面向数据自治开放的数据盒模型

熊赟[1,2]，朱扬勇[1,2]
1. 复旦大学计算机科学技术学院，上海 201203；2. 上海市数据科学重点实验室，上海 201203

摘要
在数据自治开放模式中，数据使用者可以将数据上传到应用软件或下载数据到其设备中。其中面临的问题首先是让数据使用者看得见有哪些数据，其次是如何合理组织数据方便应用软件使用、计量和管控。提出了一个数据盒模型作为数据自治开放环境下的基本数据模型，介绍了数据盒的基本要素组成及运作系统架构，说明了该数据模型的具体实施方式。

关键词
数据开放；数据共享；数据自治开放；数据模型

中图分类号：TP311 文献标识码：A doi: 10.11959/j.issn.2096-0271.2018015

Data box: a novel data model for self-governing openness of data

XIONG Yun[1,2], ZHU Yangyong[1,2]
1. School of Computer Science, Fudan University, Shanghai 201203, China
2. Shanghai Key Laboratory of Data Science, Shanghai 201203, China

Abstract
Self-governing openness of data makes that users can download or upload data to their applications to use data. However, there exist two key issues. Firstly, how to know the contents of the data. Secondly, how to organize data to fit the users' applications feasibly and manage data conveniently. A novel data model namely data box was presented, the elements of data box and operating system structure were introduced, and the implementation of the data model was illustrated.

Key words
open data, data sharing, self-governing openness of data, data model

1 引言

20世纪60年代国际科学理事会的国际科技数据委员会（CODATA）推动了科学数据开放①，但开放程度有限，并且大多集中于各自领域，如地震科学、水利科学、天文学等[1-3]。随着数据资源的战略性和商业价值越来越突显[4]，数据资源的开放共享变得越来越困难，目前为止，尚未形成完全开放的数据共享局面。20世纪60年代，数据库管理系统（database management system，DBMS）的研制成功[5,6]使数据脱离了具体的应用程序，由DBMS统一管理，这种数据资源管理模式是封闭的、内部授权的、有限的。而在数据开放的环境中，应用程序和用户对数据的访问和使用是开放的、不可预知的、无限的，数据的结构形式也是灵活、复杂多样的。在数据自治开放模式中，数据由数据拥有者在法律框架下自行确权和管理、自行制定开放规则（即数据自治），然后将数据开放给使用者，包括上传数据到应用软件使用数据或下载数据到使用者的设备中。其中面临的问题首先是让数据使用者看得见有哪些数据，其次是如何合理组织数据以方便应用软件使用、计量和管控[7,8]。

然而，现有的数据资源管理技术（数据库管理系统、文件系统）和应用软件技术针对处于封闭不开放状态的数据，聚焦于尽量保护系统数据不受外界侵害，即针对数据安全风险和隐私威胁进行研发，主要解决途径包括密码技术、噪音扰乱技术、访问控制和软件行为监控机制[9-13]。这些方法虽然在一定程度上保护了数据的安全和隐私，但是在保护数据权益上有以下缺陷。

- 诸多加密和多授权机制、安全策略、信任级别等方式是采用数据封闭的思路解决数据非法访问问题的，无法用于数据开放下保护数据权益和数据机密性。
- 差分隐私等隐私保护方法大多针对静态数据集，且通常是针对单点式数据需求，即假定数据只有一个使用者，因而不能很好地适应数据开放下的多用户数据访问的动态应用需求；更重要的是，由于添加噪声导致结果丢失准确性，限制了数据的使用，不利于数据开放。
- 开放数据的访问/应用软件往往是事先未知的，当使用方在特定条件下对数据资源有修改能力时，数据仍存在被零成本复制的风险，开放的数据资源需要额外保护机制，这需要制定软件访问数据的行为规范，并对数据访问行为建模和持续监控。

因此，需要有新型的数据组织方式合理组织数据，以适应数据自治开放模式。本文提出一个数据盒模型作为数据自治开放环境下的基本数据模型，称为“数据盒”，以有效组织用于开放的数据资源。数据盒是数据自治开放的基本单元，包括数据描述、数据操作和约束等基本要素，并通过在数据盒中封装数据防泄露和数据权益保护等机制，使之具有独立性、可用性、可控性，使得数据拥有者在数据开放的同时又能保证数据稀缺性不丧失和隐私不泄露，并且不影响现有系统，能有效支持数据自治开放。

