数据产品在线定制平台的探索实践

张峰[1]，孙宗哲[2]，OCHORA Dennis Reagan[2]，刘建楠[3]，宋杰[2]
1. 国家海洋信息中心，天津 300171；
2. 东北大学软件学院，辽宁 沈阳 110819；
3. 中国石油庆阳石化公司，甘肃 庆阳 745115

摘要
大数据时代，研究机构与企事业单位拥有海量的科学或产业数据（包括海洋、气象、地质、石油化工等行业的数据），可为客户提供分析后的数据产品。目前，这些机构尚未形成服务化的数据产品提供方式。分析了现有数据产品平台的缺陷与挑战，提出了数据产品在线定制平台的需求，设计了平台的体系结构，并对研究平台的服务化数据分析算法接口、数据安全性和隔离性保障、数据产品定价模型进行了详细探索实践，最后给出了在线定制平台的应用示例描述。

关键词
云计算；大数据；数据分析算法接口；数据即服务；数据产品定价

中图分类号：TP301.41　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2016072

Research and practice on online data production platform

ZHANG Feng[1], SUN Zongzhe[2], OCHORA Dennis Reagan[2], LIU Jiannan[3], SONG Jie[2]
1. National Marine Data and Information Service, Tianjin 300171, China
2. Software College, Northeastern University, Shenyang 110819, China
3. Petro China Qingyang Petrochemical Company, Qingyang 745115, China

Abstract
In the big data era, the research institutes and enterprises manage massive scientific data of various disciplines, such as oceanography, meteorology, geology and petroleum, and service the clients with the analysis results that are treated as data products. Traditionally, these institutes do not provide the service-oriented production platform for data products. The shortages and challenges of the production process were analyzed, the new requirements of online production platform for scientific data product was proposed, the architecture of platform was designed, and key techniques including the service-oriented interface for data analysis algorithms, mechanisms of data security and isolation, and the pricing model for data product were studied. Finally, the application of online production platform as a case study was explained.

Key words
cloud computing, big data, interface for data analysis algorithm, data as a service, data products pricing

1 引言

海洋、地质、气象、石油化工等研究部门和企事业单位在采集、管理、处理和分析海量数据的同时，也向外界提供了数据产品的定制服务，将此类数据加工成完整产品并发布、共享给客户的服务称之为数据服务。然而笔者认为，目前数据产品的制作、发布和入库机制尚存在以下问题：产品制作过程以及产品本身均无法复用；产品制作尚未服务化；无法实现高度可定制的产品制作方法；产品制作效率不高；缺乏完备的、基于计算平台的产品制作和发布流程。本文从数据产品的在线制作与发布技术展开，基于目前主流的云计算技术，设计了一种集中式、服务化、高效、产品可定制且安全的数据产品在线定制平台。

本文首先提出了数据产品在线定制平台的体系结构，然后从平台的关键技术入手，设计了用于支撑平台定制服务的数据分析算法接口和表述性状态传递（representational state transfer，REST）风格的算法描述方式、安全隔离与数据交换系统以及定价模型。最终以海洋环境基础数据中的水文数据为例，研发并测试能够实现关键研究成果的数据产品在线定制平台原型，达到一定的示范性。

2 研究动机

数据体系结构大体可分为4层，从下到上依次为原始数据层、基础数据层、数据集成层、产品库。产品库存放着由各个数据层产生、只读、用户定制的数据产品，如基础数据库的查询结果以及数据分析后产生的相关结果数据集或图形图标等[1]。然而，就目前的认知，笔者认为数据产品的制作、发布和入库机制尚存在以下问题。

一是产品制作过程以及产品本身均无法复用的问题。前期探索发现，目前尚缺乏一种集中式产品定制平台。产品制作是分散、手工化的，有很多产品名称不同，但其数据相同或相似，这些产品会视为不同的产品被不同的制作人重复地制作和保存。有些产品已经制作过或可以通过现有产品的再次加工得到，但因为缺少一个在线、共享的定制平台，无法得知该产品的制作信息，因此只能重新制作。此外，由于产品以一种手工或半自动化的方式制作，因此一些复杂产品的制作流程以及产品制作时产生的知识和数据，都无法统一管理和复用。

二是产品定制的服务化问题。前期探索发现，目前尚无面向用户提供在线产品制作的服务。对于服务产品，用户需要通过邮件、表单提交、电话或实地到访的方式提出自己的产品制作需求，获得自己需要的数据产品，并支付费用。用户尚无法在线使用产品制作服务。

三是无法实现高度可定制的产品制作[2]。前期探索发现，目前尚无可定制的产品制作过程。用户更多的是选择现有的常规产品，而无法定义新的产品需求。一些简单的需求则只能通过文字或口头描述，由产品制作人员理解后加以实现。“产品在线制作技术”和“商品在线销售技术”不同，前者需要用户能够灵活地在线定义产品制作过程，后者则仅仅需要商品信息的发布和用户数据采集。

