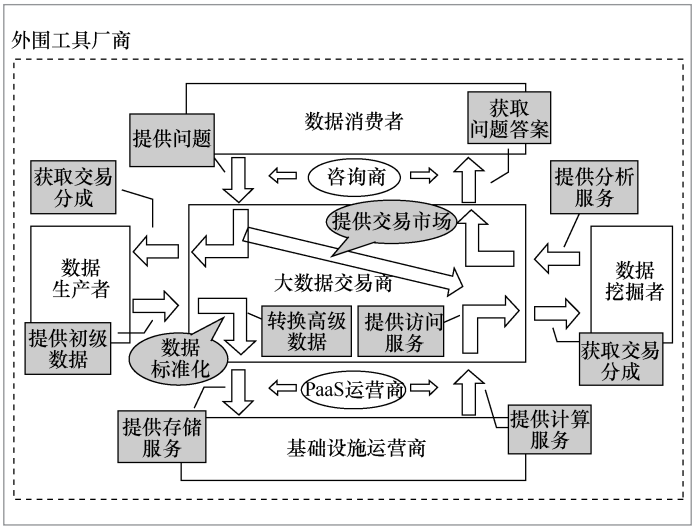


图 3 大数据交易生态系统模型

手中。另外，随着智慧城市系统的建设，政府手中的大数据会越来越完善，越来越动态。

（2）政府是最大的数据消费者

我国的社会治理模式是“大政府、小社会”的模式，要求政府对国计民生方方面面做好保障与服务工作。这些工作的顺利开展都离开不科学决策，离不开数据的支撑。随着大数据价值的不断发现，政府治理的效率也将不断完善。

（3）政府是城市公共设施的提供者

政府也将持续为每个城市的健康运行提供必要的基础设施，如交通、水利、学校、医院、水电煤气等。而信息基础设施正在成为城市越来越重要的基础设施之一。到2013年上半年，全国共规划建设数据中心255个，已投入使用173个，总用地约713.2万平方米，总机房面积约400万平方米。

可见，在大数据生态的5个核心角色中，政府已经身兼3个角色。

另一方面，社会资本在目前的产业环境下承担大数据交易商角色仍有很多挑战。

- 大数据交易的法律法规、信息安全

标准等宏观政策还不完善，导致各种市场主体对参与大数据交易持观望态度。

● 大数据交易还没有看得见的成熟的商用模式，能否在预期的投资周期里获得投资合理回报是一个很大的问题。

● 由于大数据交易对象的高度技术化，如何吸引大规模的用户，认同交易物有所值，需要强大的信用支撑来鼓励各种用户先行尝试。

在大数据产业初期，通过政府投资，其他社会资本参与成立数据交易商是一个比较现实的选择。政府可以在实践过程中打通产业各个环节，迅速完成法律法规建设，通过PPP（public-private-partnership,公私合作）模式、政府采购服务以及财政补贴的方式来为新兴产业提供资本和信用保证。

3 实践案例

中兴通讯股份有限公司和银川市政府共建智慧城市是大数据生态系统产业模型的一个实践案例。其中，中兴软创科技股份有限公司作为数据挖掘者参与了银川市城市大数据的合作开发；银川市政府承担数据提供者和数据消费者的角色；银川市与中兴通讯合作组建的中兴（银川）智慧产业有限公司承担了交易商和基础设施运营商的角色。合作开发过程主要围绕“基础设施、技术架构、获取数据、分析列表、分析人才、分析过程和决策应用”7个方面展开。

3.1 基础设施

在目前的技术条件下，获得城市大数据运营所需的基础设施其技术困难不大。以银川市为例，从动土开工到大数据中心投入使用，整个工期不到一年，总体成本对于一个城市而言不高。也可以采取租赁互联网公司数据中心的方式，但考虑数据安全、运维成本、区位优势等因素后，城市自建大数据中心仍然是主流选择。

3.2 技术架构

满足城市大数据开发需要的技术平台也不难搭建。以笔者研究团队的经验，这个平台应该包括大数据采集器、数据中心、主数据管理、大数据分析器、大数据服务器、可视化服务器、大数据客户端7个部分，技术才算是比较完整的，如图4所示。