2 面向数据自治开放的数据模型

面向数据自治开放的数据模型的核心思想是建立基于“数据盒”模型的开放数据组织，按照数据盒的方式向数据使用者开放数据，即呈现给数据使用者的是一定

① www.codata.org

数量的数据盒，并且，在数据盒中封装数据防泄露和数据权益保护等机制。也就是说，数据盒是带有自主程序单元和内在计算能力的数据组织存储模型。数据拥有者将数据灌装入数据盒中，封装的数据只能通过数据盒中的自主程序单元接口进行受控的访问。这样做的好处是既方便数据使用者使用开放数据，即外部可见、可理解、可编程，又能防止数据拥有者权益受到侵犯，即内部可控、可跟踪、可撤销。面向数据自治开放的数据模型包括以下几个方面。

- 为数据使用者提供开放数据的基本组成单元——数据盒，如**图1**所示，包括数据盒的数据描述、数据操作和数据约束等基本要素以及数据盒的性质。
- 封装在数据盒中的数据防泄露和数据权益保护机制，并设有供外部软件使用的接口。
- 数据盒的计量与定价，即根据数据使用者提出的要求和目标，计算数据使用者所需数据盒的数量和时间等，并进行定价。

2.1 数据盒基本要素

（1）数据盒的数据描述

因为数据盒的设计目标是面对数据开放环境的，数据来源复杂多样、不可预知，所以数据盒的数据描述包含数据结构。数据结构采用BSON（binary serialized document format）存储复杂类型的数据。BSON使用C语言风格的数据表现形式，其编码和解码都是快速的。每个数据盒设有唯一的标识名（字符串类型），称为数据盒标牌。每个数据盒还设置数据使用的相应软硬件环境。由于开放数据盒为数据使用者服务是在分布式环境下进行的，所以数据盒的性质也参考分布式存储的性质，即数据盒性质包括最终一致性和基本可用性。其中，最终一致性是指要求系统数据副本最终能够一致，而不需要实时保证数据副本一致；基本可用性是指出现故障时，允许损失部分可用性，保证核心功能或当前重要功能可用。

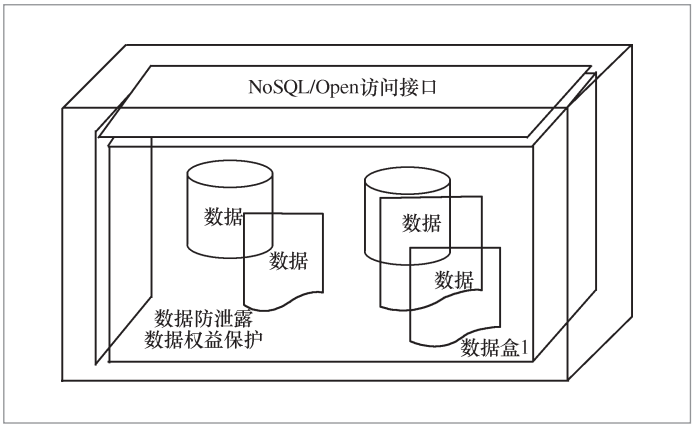

图1 数据盒的组织结构

（2）数据盒的数据操作

数据资源以数据盒的形式存放在数据站中，与传统的DBMS相比，数据开放自治下的数据资源管理不涉及事务处理，只有数据使用，但也不同于数据仓库，数据仓库是不对外开放的。数据盒的数据操作包括数据盒的创建、数据盒的删除（当数据盒不再被使用时，则需要删除该数据盒）、数据盒的切块和切片（数据盒中存放了多条数据记录，每条数据记录有多个字段/属性，数据盒使用者可能只需要使用该数据盒中的部分数据记录和部分数据字段，这需要数据切块或切片操作，即选取数据盒中的部分数据记录和部分数据字段，形成新的数据盒，切块或切片后的数据盒是原始数据盒的一部分，包含原始数据盒的部分数据字段和数据记录）、数据盒的拼接（数据盒的使用者可能会使用多个数据盒中的数据记录，并且可能是这些数据盒中的部分数据字段，因此，需要将多个数据盒拼接起来，形成新的数据盒）、数据