四是产品制作效率不高。前期探索发现，目前部分数据产品制作效率低，很多产品的计算机制作时间超过20 min甚至更长，算法执行效率低，导致在线产品制作流程产生瓶颈。只有提高产品制作效率，将复杂产品的制作过程简化或分割为较小的产品，充分利用现有的计算结果避免重

复计算，采用更高效的算法或计算平台，才能成功实现在线产品制作。

五是缺乏完备的、基于计算平台的产品制作和发布流程。类比工作流的定义，即使存在一个在线产品制作平台，仍然需要定义完备的产品制作和发布流程，使“多个参与者之间按照某种预定义的规则传递文档、信息或任务”的流程可以自动进行，从而实现预期制作目标，或者促使此目标的实现。需定义整个流程的关键节点、节点前置条件和后置条件，并研究用户授权、付费方式等关键问题。

因此，本文提出的数据产品在线定制平台，将重点关注和解决以下问题：满足用户高度可定制的在线制作产品要求，通过多种算法组合完成需求；提供直观的产品价格展示；提高产品制作效率和产品利用率。

3 相关工作

国内的海洋、气象、地质、石油化工等科研和企事业部门经过几十年的发展，已积累了大量可用的数据，在提供数据共享、产品定制服务方面已有不少研究。回顾前人工作，数据共享方面已有大量先例，如南海海洋科学数据库①、国家林业科学数据平台②、国家地震科学数据共享中心③等，以上网站均在一定科学领域内提供相关数据及其数据产品的在线共享，但并未提供产品的在线定制功能。在数据的定价方面，大部分数据共享网站的数据均为无偿共享，无偿共享一方面能降低数据的共享门槛，另一方面却会增加平台的负载。简单的数据共享可以采用无偿共享的方式，但本文平台共享的是数据产品的定制服务而非数据，产品的定制需要耗费大量的计算资源，更适合有偿共享，也更符合数据即服务的概念，所以设计一个计算服务价格的定价模型是很有必要的。Youseff L等人[3]提出了定价模型的三大形式：每单位定价（per-unit pricing）、分级定价（tiered pricing）和预订定价（subscription-based pricing），并指出任何定价模型都至少使用了其中一种形式。本文采用的以数据价值为主导的定价模型属于每单位定价，但与之不同的是本文的定价单位并非只是常规意义上的计算资源，更多的是平台采集的海量数据。

在算法接口的设计上，REST架构是描述算法、传递消息的主流选择。REST风格的架构具有更高的可伸缩性和更低的开发复杂度[4]，REST架构是基于超文本传输协议（hypertext transfer protocol，HTTP）的，任何对资源的操作行为都是通过HTTP来实现，通过通用的链接器接口对资源进行操作，对统一资源标识符（uniform resource identifier，URI）的操作限制在4个方法内：get、post、put、delete，对应资源的增加、读取、更新、删除（create，read，update，delete，CRUD）操作[5]。本文采用的正是REST架构的接口设计，用户所需数据被转换成了数据服务，减少了开发难度。

4 体系结构

本平台的体系结构如**图1**所示，水平方向体现用户请求在平台内层次递进的过程，主要包含4个组件，分别是Web模块、外部处理模块、内部处理模块、存储模块。各组件内部以功能分层，在垂直方向上加以区分。

Web模块位于用户访问的前端，它负责与用户进行交互以完成数据收集与交换工作。Web模块采用B/S结构，支持数据转发，即将用户提交的信息转发到外部处理

① http://www.ocdb.csdb.cn/
② http://www.cfsdc.org/
③ http://westdc.westgis.ac.cn/

