ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2012/06(05)-0385-12
DOI: 10.3778/j.issn.1673-9418.2012.05.001

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-51616056

云计算中面向隐私保护的查询处理技术研究*

霍　峥[1+],孟小峰[1],徐建良[2]
1. 中国人民大学 信息学院,北京 100872
2. 香港浸会大学 计算机系,香港

Privacy-Preserving Query Processing in Cloud Computing*

HUO Zheng[1+], MENG Xiaofeng[1], XU Jianliang[2]
1. School of Information, Renmin University of China, Beijing 100872, China
2. Department of Computer Science, Hong Kong Baptist University, Hong Kong, China
+ Corresponding author: E-mail: huozheng123@gmail.com

HUO Zheng, MENG Xiaofeng, XU Jianliang. Privacy-preserving query processing in cloud computing. Journal of Frontiers of Computer Science and Technology, 2012, 6(5): 385-396.

Abstract: A vital concern in cloud computing is how to protect both data privacy and query privacy while providing query services for users. This paper surveys several critical techniques of privacy-preserving query processing in cloud computing, which include cloud database indexing, query optimization, encryption-based privacy-preserving techniques, privacy-preserving techniques based on secure multi-party computation and authorization auditing techniques. Finally, the paper analyzes the challenges of privacy-preserving query processing in cloud computing and figures out the trend of this area.

Key words: cloud computing; query processing; privacy-preserving; index

摘　要: 在云计算环境中既能同时保护数据隐私和用户查询隐私,又能提供给用户满足需求的查询结果是云计算中面向隐私保护的查询处理的关键问题。对云计算中面向隐私保护的查询处理技术的若干关键问题进行了全面的调研,包括数据库索引技术与查询优化、基于加密的隐私保护技术、基于安全多方计算的隐私保护技术以及查询结果完整性验证技术。分析了云计算中面向隐私保护的查询处理技术的挑战性问题,指明了未

* The National Natural Science Foundation of China under Grant Nos. 60833005, 61070055, 91024032 (国家自然科学基金); the National Science and Technology Major Special Project of China under Grant No. 2010ZX01042-002-003 (国家科技重大专项"核高基"项目); the Research Funds of Renmin University of China under Grant No. 10XNI018 (中国人民大学科学研究基金).
Received 2012-01, Accepted 2012-03.

来研究方向。

关键词:云计算;查询处理;隐私保护;索引
文献标识码:A **中图分类号**:TP391

1 引言

随着信息产业的发展,企业和政府机构产生的数据量快速增长,如何管理和分析海量数据是目前医疗、通信、交通及互联网等很多领域面临的问题。传统的数据管理系统对于如此大规模的数据管理已不再有效,即便它们能够管理大规模数据,但所花费的相关软硬件以及维护成本让大部分企业望洋兴叹。自从2006年谷歌公司推出BigTable以来,云计算概念呈现在大众面前。作为云计算基础的云数据库系统是由大量性能普通、价格便宜的计算节点组成的一种无共享大规模并行处理环境,它克服了管理海量数据成本过高的缺点。另外,云数据库系统结合了网络化和虚拟化技术来实现超级计算和存储能力,具有高可靠性、高扩展性、通用性、按需分配等优点。

近几年,云计算已经成为当今信息产业最受关注的一种全新的计算模式。据IDC研究机构预计,2013年云计算服务的市场将达到146亿美元的规模,而其中数据市场有39.8亿美元,约占整个数据管理系统软件市场的27%,并且以每年10.3%的速度增长[1]。不可否认,云计算已成为未来海量数据管理的必然趋势。从成本和性能两方面考虑,越来越多的企业愿意把自己的数据中心从昂贵的高性能计算系统转移到公有的云平台或私有的云平台上。然而,企业以及个人用户在使用云计算带来的便利的同时,也面临着隐私泄露的风险。研究者们认为,云计算中的查询和数据隐私保护已经成为云计算研究中的关键问题之一[2]。

本文对云计算中面向隐私保护的查询处理技术进行了综述。第2章介绍云计算中面向隐私保护的查询处理技术中的基本概念及需要解决的关键问题;第3章分析云数据库系统中的关键问题;第4章介绍三种主流的查询隐私保护技术;最后展望未来研究工作。