盒的交(数据盒的使用者可能需要两个数据盒共同的数据记录,数据盒的交实现共同数据记录的提取,形成新的数据盒)、数据盒的并(数据盒的使用者可能需要两个数据盒的所有数据记录,数据盒的并实现两个数据盒中的数据记录的合并,形成新的数据盒)以及数据盒中数据的查询(通过设置一定的条件,对数据盒中的数据进行查询,例如某个字段满足一定的阈值作为查询条件等)、插入(当数据盒有新的数据装入时,需要执行插入操作)、删除(数据盒中的数据不再被需要的时候,需要执行删除操作)和更新(数据盒中数据记录的某些值发生变化时,需要执行更新操作)等。此外,数据开放的粒度是有差异的,数据盒粒度划分也是数据盒的重要数据操作。

(3)数据盒的数据约束

数据盒的数据约束包括数据盒约束、数据盒内的数据约束、数据盒间的数据约束、数据盒多副本控制约束等。数据盒内的数据约束类似于传统数据库中的数据约束;数据盒间的数据约束主要是指数据盒之间的关联约束;数据盒多副本控制约束主要是指数据盒可以拥有多少个副本,这些副本可被哪些数据使用者使用(例如,对使用数据盒的机器的MAC地址加以限制等)。

2.2 数据权益保护机制

数据权益保护机制是将数据盒及其权属绑定为有机整体,以明确数据权益的保护对象、保护等级以及保护粒度等。其中,保护粒度是指数据盒中的数据记录可被访问的粒度,例如,数据字段“薪水”的访问可以分为访问每一条数据记录“薪水”的具体值、访问“薪水”的平均值或统计值等。数据盒权属包括数据所有者的信息标记、开放数据的使用权限以及相应的使用许可范围、定位跟踪标识等。外部软件可通过数据盒接口访问数据盒中的数据,因此,数据权益保护机制还包括外部软件行为监控机制,对外部软件行为进行监控,评价软件的言行一致性,甄别权益受损的软件行为模式。

2.3 数据防泄露机制

由于数据使用者可能通过多次获取数据片段实现对数据盒中数据的拼图,形成一个完整的数据盒,导致数据泄露。数据防泄露机制包括:数据盒正常使用情况下的数据加密保护,实现关键数据密文形式共享;数据盒被盗取或控制情况下的数据机密性保护,提供数据盒抗盗取和抗逆向拆解能力;数据监控保护,发现数据盒的不正当使用时,启动数据盒自毁机制。

2.4 数据盒的计量与定价

数据盒的计量是指根据数据使用者的要求,对数据使用需求建模,根据不同的需求粒度切分数据盒,为数据使用者提供数据资源;数据盒的定价是指按照数据盒的定价规则对数据使用者进行定价。

综上,按照数据盒为基本单元组织数据,并封装数据防泄露和权益保护机制以及提供数据盒计量和定价策略,为数据开放提供技术保障。

3 数据盒运作系统框架

以数据盒形式展示的数据资源存放在数据站中,每个数据站配备一套数据资源管理系统,用以管理该站下的所有数据

盒，并通过数据盒虚拟化、应用装载等功能供外部使用数据。其具体运作的系统如**图2**所示，包括数据源管理模块、数据盒构建与环境配置模块、数据灌装模块、交互模块。

各个模块介绍如下。

（1）数据源管理模块

数据源管理模块即数据站配备的数据资源管理系统，用于管理数据拥有者的数据源，这个数据源可能是单点的，也可能是分布式的。数据拥有者想要开放的数据都从该数据源中获得，而且数据拥有者本身并不需要关心数据源的组织方式和存储格式。

（2）数据盒构建与环境配置模块

数据盒构建与环境配置模块负责数据盒的生成以及数据盒环境的配置。

依据数据盒的大小（数据盒大小由数据拥有者提供的该数据盒中的数据记录的多少决定，即由数据记录的字段数、数据记录的条数决定数据盒的大小）、约束条件、接口和监控要求以及环境需求（由数据拥有者提供的该数据盒初始使用环境，即使用这些数据可用的软硬件环境，例如，GPU、内存、操作系统配置等）构建数据盒，并配置数据盒环境。数据盒大小实际是根据数据记录的字段及其条数、值决定的。