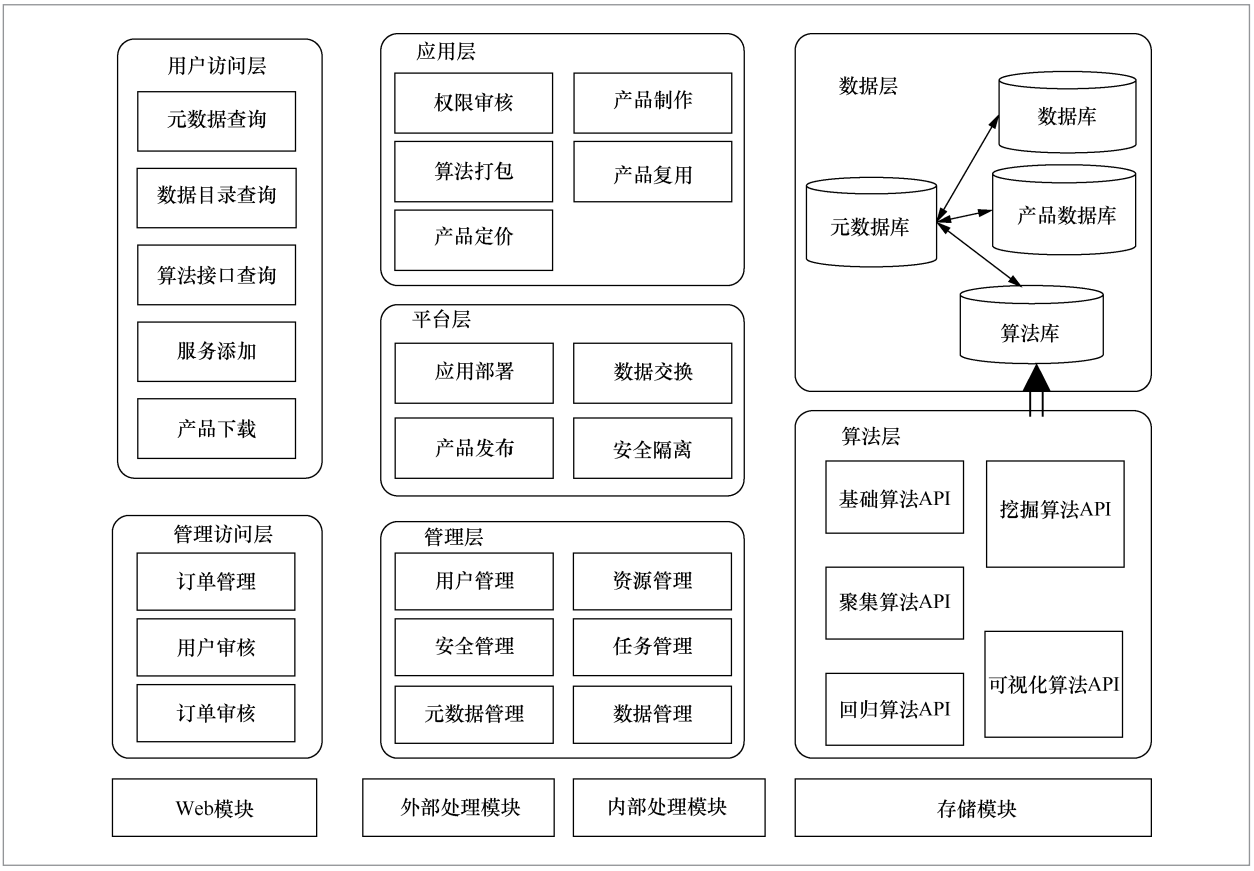


图1 软件体系结构

平台处理。在功能结构上分为用户访问层和管理访问层，用户访问层满足用户申请服务的需求，管理访问层满足管理员和外部服务平台管理层交互的需求。

外部处理模块面向外网，用于处理非隐私数据，数据处理量较小，主要完成数据交换和信息处理。其分类处理的数据和信息主要从Web平台和内部处理平台获取。

内部处理模块面向内网，是包含着大量主机的服务器集群，工作环境较为安全，用于部署应用程序，处理涉及隐私数据的应用请求，其主要完成数据产品和安全性的相关工作。

外部处理模块和内部处理模块在功能结构上一致，故功能分层统一分为应用层、平台层和管理层。应用层负责应用执行、数据计算等，平台层负责应用部署、数据交换、安全隔离控制等，管理层负责用户管理、数据管理、资源控制等。

存储模块用于存储所有数据和算法，一般分为两个部分：一部分用于存储外部可访问到的元数据，采用关系型数据库，使用HTTP访问，作为外部存储平台；另一部分存储隐私的数据、数据产品和算法，多为非结构化或半结构化数据[6]，同时使用关系型和非关系型数据库，与外部网络隔绝开，作为内部存储平台。

存储模块分为数据层和算法层，数据层存储原始数据和元数据，算法层存储制作产品所需算法API。元数据用于描述数据，如内容、结构、来源、质量和访问方法，面向外网的元数据库存储了基础数据、数据产品、算法的核心元数据，也是用户唯一能直接访问的数据库。用户依靠核

心元数据库获得算法类型、算法的可用数据范围、算法接口的引用形式、数据范围、数据的可用算法范围，面向内网的元数据库存储了上述数据的完整元数据。

系统体系结构如**图2**所示，自底向上分别为物理层、数据层、平台服务层、应用层。

物理层中，内网服务器的运算服务器通过交换机相连，形成一个含多个主机的计算集群，用于处理Web服务器传入的运算请求，同时完成对集群的管理和控制，如资源调度和任务调度等。内网数据服务器保存完整元数据、基本数据、算法、可复用的数据产品等。