2 基本概念与关键问题

本章主要介绍云计算中面向隐私保护的查询处理技术中的基本概念以及需要解决的关键问题。

2.1 基本概念

云计算中一般有三个角色:数据拥有者、查询用户和云计算平台。数据拥有者将数据提供给云计算平台进行存储,查询用户通过云计算平台提供的查询接口对数据进行查询。三方参与的云计算模型如图1所示[3]。

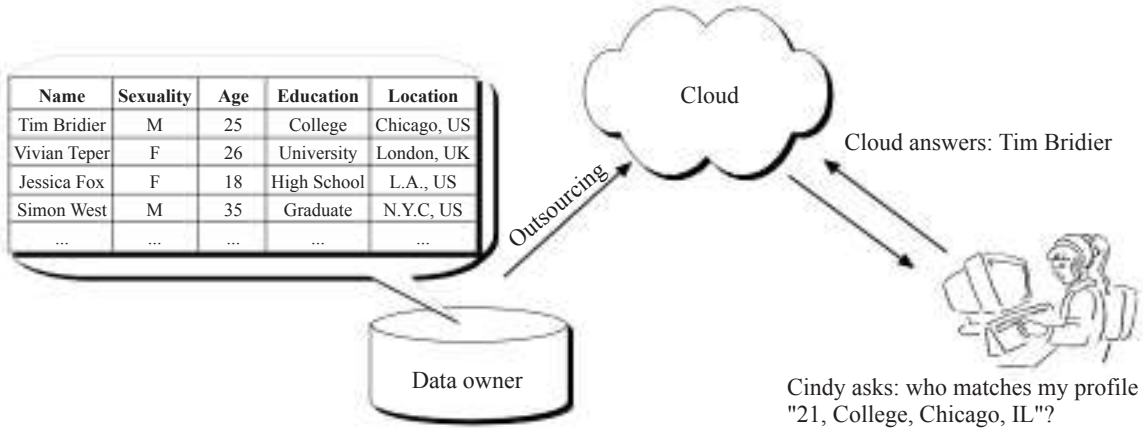

Fig.1 The architecture of cloud computing

图1 云计算模型

假设某查询用户Cindy打算在某社交网站(数据拥有者)上查找和她有相同背景(比如年龄、教育背景、所在城市)的人交朋友。然而,在返回的查询结果中,除了目标用户的姓名及联系方式以外,网站不能将其他用户的年龄、教育背景、所在城市等个人信息透露给Cindy(数据隐私),用户Cindy也不愿意将自己的个人信息作为查询内容的一部分透露给网站(查询隐私)。这就需要在返回满足用户查询需求的情况下,同时保护网站本身的数据隐私和用户的查询隐私。在其他商业应用中也有类似的问题。由于某些查询可能泄露商业机密,数据隐私或查询隐私的泄露可能带来更加严重的后果。例如,某零售商计划在某个区域开设分店,需要事先评估该区域的目标客户群。假设该区域的人口数据信息都存储在云数据库系统中,该零售商需要查询该区域的人口统计数据。人口信息的个体信息不能暴露给云计算平台或该零售商(数据隐私)。同时,该区域的名称也不能暴露给云计算平台(查询隐私),否则会造成该零售商的商业计划泄露。此外,由于云计算平台可以同时为多个数据拥有者存储数据,使得隐私泄露的情况更加严重。例如,某用户发出两个查询,一个查询本地的药店地址,另一个查询治疗糖尿病的药品,攻击者可以结合两个查询推断出该用户可能患有糖尿病。

综上所述,云计算中面向查询处理的隐私保护技术主要关注以下两个方面。

2.1.1 用户的查询隐私

一般来说,隐私是指个人或组织不愿意被外界知晓的信息。在云计算环境中,查询用户通过向云计算平台发出查询来获取服务。然而,用户提交的查询有可能暴露用户的个人隐私。在上述例子中,用户Cindy如果将自己的个人资料作为查询内容提交给云数据库系统,则可能暴露其个人隐私,如个人信息、兴趣爱好、教育背景等。用户在享受查询服务的同时,更希望自己的查询隐私能得到保护。用户的查询隐私保护是指通过采用隐私保护技术,使云数据库系统和数据拥有者不能获知用户的查询内容,也不能通过用户的查询推导出关于用户的任何信息。