（3）数据灌装模块

数据灌装模块负责将数据源进行格式转化，导入数据盒。从数据源中获取的数据可能是有不同格式的，在经过一定的格式转化之后，形成BSON以支持数据盒的数据灌装。

（4）交互模块

交互模块负责承载数据使用者使用数据盒的展示、交互。数据使用者可以通过该模块的交互来获得和使用数据盒。该

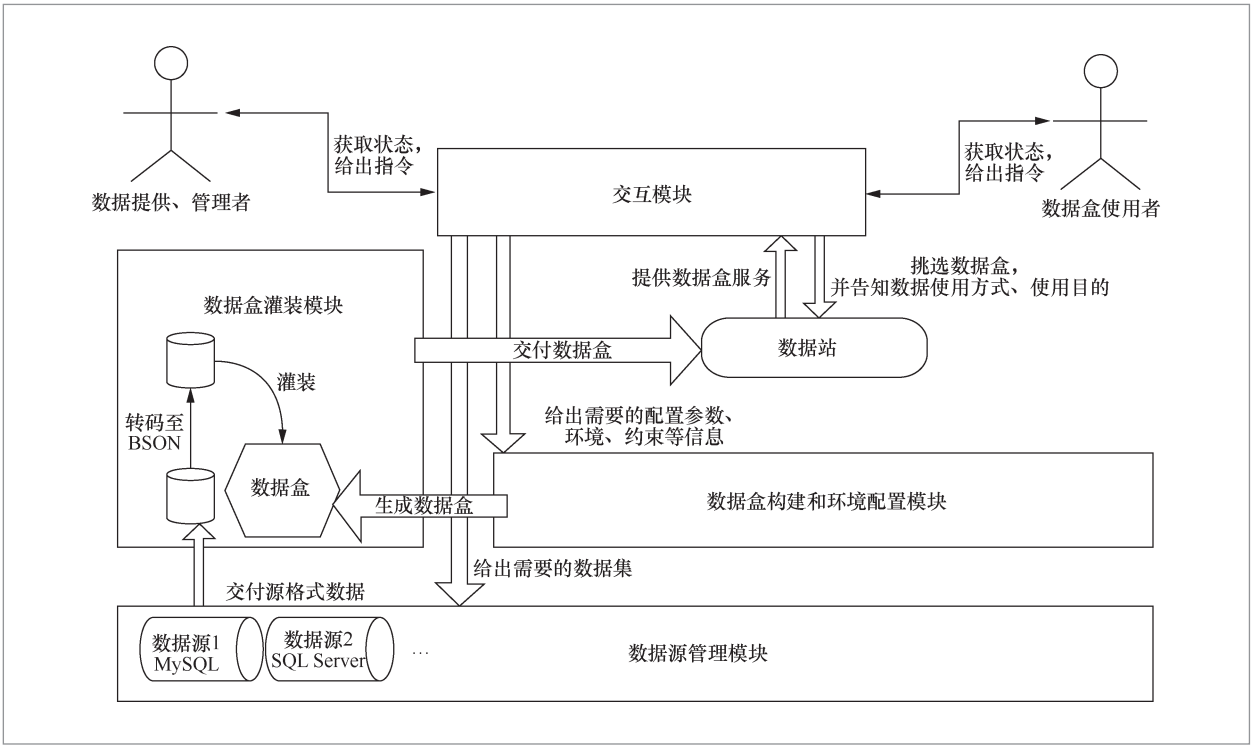

图 2 系统模块组成

模块把最后得到的数据盒交付给数据使用者。

该系统的运作流程如下。

● 准备数据：数据拥有者选定希望开放的数据，确定数据字段和数据记录，并给出数据的使用约束，通过数据源管理模块将数据交付给数据盒灌装模块。

● 数据盒构建与环境配置：配置数据所需的软硬件环境，通过数据盒构建与环境配置模块将软硬件环境配置信息封装到数据盒中，生成数据盒，并分配数据盒标识。

● 数据灌装：在数据盒灌装模块，实现原始数据格式转换，转码为BSON，将数据拥有者想要发布的数据记录按照其约束条件实施数据盒的灌装，并且封装数据访问控制和数据防泄露机制。