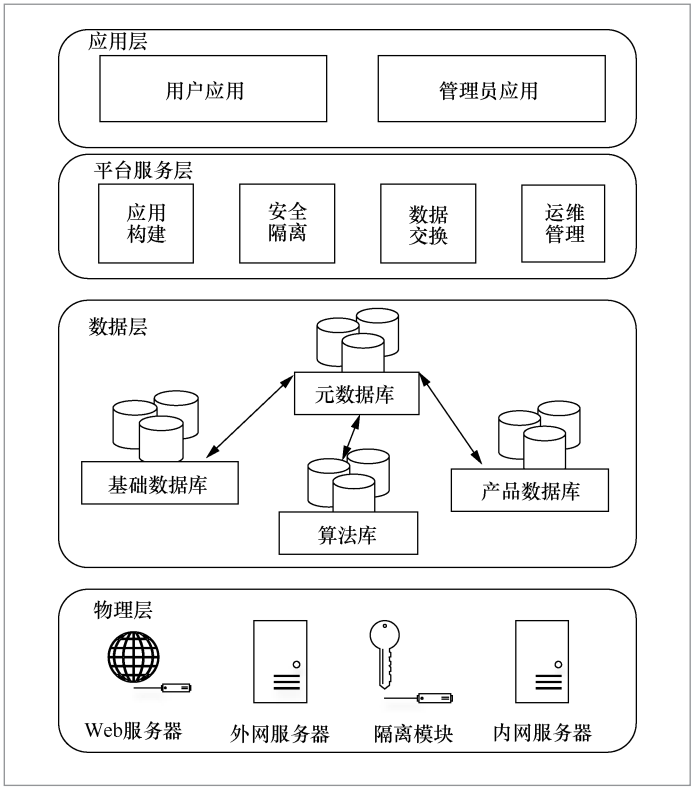


图 2 系统体系结构

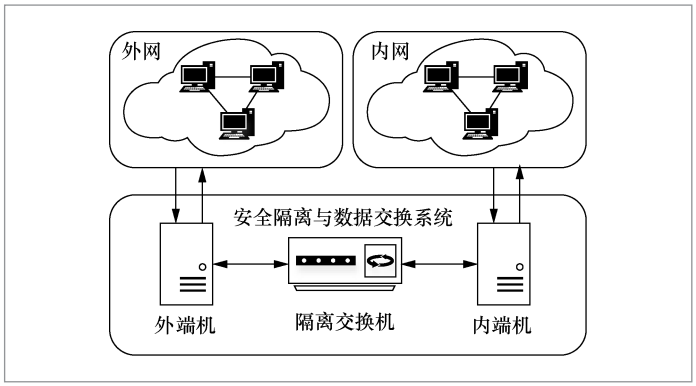


图 3 安全隔离和数据交换系统硬件结构

Web服务器使用Apache搭建完整的网站，搭建用户与内网服务器之间的桥梁，用户可以通过Web进行交互，用户完成产品制作算法的设计并将其上传到Web服务器上，Web服务器转发算法到内网服务器中进行部署和执行，并将执行结果返回给用户。外网数据服务器负责存储算法和基础数据的核心元数据，核心元数据是使用频率最高、用户可以直接访问的元数据，是完整元数据的子集。隔离模块是实现系统安全隔离和信息交换[7]的硬件基础，是由两台主机和一台隔离交换机组成的，硬件结构如**图3**所示。

数据层中分布了数个数据库，包括元数据库、基础数据库、产品数据库以及算法库，除核心元数据库可直接与外网相连外，其他数据库均处于内网。

平台服务层中，除了集成了操作系统、数据库、中间件等平台软件以外，还集成了物理隔离网间技术（gap technology，GAP）系统所需的软件，实现安全隔离前提下数据的双向交换。

应用层提供了多种算法API，通过API构建完整的产品制作算法，向普通用户以及管理员提供搜索、查询等服务。

4.1 核心模块

在软件体系结构中，大部分功能模块的实现比较常规，除了一些约定俗成的功能模块，需要对剩下的核心模块进行解释。

数据目录检索是用户选择数据的窗口，此目录的检索内容来自于基础数据和

算法API的元数据（算法API可视为另一种形式的数据）。

元数据检索是平台极为重要的功能模块，元数据是描述数据模型的数据，它是关于数据模型的基本概念、基本关系、基本约束的语义，其内容包括：数据描述，数据来源、数据所有者及数据序列（数据生产历史）等的说明，数据质量描述，数据分析信息说明，数据转换方法描述等[8]。