2.1.2 数据拥有者的数据隐私

在云计算环境中,数据拥有者将自己持有的数据存储到云计算平台上,通过用户有偿地使用云计算提供的服务而获益。因此,数据拥有者的数据一方面不能暴露给云计算平台,另一方面也不能暴露给查询用户,也就是说,查询用户只能得到和查询相关的结果,不能额外获得任何与查询结果无关的数据,否则就损害了数据拥有者的利益。在上述例子中,用户Cindy只能得到和自己背景相近的人的姓名和联系方式,但不能获取他们具体的年龄、教育背景等信息。简单地说,数据拥有者的数据隐私保护是指通过隐私保护技术,防止查询内容以外的数据泄露给查询用户或者云计算平台。

综上所述,对云计算的各个参与方而言,面向隐私保护的查询处理都是迫切需要解决的问题:对查询用户而言,如果在查询处理中隐私保护机制不完善,用户由于担心查询隐私的泄露,将尽量减少使用云计算服务;对数据拥有者来说,若查询处理暴露其拥有的数据的隐私,不仅涉及商业利益的问题,而且还可能面临法律诉讼的风险;对于云数据库系统而言,如果用户隐私得不到保障,其服务的可靠性将会受到质疑。因此,迫切需要一种能在云计算中同时保护查询隐私和数据隐私的新型查询处理技术,以全面保护查询用户、数据拥有者和云数据库系统的隐私,云计算中面向隐私保护的查询处理技术的研究应运而生。

2.2 关键研究问题

云计算环境中有三个参与方:数据拥有者、云计算平台和查询用户。云数据库系统是云计算中用来存储数据的关键组件。近年来,云数据库系统引起了IT公司和研究机构的广泛关注。各大公司如微软、亚马逊、谷歌等也纷纷推出了自己的云数据库系统,云数据库系统是云计算中面向查询处理隐私保护技术的关键研究问题之一。因此,本文针对云数据库系统的实现结构、数据模型、容错性与可用性、一致性处理、索引设计和查询优化、隐私和安全等问题进行了调研,分析了国内外研究现状。

云计算中的隐私问题近年来才逐渐凸显,针对这方面的研究并不多见。云计算中面向查询处理的隐私保护需要借鉴外包数据库和分布式数据库中的隐私保护技术。因此,本文针对相关研究领域进行了综述,具体包括:基于加密的隐私保护策略、基于安全多方计算的隐私保护策略和查询结果的完整性验证等。

3 云数据库系统的实现

3.1 系统结构

从系统结构来看,目前的云数据管理系统有三种实现方式:基于MapReduce的系统结构、基于数据库管理系统(database management system,DBMS)的系统结构和将两者结合的系统结构。

基于MapReduce框架的系统具有良好的扩展性、容错性和并行性,但是对结构化查询语言(structured query language,SQL)的支持比较困难,大部分该类型的数据管理系统还不支持SQL语句,仅有Hive支持部分SQL语句。基于DBMS架构的系统能够充分利用DBMS中的查询优化策略,但是不利于对半结构化和非结构化数据的支持,且数据的冗余无法交给底层的DBMS实现,只能放到其上层来实现,增加了系统的复杂性。HadoopDB[4]利用MapReduce技术将单个节点上的DBMS联系起来,希望能够充分利用两者的优势,但是这种方式增加了SQL语句解析的工作,导致SQL语句的执行效率降低。另外,由于Hadoop的数据存放在DBMS中,而非分布式文件系统(Hadoop distributed file system,HDFS)上,其对数据冗余的处理也只能像基于DBMS结构的系统一样无法避免。