● 数据盒交付：将生成的数据盒交付给数据站。数据使用者通过交互模块，挑选所需的数据盒，并告知数据盒的使用方式和使用目的，根据数据使用者提交的数据使用需求和数据访问程序及其声明，数据拥有者对数据盒进行定价，数据站为数据使用者提供数据盒及其服务。

通过上述步骤，数据拥有者可以以数据盒为基本组织单元组织所拥有的数据资源，放在数据站中，对外开放发布；数据使用者可以发出数据使用需求，从数据站中申请使用数据盒。

以数据盒为基本单元的数据自治开放数据资源组织使用示意如图3所示。

4 数据盒模型的特性和效果

从数据盒的基本要素组成以及运作系统架构可以看到，数据盒是一个类似于现实世界中用来盛放物品的盒子的数据装置，盒中存放的是数据源。盒子便于运输，也便于数据的流通。数据盒可以配置不同大小的尺寸，可以根据数据使用者的需求装入不同大小的数据。每个盒子设置一个标识，称为数据盒标牌，用于识别数据盒。

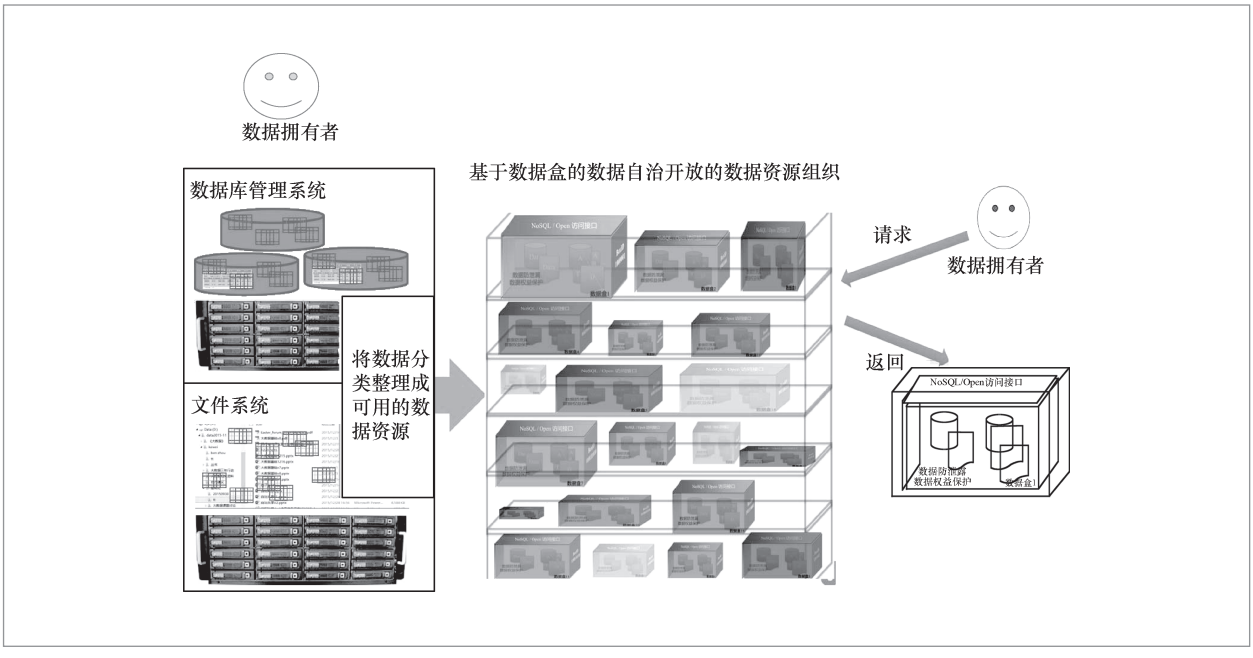

图 3 以数据盒为基本单元的数据自治开放数据资源组织使用示意

数据盒为数据使用者提供数据的访问接口，数据使用者可以声明其访问需求，通过接口访问数据盒中的数据。考虑到数据盒的使用权限，数据盒还存放数据防泄露和数据权益保护程序。每个数据使用者可以申请获取一系列（一个或多个）数据盒。数据盒将数据拥有者的数据、访问权限等进行了有机合理的组织，使得数据使用者能够很方便地看见有哪些数据被开放、数据的结构和组成等，按照数据盒的方式向数据使用者开放数据，使数据拥有者可以根据数据的开放程度进行数据的把控，实现了数据内部可控、可跟踪、可撤销；并且，数据盒的集成形式为数据使用者/应用软件的使用、计量、管控提供了便利。数据使用者看到的是一系列封装好的数据盒，可以根据数据使用需求方便地选择使用数据盒。对数据使用者而言，开放的数据是外部可见的、可理解的，数据使用者也可以提出自己对数据的访问需求和程序，让数据盒拥有者进行审核，实现数据盒可编程。这种数据资源组织形式提升了数据的独立性、可用性和可控性，实现了数据自治。