大数据采集器能够实现海量数据的收集，不管是结构化数据还是非结构化数据，文本、语言、视频都能实现数据的采集、清洗、整合、转换和装载，这些数据最终存储在数据中心。

数据中心从软件与硬件层面实现对海量数据的存储和访问，同时注重能耗与安全。

主数据管理则实现数据的编目、管理、授权、共享和交换，维护城市数据模型，形成五大库（即人口库、法人库、地理信息库、建筑物库和宏观经济库），并维护各自的过程库、业务数据库和主题应用库等。

大数据分析器根据问题、目标，设计出分析模型及数据处理、训练、检验过程，将设计好的蓝图交给大数据服务器计算。

大数据服务器管理所有的计算资源，实现分布式计算、海量数据即时处理。

可视化服务器把大数据分析结果转换成图形，直观地告诉客户所拥有数据的形态和关键特征，这些图形最终通过大数据客户端向用户呈现。

大数据客户端包括如下3类。

● 数据服务平台：面向公众，以网站的形式向公众提供大数据开放服务，鼓励大众参与城市服务。

● 决策服务平台：面向各级领导，通

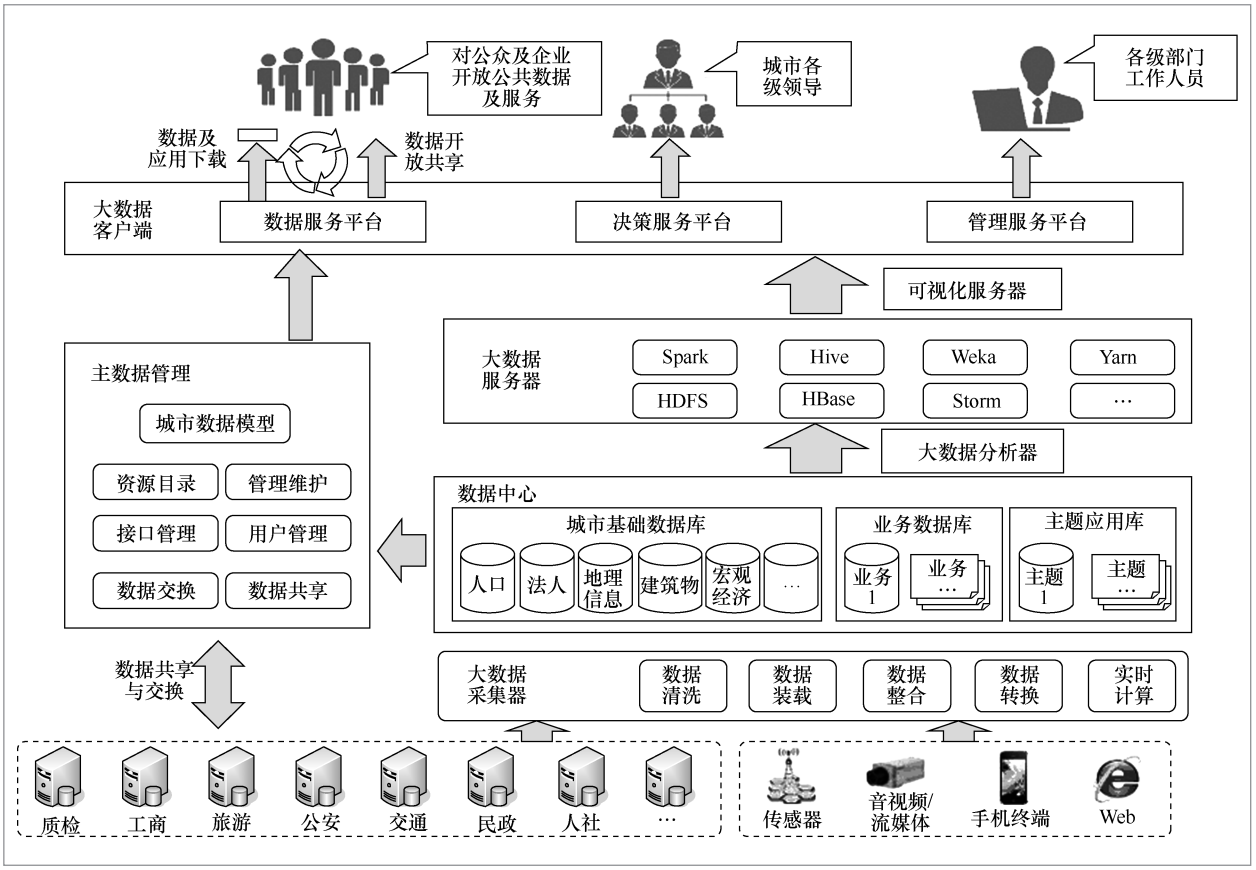


图 4 满足城市大数据开发所需要的技术平台架构

过图表方式呈现经济、民生等数据的分析结果。

● 管理服务平台：面向政府工作人员，通过缩放地图、拉动时间线来查看其感兴趣的数据，如街道主任可以限定自己所处街道查看人口出生率，而同级教育主任可能更关心扫盲率。

3.3 获取数据

在城市大数据开发过程中，真正的困难是从获取数据开始的。从产业模型角度看，属于培育大数据生产者的工作。

首先，政府部门开发自己的数据意愿很低。这其中的原因非常多，包括政策上的顾虑、部门立场的考虑以及公开数据可能带来的种种问题和对变化的担忧。在这些因素里，数据安全是一个绕不开的话题。2015年刑法修正案在信息安全领域明确扩大了犯罪主体的适用范围[③]，使得部门主要领导和相关负责人都不愿意承担因数据泄漏风险引发的连带责任。为了让政府部门的数据能够更有效地共享，除了技术上不断提高，加大数据保护的力度之外，在法律、制度上进一步细化和松绑已成为不可缺失的一环。商业上的创新也比较关键，比如考虑一种保险制度来解除大数据共享过程中所引发的安全责任风险。