通过元数据查询检索系统，用户可以知道数据的产生背景、数据质量、数据格式、数据量大小、数据单价等信息，从而决定是否需要使用该数据制作数据产品。

数据产品存储及复用功能将根据用户制作的数据产品及其相关信息自动生成相应数据产品元数据，例如数据范围、数据量、产品制作算法等，产品可复用的算法范围可以根据产品的制作算法得出，产品标签则由上传者或管理员手动添加。

4.2 控制流和数据流

平台的数据流展示了平台数据流动的方向，控制流说明了平台及用户前后台交互的步骤。数据流动较为复杂，具体的数据流动如**图4**所示。

平台的数据流动主要步骤如下。

步骤1 最基础的数据存储在基础数据库、算法库、产品数据库中，抽取这些数据库的所有元数据形成一个完整的元数据。

步骤2 完整元数据库中抽取核心元数据形成一个核心元数据库，供用户查询数据目录、元数据和算法接口。

步骤3 用户提交算法后，首先从元数据库中提取产品元数据，验证当前算法是否有现成的产品可复用，并从基础数据库和算法库中提取所需数据，并打包成应用，进行产品制作，制作完成后提供下载。

步骤4 根据步骤3提取的算法和数据量，结合核心元数据中的算法及数据单价，通过定价模型计算出产品定价，并在产品完成后支付下载。

平台的控制流程主要分为三大部分：用户在前台提交算法、后台定价并制作产品、用户在前台付费并下载产品。具体的控制流程如**图5**所示。

5 关键技术

海量数据挖掘的关键问题是数据挖掘

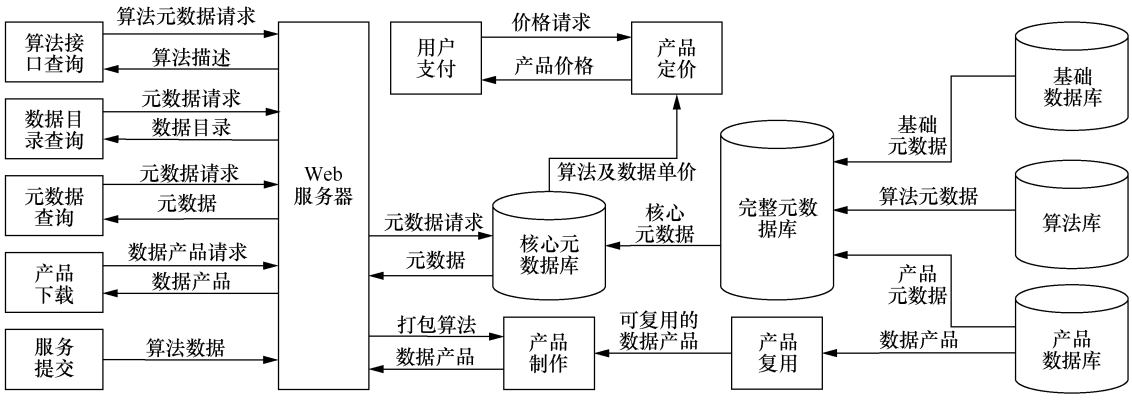


图 4 数据流程

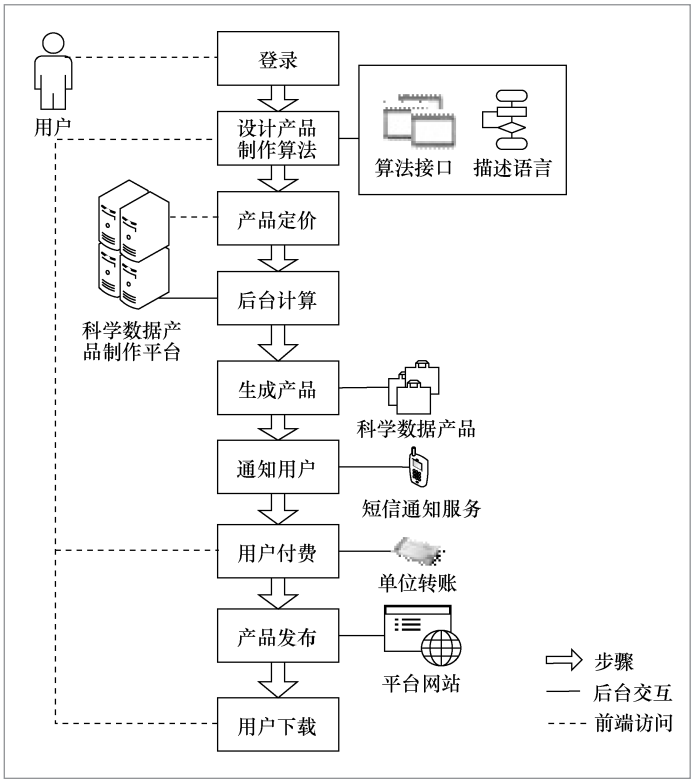


图 5 控制流程

算法的并行化。而云计算采用 MapReduce 等新型计算模型，这意味着现有的数据挖掘算法和并行化策略不能直接应用于云计算平台进行海量数据挖掘，需要进行一定的改造。