5 具体实施方式

数据盒的生命周期包括以下几个阶段。

- 创建：选择/生成一个具有指定存储和计算能力的基础映像实例。
- 灌装：向映像实例中添加用户所选的数据，还可以添加用户根据数据自治开放标准编写的程序模块。
- 运输：将灌装了数据的数据盒运送给用户，运输方式可以为在线，也可为离线。
- 加载使用：用户根据使用说明在本地加载数据盒，所有访问受到数据盒访问控制的限制。用户可以要求重新灌装数据或自定义模块。
- 销毁：到达指定的访问期限后，数据盒自动销毁内部数据。

面向数据盒的软件开发和运行平台主要为数据盒的加载和使用环节提供支撑，着重解决数据盒应用开发和运行的技术问题，对传统的集成开发环境和运行环境进行了必要的技术改造和扩展。

下面以数据拥有者如何将其数据资源构建为若干数据盒用于开放为例，说明本数据模型的具体实施方式，如**图4**所示。

（1）生成一个初始数据盒

使用Java中的Docker-Java-API从基础镜像创建容器，基础镜像中有Java、Tomcat、MongoDB等软件环境，生成一个初始数据盒。

（2）确定要装入数据盒的数据

如**图4（a）**所示，数据使用者选定需要装入数据盒的数据，确定需要开放的数据字段、数据记录等。例如，数据使用者有一个用MySQL存储的医疗数据库，包含病人基本信息、医生基本信息、门诊就诊信息等多张表。数据拥有者希望开放的是用药数据，这涉及病人基本信息表的非敏感字段，例如数据表中的病人就医顺序号（非身份证号或医保卡号等）、病人疾病诊断、病人用药信息等，这些数据分别在不同数据表中。数据使用者首先选定这些所需字段以及想要开放的数据记录条数（或时间区间等），例如，开放2017年1—6月的数据。

（3）根据选定的数据，设置数据盒的大小、约束条件等

如**图4（b）**所示，若数据使用者要开放的数据是2017年半年的数据，那么，计算2017年半年的数据量（例如20 GB），根据数据量，设置该数据盒的大小。数据拥有者根据数据的特点和使用权限，设置约束