3.2 基于云计算的数据模型

目前云数据管理系统中存在两种数据模型:一种是来源于Google的BigTable所采用的<*key*, *value*>数据模型,这是大多数云数据管理系统采用的数据模型;另一种是简单的关系数据模型,数据被组织成表中的记录。<*key*, *value*>模型能够更有效地使用MapReduce编程框架,并且使用灵活,也有利于存储半结构化数据和非结构化数据。但是由于缺少像关系模型背后的坚实理论基础的支持,<*key*, *value*>模型对复杂查询的支持以及事务处理的支持显得力不从心。

关系数据模型在传统的关系数据库中得到广泛应用,但是直接放到云计算环境中又缺乏对扩展性的支持,另外关系数据模型不能较好地描述非结构化数据。

3.3 容错性与可用性

大多数云数据管理系统应用要求高可用性。即,所有的应用需要在错误出现的情况下能够继续进行数据读操作,而某些应用甚至要求在出错情况下能够继续进行数据写操作。云数据管理系统中的容错性与传统的关系数据库系统的容错性有所不同。对于事务处理过程来说,一个具有容错性的DBMS能够从错误中恢复到错误发生前的状态,且不丢失数据和修改;对于分布式数据库系统来讲,一个容错的系统能够在工作节点出错的情况下继续进行事务的提交操作。而对云数据管理系统,容错性不仅仅表现在出错情况下用户操作的正常进行,更体现在当参与执行用户查询的节点出现问题时,该数据查询操作不需要重新启动。云计算平台往往采用廉价、不可靠的主机搭建无共享集群,出错几率高于传统的分布式数据库中的高性能服务器。出错问题随着集群规模的增大显得尤为突出:查询涉及到的云数据规模越大,需要工作的节点就越多,而在查询中节点出问题的概率也会越大。文献[5]指出,Google公司平均每个分析任务都会遇到1.2个节点错误。如果每次查询出错时操作都需要重新启动,那么一个分析任务很有可能永远无法完成。

在基于MapReduce的系统中,保证查询的容错性是比较容易的。如果在Map任务或者Reduce任务过程中出现了问题,任务追踪机制可以检测到错误,并将相应的任务交给正常的服务器完成,这些操作对客户端是透明的。但是在传统的并行数据库中,错误被认为是一种极少发生的事件,所以错误发生时大部分并行数据库会重新进行整个查询操作。因此在基于DBMS架构的系统中,如果要满足云数据管理系统的这种容错性,必须构建一层中间件服务

来进行容错处理。

3.4 一致性处理

Brewer在2000年提出了著名的CAP理论，后人也论证了CAP理论的正确性。CAP理论指出：一个分布式系统不可能同时满足一致性（consistency）、可用性（availibility）和分区容忍性（partition tolerance）这三个需求，最多只能同时满足其中的两个。作为一种特殊的分布式数据管理系统，云数据管理系统可以提供强一致性或者某种形式的弱一致性。最强的一致性就是数据库理论中常说的ACID特性，即原子性（atomicity）、一致性（consistency）、隔离性（isolation）、持久性（durability）。ACID特性广泛应用在传统的关系数据库中。与之相对应的另一个极端就是基本可用的最终一致性。最终一致性通常表现在数据冗余系统中，最新版本的数据出现在某几个数据节点上，在一段时间内有些节点的数据是过期版本，但它们最终会更新成最新版本。不同的应用对于一致性的要求也不同，因此系统开发者们根据需求在一致性、高可用性和分区容忍性中进行权衡。类Big-Table系统是一种CA系统，即它们具有高一致性和高可用性，但是不具备分区容忍性。雅虎公司的PNUTS（platform for nimble universal table storage）则选择了另外两个特性：可用性和分区容忍性。文献[6]指出，PNUTS使用了一种介于可串行性和最终一致性之间的一致性模型——基于时间轴的一致性模型，保证每个节点上的数据均按照同一个时间顺序进行修改。使用这种模型，PNUTS支持一系列不同程度一致性要求的接口。