数据产品的在线制作对产品制作过程提出可定制性要求，用户可以根据数据分析算法库提供的轻量级接口，运用一种简单的过程描述语言，通过算法接口组装成用户自定义的制作过程。本文定义了一种数据分析算法接口模型和REST风格的算法描述方式。

数据安全性是产品制作平台的基石，保障数据的安全性意味着保证用户只能通过平台指定的方式进行数据访问，同时用户只能在平台指定的数据存储区内保存请求之间的数据，以保障用户定制的产品制作过程对数据安全性的要求。本文提出了一种在保证安全隔离的前提下，实现可控的数据交换的方法。

数据定制平台的商业价值不仅在于普通平台一段时间内独占的平台软硬件资源，还包括数据的价值以及产品制作过程赋予基础数据的附加价值。本文设计了一个具有针对性的定价模型。

5.1 数据分析算法接口和REST风格的算法描述

为了准确支持数据分析算法，本文设计了一种数据分析算法接口模型，使其能够统一地对产品定制平台提供的各类数据分析算法进行描述。主要作用有：为用户提供便利性和数据访问限制，用户能且仅能通过算法接口使用产品制作服务，进而访问数据；为服务方提供扩展性，服务方可以按接口扩展现有算法库，提供更多的产品生成算法。

数据分析算法接口是一系列自定义的面向过程的Hadoop算法接口，主要包含复杂查询、聚集、统计、回归、挖掘、可视化等算法种类，每种算法种类作为一个接口模块，用户只需设置简单参数即可使用提供的算法。同时，考虑到算法接口的动态扩展，在设计算法过程中，各个模块间耦合度小，方便后期的算法添加。以“查找（query）特定范围的数据，计算它们的方差（variance），如果方差小于给定值，则返回数据均值（mean），否则返回数据直方图（histogram）”为例，query、variance、mean和histogram均为算法库提供的轻量接口，参数为数据的范围。

为了解决数据的授权访问问题，笔者在设计算法时，给每个算法都设定了数据访问范围，在算法组合后，任一算法只可访问算法本身规定范围内的数据和在制作过程中产生的数据。

为了实现算法描述的简单、直观、高效，采用REST作为算法描述方式的规范，REST的设计理念为：不区分事物，统一抽象为资源；资源使用统一资源标识符（uniform resource identifier，URI）为计算机指引资源所代表的文档或对象的具体位置，将URI作为唯一标识；采用通用接口操作资源；URI不随资源的操作而改变；采用无状态通信。基于REST，本文提供的接口见**表1**。接口分为数据服务和算法服务，如何使用接口也将在**表1**中描述。

本平台提出的接口以URI形式提供给用户使用，用户可以直接通过HTTP请求与资源进行交互，也可以将其嵌入系统后进行二次开发，用户在系统窗口内输入URI进行交互。

以海洋环境基础数据中的水文数据为例，制作“2010—2014年渤海深度在100 m内的平均水温”的URI可描述为：/data/temperature/use/year/2010/2014/use/sea/bohai/use/depth/0/100/algorithm/average/use_data/1，其中不同data的多个type顺序可替换。

5.2 定价模型

考虑到数据产品制作成本的特殊性，即主要成本由采集数据产生，产品制作过程中消耗的计算资源可忽略不计，因此在本平台的定价模型中，将定价粒度（即定价的基本单位）定义为制作算法，这种设计避免了为了收集和计算占用资源，使得一个订单服务在一段时间内独享平台软硬件资源造成的资源浪费，还为用户提前得知定价以便修改定制算法提供了可能性。

以制作算法作为支付单位的定价算法描述如下。

假设产品定价为P，产品权值为R，第i个产品制作算法单价为p_i，采用制作算法个数为n，可得计算式：

$$P = R \times \sum_{i=1}^{n} p_i \qquad (1)$$

其中，产品权值R表示不一样的产品结果需要不一样的定价方法，例如，同样的算法和数据，获得一系列连续数值、连续数值的分布图像、连续数值的三段分布图像的定价必然是不同的。假设产品有n种形式的结果，第i种形式分别有n_i个结果，每个