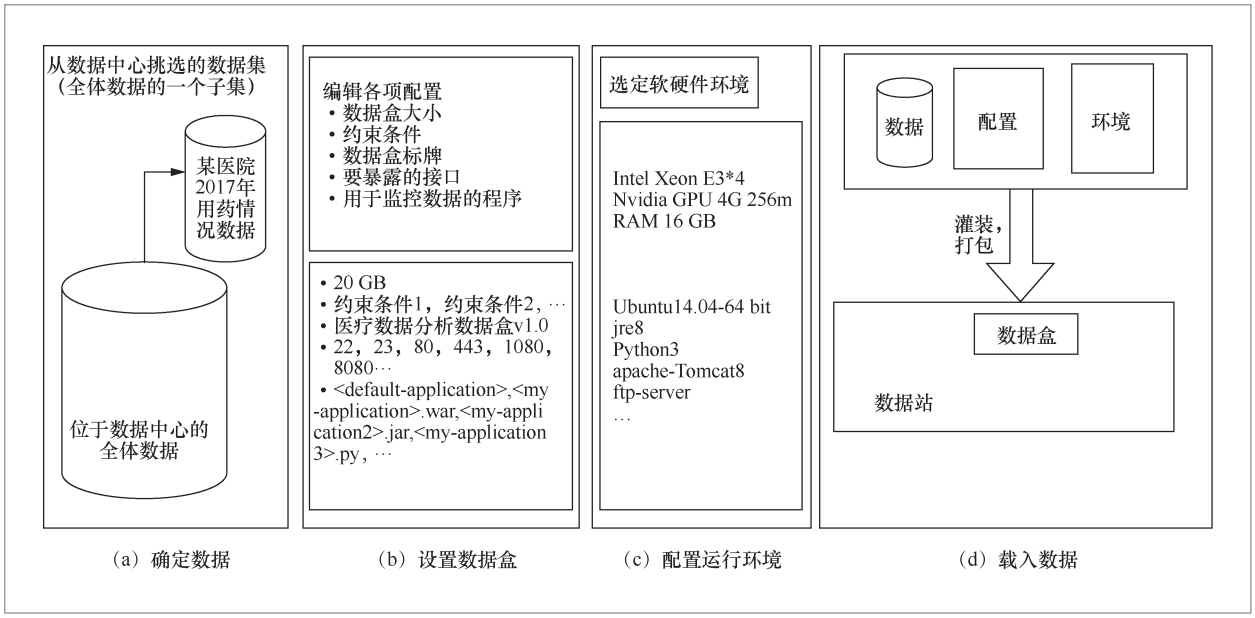

图 4　数据盒的实施步骤

条件，例如，不允许将这个数据盒与涉及病人个体信息的数据盒进行拼接，或不允许对该数据盒中的某个字段（如年龄字段）进行逐条访问；数据盒的多副本约束即约束该数据盒的使用者数量等。

（4）在灌装数据拥有者数据时，配置数据盒对应的运行环境

如**图4（c）**所示，选定需要的软硬件环境，例如，Intel Xeon E3*4、Nvidia GPU 4G 256 m、RAM 16 GB等硬件环境以及Ubuntu 14.04-64bit、jre8、Python3等软件环境。

（5）灌装和打包数据盒，把数据装载进数据盒

如**图4（d）**所示，例如对于数据源为MySQL的数据，可以使用Java中的MySQL-connectorAPI来读取数据源中的数据，并进行数据格式转换，将这些数据输出到文件中，保存为BSON。将BSON文件复制到创建好的容器中，并将BSON文件读入。将一个用于向外界暴露数据获取接口的网站应用以war包的形式复制到容器中，并在Tomcat中启动。建立数据盒标牌。打包这个镜像，并将镜像以文件形式提供给数据使用者。

（6）在该镜像中封装数据盒防泄露机制

数据盒防泄露机制作用如下：数据盒正常使用情况下的数据加密保护，实现关键数据密文形式共享；数据盒被盗取或控制情况下的数据机密性保护，不允许数据盒抗盗取和抗逆向拆解；根据数据拥有者的数据约束，不允许数据使用者对数据盒进行非法拼装；控制数据使用者的数据记录数量等。

（7）在该镜像里封装数据盒访问监控机制，实施对数据盒中数据访问的控制

例如，数据使用者声明只对数据进行聚类分析，但在使用数据过程中，如果数据使用者还做了其他数据访问操作，不正当使用数据时，那么封装在数据盒中的访问监控程序将报警，并阻止非法访问操作，甚至执行数据盒自毁操作。

（8）数据盒灌装完成后，外部数据使用者可以访问和使用该数据盒

数据使用者提出对数据盒的使用需求

以及使用方式(例如聚类分析、逐条读取、只利用其中的某些字段等),数据拥有者根据这些需求对数据盒进行定价,通过数据站为数据使用者提供数据盒服务。数据使用者可以利用Docker在其本地启动这个镜像,并通过提供的接口获取数据,即获取数据盒,实现数据开放;同时,数据盒中封装的访问控制和防泄露机制又能保证数据使用者无法直接接触到数据,保证数据在开放时的安全性。

6 结束语

本文提出的面向数据自治开放的数据盒模型,将数据拥有者的数据、访问权限等进行了有机合理的组织,使得数据拥有者可以根据自己对数据的开放程度进行把控。数据使用者也可以提出自己对数据的访问需求和程序,让数据盒拥有者进行审核,实现数据盒可编程。这种数据资源组织形式提升了数据的独立性、可用性和可控性,实现了数据自治,使得数据拥有者在数据开放的同时又能保证其数据稀缺性不丧失和隐私不泄露,有效地支持了数据开放。

参考文献:

[1] MINSTER B, CAMPBELL J W, DOZIER J, et al. Earth observations from space: the first 50 years of scientific achievements[C]// AGU Fall Meeting, December 10-14, 2007, San Francisco, USA. Washington DC: National Academies Press, 2008: 1-8.