其次，数据预处理（即把低级数据加工成高级数据）的工作量非常大。一方面，政府的系统建设过于分散，都是大量的小厂商开发出来的，数据规范性一开始就不高；另一方面，这些政府系统一开始没有考虑向大数据分析优化，缺失数据严重，

③ http://legal.gmw.cn/2015-11/12/content_17705238.htm

而不同系统之间的数据一致性更加没有保障。这就要求厂商花出大量的时间进行数据查漏补缺，通过不同的数据源进行相互验证来获得更加完整、准确的数据集。在这个预处理过程中，本身也有一些大数据技术在其中应用，比如通过数据分析来判断哪些数据集准确性更高，从而替换其他重复数据。

另外，数据格式五花八门，有很原始的表格、文本数据，也有纸质数据，需要通过OCR扫描识别入库。

3.4 分析列表

有了数据之后，接下来就是要有分析目标。从产业模型角度看，属于培育大数据消费者的工作。

在培育消费者方面，目前比较新颖的做法就是大数据竞赛。例如，2015年8月在上海举行的开放数据创新应用大赛，奖金最高达20万元④。

传统的做法是需求调研，通过和各委（员会）、办（公室）、局座谈来发现他们工作中的难题，并从中找到大数据可以胜任的问题列表。这种方式由于信息不对称，导致效率比较低。因此，在选择部门时应该考虑部门的业务特点、部门积极性和领导人风格来安排优先次序。

3.5 分析人才

企业获得合格的大数据分析人才不是一件容易的事情，主要是因为大数据分析人员不仅要熟悉大数据工具、技术，还需要精通数理统计以及有足够的社会通识，才能通过一层层数据关联关系找出问题的答案。

一种可行的办法是通过2~3个小团队高效协作的方式来解决，类似“戚家军”的战斗组织形态，这样可以整体降低对人才的需求门槛，使得产业模型里的数据挖掘者可以规模化。

3.6 分析过程

分析人员在针对具体问题进行分析前要学习很多算法，除此之外还要关注如下重要的问题。

（1）评估方法是关键

算法要在新数据上的表现和在样本数据上的表现几乎一样好。比较好的做法是把数据集一分为二，一部分用于训练模型，一部分用于模型评估。交叉验证，观察算法的稳定性。如果算法不能稳定下来，那么结果是非常可疑的。因为服务的领域是公共服务领域，如果一旦错误执行，就会存在很大危害。另外，训练模型也不能训练过度，避免出现过度拟合的问题。