表1 对外URI描述

URI	描述
ID	返回所有可使用的数据类型及其ID
/data/{type_ID}/	返回指定ID类型的数据集的核心元数据，包括数据集可操作范围和描述
/data/{type_ID}/use/{range_ID}/{range_head_ID}/…{range_tail_ID}/…	使用指定ID类型下指定范围的数据集作为算法分析的数据来源，可重复添加以使用不同数据
/algorithm	返回所有可用的算法接口及其ID
/algorithm/{algorithm_ID}	返回指定ID的算法接口描述，包括输入的数据类型和数目以及输出的数据类型和数目
/algorithm/{algorithm_ID}/…/use_data/{use_data_ID}/…/use_algorithm/{use_algorithom_ID}/…	使用指定ID的算法接口，输入数据为指定序号的数据集、指定序号的算法返回数据的集合，第i个URI标识为“data”的数据集序号即为i，第j个URI标识为“algorithm”的算法序号即为j

结果的权值为r_i(0≤r_i≤1)，可得计算式：

$$R=1+\sum_{i=1}^{n}(n_i\times r_i) \quad (2)$$

此外，假设第i个算法基础价格为a_i，该算法采用数据条数为n_i，其中第j条数据的数据单价为d_{ij}、第j条元数据价格为d_{ij}'，第i个产品制作算法单价p_i的计算式如下：

$$p_i=a_i\times\sum_{j=1}^{n}(d_{ij}+d'_{ij}) \quad (3)$$

其中，第j条数据的数据单价d_{ij}需要在数据导入的同时给定值，第j条元数据价格d_{ij}'是考虑到有部分元数据可能单独或和数据联合起来进行运算而设计的，需要在元数据导入同时给定，将一个已有的可复用的数据产品也视作一条数据。

6 应用实例

笔者开发了国家海洋局的海洋信息产品在线制作与发布系统，系统实现了元数据的检索、数据产品的定制、数据产品的浏览与获取等功能。

用户编写符合数据分析算法接口标准的算法语句，通过网站窗口提交到网站后台，网站后台在审核数据调用权限后，传递到数据分析平台进行相关计算，生成产品。**图6**为示例算法“2010—2014年渤海深度在100 m内的平均水温”的URI的解析，解析时首先识别标识符，然后识别标识符后跟随的ID，用来限定数据范围、选择分析算法。

用户在订单管理页面可以有效管理订单。订单号由订单生成时间及订单内容组成，订单价格在用户提交算法后由服务器根据定价算法得出，订单状态实时更新。用户支付后，点击下载即可下载产品，产品根据用户提交的算法不同，可分为单个数值、数据集、图像集。数据产品在线定制平台目前在国家海洋信息中心相关部门初步应用，**表2**总结了应用前后的效果对比。可见，平台能够完成设计目标，但产品定制效果有待提高。最终，笔者预期可以实现部分产品生成算法的定制。

7 结束语

本文建立了数据产品在线定制平台的体系结构和流程，实现了对数据的产品定制，完成了算法接口和定价模型的设计和原型系统示例，对于完善“数字海洋”系统的数据管理与共享、建立完备的数据服务体系起到一定作用。