[2] MAGEE A F, MAY M R, MOORE B R. The dawn of open access to phylogenetic data[J]. PloS One, 2014, 9(10): e110268.

[3] KAUPPINEN T, DE ESPINDOLA G M. Linked open science-communicating, sharing and evaluating data, methods and results for executable papers[J]. Procedia Computer Science, 2011, 4(4): 726-731.

[4] 朱扬勇, 熊赟. 数据资源保护与开发利用[M]. 上海: 上海科技文献出版社, 2008: 133-137.
ZHU Y Y, XIONG Y. Protection and utilization of data resources[M]. Shanghai: Shanghai Scientific & Technical Publishers, 2008: 133-137.

[5] CODD E F. A relational model of data for large shared data banks[J]. Communications of the ACM, 1970, 13(6): 377-387.

[6] JEFFREY U. First course in database systems[M]. New Jersey: Prentice-Hall Inc, 1997.

[7] 郑磊. 开放政府数据研究:概念辨析、关键因素及其互动关系[J]. 中国行政管理, 2015(11): 13-18.
ZHENG L. Study on open government data: definitions, factors and interactions [J]. Chinese Public Administration, 2015(11): 13-18.

[8] 郑磊. 开放政府数据的价值创造机理:生态系统的视角[J]. 电子政务, 2015(7): 2-7.
ZHENG L. Value creation mechanism of open government data: from ecosystem perspective[J]. e-Government, 2015(7): 2-7.

[9] GENTRY C. Fully homomorphic encryption using ideal lattices[C]//STOC, May 30-June 2, 2009, Bethesda, USA. New York: ACM Press, 2009(4): 169-178.

[10] SAHAI A, WATERS B. Fuzzy identity-based encryption[M]. Heidelberg: Springer, 2005: 457-473.

[11] DWORK C. Differential privacy[C]// The 33rd International Conference on Automata, Languages and Programming. July 10-14, 2006, Venice, Italy. Heidelberg: Springer, 2006: 1-12.

[12] CARMINATI B, FERRARI E, MORASCA S, et al. A probability-based approach to modeling the risk of unauthorized propagation of information in on-line social networks[C]//

The 1st ACM Conference on Data and Application Security and Privacy, February 21-23, 2011, San Antonio, USA. New York: ACM Press, 2011: 51-62.

[13] GORLA A, TAVECCHIA I, GROSS F, et al. Checking App behavior against app descriptions[C]//The 36th International Conference on Software Engineering, May 31-June 7, 2014, Hyderabad, India. New York: ACM Press, 2014: 1025-1035.

作者简介

朱扬勇(1963-),男,博士,复旦大学计算机科学技术学院教授、学术委员会主任,上海市数据科学重点实验室主任。1989年起从事数据领域研究,2008年提出数据资源保护和利用,2009年发表了数据科学论文“Data explosion, data nature and dataology”,并出版专著《数据学》,对数据科学进行了系统探讨和描述。2010年创办了“International Workshop on Dataology and Data Science”,2014年和石勇、张成奇共同创办了“International Conference on Data Science”。担任第462次香山科学会议“数据科学与大数据的理论问题探索”的执行主席、“大数据技术与应用丛书”主编。目前主要研究方向为数据科学、大数据。

熊赟(1980-),女,博士,复旦大学计算机科学技术学院教授。2004年起从事数据领域方面的研究工作,作为项目负责人主持国家自然科学基金、上海市科学技术委员会发展基金以及企业合作项目。相关研究成果在国际权威期刊和会议发表论文40余篇、出版著作3本。目前主要研究方向为数据科学和大数据。

收稿日期: 2018-02-01

基金项目: 国家自然科学基金资助项目(No. 91546105, No. U1636207);上海市科技发展基金资助项目(No. 16JC1400801)

Foundation Items: The National Natural Science Foundation of China (No. 91546105, No. U1636207), Shanghai Science and Technology Development Fund (No. 16JC1400801)