（2）特征提取是根本

分析人员不要迷信算法，大多数复杂算法效果大同小异。但要确保完全理解这些等价算法中的一种，然后一直用下去。

在分析过程中如果能找到合适的特征，对于达到分析目标所需的样本数据量就能大大缩减。数据分析人员需要完整地掌握各种特征工程来快速找到样本数据的特征向量。如果分析人员非常懂业务，也可以弥补特征工程经验不足的短板。特征提取是大数据分析非常重要的成功因素。

（3）时间瓶颈是模型训练，而不是数据集规模

在模型训练过程中，需要花费大量的精力进行参数优化，从而得出比较合理的解。在承诺给政府部门一个分析结果之前，应该充分留有这部分的时间。

另外，还有“数据自大”问题，很多人拿到了数据以为很大，其实这只是很小的部分，但他自己不知道，所以结果会出现偏差。还有就是算法演化问题和数据

④ http://www.chinanews.com/auto/2015/08-18/7474445.shtml

生产者的看不见的动机，这些都会导致分析结果和实际出入较大，分析时需要仔细甄别。

3.7 决策应用

当数据分析人员把一个分析结构给政府相关部门，报告里面的结论是否就会很快被采纳？其实不一定。分析结果不能及时应用主要包括如下原因。

（1）大数据分析透明度不足

大数据分析由于算法上的艰深难懂，除专业人士之外，其他人很难搞懂，导致最终的分析结果很难证明其结果是正确的、中间的分析过程是可靠的，使得政府不是非常愿意主动采信这样的分析结果。

（2）缺乏第三方机构的验证

如果有第三方机构验证也能促进政府放心使用大数据分析结果，使政府决策更具科学性。但企业因为商业机密方面的原因，不愿意公开分析过程中的数据模型，导致第三方没有合适的验证方式。

（3）多方位分析结果相互不支持

有时确实会出现多个分析结果打架的情况，这时候需要仔细排查，分析是数据上的原因还是算法上的原因。但有时这样的分析结果没有及时发现就报给政府部门，将导致相关部门对分析结果的可信度更加担忧。

如何提高大数据分析结果的可信度，笔者认为可行的办法是改变应用方式。由传统的“报告式”结果呈现转变为“探询式”结果呈现，中兴软创科技股份有限公司在这方面正在进行较大的技术创新。例如，对政府行政审批数据进行了一个预测分析，通过KNN回归模型来预测每一类行政审批事项当前最合理的办件承诺时间。这个承诺时间是动态变化并适配外部环境变化（如收件量、工作人员状况、时节、社会热点等）的，从而让这个时间更加科学。

4 结束语

本文介绍了笔者在城市大数据开发方面的一些经验。这个领域还有很多重要问题需要一一面对，如能耗与污染、信息模型与标准库、可靠性与可用性等。在工作开展的过程中会遇到很多现实困难，但更多的是解决办法。其中，发展大数据生成者和消费者并建立完整生态依然是发展大数据产业的重中之重。

国务院发布了《国务院关于印发促进大数据发展行动纲要的通知》⑤，中国的大数据产业已经势不可挡，必将开始一个新的智慧城市时代。

⑤ http://politics.people.com.cn/n/2015/0905/c1001-27545655.html

作者简介

邓晖（1974-），中兴软创科技股份有限公司智慧产品部副部长。1999年毕业于哈尔滨工业大学机器人研究所，加入中兴通讯股份有限公司计费产品线。有15年的电信行业产品研发、交付及管理从业经验。历任高级研发工程师、系统架构师、大项目经理、产品经理、客服部长、市场总监等职务。有丰富的国内（国际）市场、研发、交付工作经验和国家（行业）标准编写经验，多次参与智慧城市的顶层设计。

收稿日期： 2016-01-13

* 本文为2015中国大数据技术大会（BDTC）演讲约稿