进一步的工作有：首先，定价模型需要根据进一步实验给出数据之间、算法之间的基础比值，以便为实现具体模型时提供定价参考；其次，用户申请数据的过程需

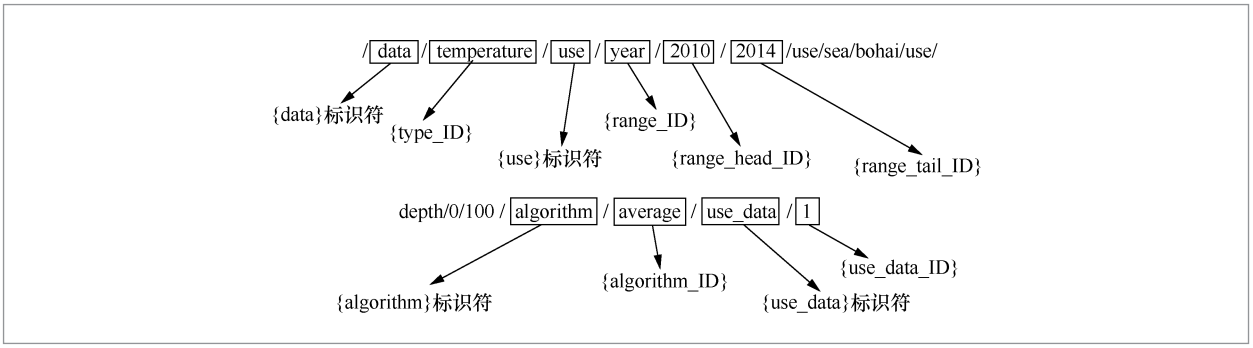


图 6 URI 示例解析

表 2 应用效果比较

应用前	应用后	效果
产品名称不同但数据相同或相似，这些产品会被视为不同的产品被不同的制作人重复地制作和保存	产品或复用或可以通过现有产品的再次加工得到	产品可复用
产品制作方法的分享只能依靠口头阐述或文档规范	产品制作时产生的知识和数据，如算法和流程，都可以统一地管理和复用	产品制作过程可复用
用户更多的是选择现有的常规产品。新需求则只能通过文字或口头描述，由产品制作人员理解后加以实现	大部分产品可以由用户通过选择算法来定制	产品可定制
由于自动化程度不高导致产品制作效率低	减少了数据准备、算法实现、流程控制等阶段的人工参与，制作效率明显提高	产品制作效率提高
很多大产品都需要从零做起，算法执行时间长	化整为零，大产品由小产品组成，小产品可以复用现有产品或重新制作	产品制作效率提高
工作流程基本是一种约定俗成的现行做法，关键步骤有相关条例约束	整个产品生成流程均制度化，关键步骤更为丰富，且原来零散的条例更整体化	流程更为规范

引入合适的权限控制系统以控制资源的授权，避免因此带来的安全问题；最后，需要提升数据分析算法接口定制复杂分析算法的能力，以适应用户对于“高度可定制”的产品制作需求。

参考文献:

[1] 张峰, 石绥祥, 殷汝广, 等. 数字海洋中数据体系结构研究[J]. 海洋通报, 2009, 28(4): 1-8. ZHANG F, SHI S X, YIN R G, et al. A study of data architecture in digital ocean[J]. Marine Science Bulletin, 2009, 28(4): 1-8.

[2] ALBERTO P. Technologies for large data management in scientific computing[J]. International Journal of Modern Physics C, Physics and Computers, 2014, 25(2): 343-352.

[3] YOUSEFF L, BUTRICO M, SILVA D D. Toward a unified ontology of cloud computing[C]//The Grid Computing Environments Workshop, Nov. 12-16, 2008, Texas, USA. New Jersey: IEEE Press, 2008: 1-10.

[4] AZEEZ A, PERERA S, GAMAGE D, et al. Multi-tenant SOA middleware for cloud computing[C]//The 2010 IEEE 3rd International Conference on Cloud Computing, July 5-10, Miami, Florida, USA. New Jersey: IEEE Press, 2010: 458-465.

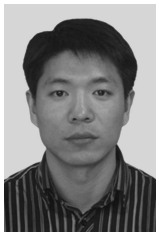
[5] FENG X, SHEN J, FAN Y. REST: an alternative to RPC for web services architecture[C]//The Future Information Networks, October 14-17, 2009, Beijing, China. New Jersey: IEEE Press, 2009: 7-10.

[6] LOMOTEY R K, DETERS R. Analytics-as-a-service (AaaS) tool for unstructured data mining[C]// 2014 IEEE International Conference on the Cloud Engineering (IC2E), March 11-14, 2014, Boston, MA, USA. New Jersey: IEEE Press, 2014: 319-324.

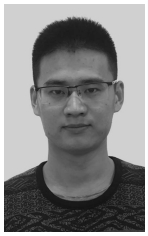
[7] KULKARNI G, GAMBHIR J, PATIL T, et al. A security aspects in cloud computing[C]// 2012 IEEE 3rd International Conference on the Software Engineering and Service Science (ICSESS), June 22-24, 2012, Beijing, China. New Jersey: IEEE Press, 2012: 547-550.

[8] CAZEMIER H, RASMUSSEN G D. Query engine and method for querying data using metadata model: U.S. Patent 6,609,123[P]. 2003-8-19.

作者简介



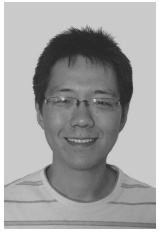
张峰（1978-），男，博士，国家海洋信息中心副研究员，主要研究方向为云计算与数据服务。



孙宗哲（1991-），男，东北大学软件学院硕士生，主要研究方向为高能效计算。

Ochora Dennis Reagan（1990-），男，东北大学软件学院硕士生，主要研究方向为大数据处理。

刘建楠（1963-），男，就职于中国石油庆阳石化公司，主要从事企业经营和信息化管理工作。



宋杰（1980-），男，博士，东北大学软件学院副教授，中国计算机学会高级会员，主要研究方向为大数据存储与管理、高能效计算、云计算。

收稿日期： 2016-10-20

基金项目： 国家自然科学基金资助项目（No.61433008，No.61502090）；数字海洋开放基金资助项目（No.KLD0201405）；教育部博士点基金资助项目（No.20130042120006）；教育部—英特尔信息技术专项科研基金资助项目（No.MOE-INTEL-2012-06）

Foundation Items: The National Natural Science Foundation of China(No.61433008, No.61502090), Digital Ocean Open Foundation(No.KLD0201405), Doctoral Fund of Ministry of Education of China (No.20130042120006), Ministry of Education - Intel Information Technology Special Research Fund(No. MOE-INTEL-2012-06)