3.5 索引设计和查询优化

为了提高海量数据的查询性能，高效并且维护代价低的索引技术必不可少。目前国内外的研究者们针对云数据管理系统的索引技术和查询优化展开了研究。

（1）一维数据索引

Aguilera等人在文献[7]中首先提出了在云数据管理系统上建立分布式B树（binary tree）的思想。该分布式B树具有高可扩展性、低代价消耗、容错性好以及易于管理的特点。他们将B树的结点分布在一个局域网络中的多台服务器上。同时，Aguilera等人使用的是B+树，叶结点存放键-值对，内部结点只存放键-指针对。该分布式B+树除了支持传统的字典操作（如插入、查找、更新、删除）以及有序遍历以外，还支持两个传统方法不支持的实际应用特性：事务访问机制和在线迁移树结点。它可以完全透明地从一台服务器向另一台服务器或新增服务器迁移树结点，同时B树仍然继续提供服务请求。这个特性帮助在线管理系统能够持续进行线上操作。然而，该方法不但给服务器带来了很大的索引维护负担，而且给客户端机器造成了很大的内存负担，因为客户端需要复制查询对应的所有内部索引结点。因此尽可能选择一个索引更新维护代价较小的索引机制是一个实际应用中必须解决的关键性问题。Wu等人在文献[8-9]中正是认识到这一点，提出了代价估计模型作为索引维护和调优的参考模型。该索引框架主要由三层构成，中间层主要是成千上万的节点为用户提供计算服务。用户数据被划分为许多的数据块，并根据分布式文件系统协议分布存储在不同的节点上。每个节点为其存储的数据建立局部索引。除了局部索引以外，每个节点分享其局部的存储用来维护全局索引。

（2）多维数据索引

以上的索引机制主要针对一维数据的查询处理。实际应用中，查询更多的是针对多个属性的复杂查询。文献[10]首先提出了在云数据管理系统中建立多维索引的思想，针对云数据管理中主从结构的这种系统，采用了两级索引机制，在每个从节点上对数据建立*k*-d树（*k*-dimensional tree）索引，在主节点上对所有*k*-d树索引建立R树（region tree）索引。该方法主要针对主从结构的系统，对于基于分布式哈希表（distributed Hash table，DHT）的云数据管理系统，多维数据索引未作考虑。文献[9]针对以上问题，提出了一个基于DHT系统的多维索引机制，它在每个存储数据的节点上对数据建立R树索引，然后所有的索引服务器通过控制区域网（controller area network，CAN）连接起来，以便查询的分发。同时，为了减少查询代价和索引维护代价，提出了一种动态调优

算法,用来选择R树结点来建立CAN网络索引。

(3)查询优化

最近,针对云计算中的查询优化工作出现了新的研究成果。Nykiel等人在文献[11]中提出了一种名为MRShare的架构,该架构针对MapReduce任务提出了一种资源共享机制,用于减少MapReduce任务中的重复工作,以提高查询性能。Dittrich等人在文献[12]中提出了Hadoop++,它在Hadoop系统的基础上增加了新的索引和连接查询优化方法以提高查询性能。文献[13]提出了影响MapReduce性能的几个因素,包括I/O模型、调度策略和数据存取方式等,并在针对这些因素进行测试的基础上提出了提高MapReduce性能的方法。文献[14]提出了一种MapReduce任务调度算法LATE,该算法优先执行剩余时间较长的任务,从而达到减少整个查询时间的目的。

4 云计算中的隐私问题

云计算中的隐私问题受到越来越多的关注。最近,研究者们针对云计算中的数据发布、数据挖掘等隐私问题展开了研究。文献[15]提出了在云计算环境中数据发布的隐私保护问题。文献[16-17]提出了云计算中面向隐私保护的数据挖掘问题。其中文献[16]提出的数据挖掘架构,可以在保证数据隐私的情况下,使数据拥有者能正确地恢复出挖掘到的关联规则。文献[18-19]提出了云计算中面向隐私保护的信息检索问题等。文献[20]提出了一种面向隐私保护的最短路径计算方法。

然而,云计算中面向查询处理的隐私保护技术需要借鉴外包数据库和分布式数据库中的隐私保护技术。在外包数据库中,隐私保护的处理主要是基于加密的方式,同时还存在着对查询结果完整性验证的机制;在分布式系统中,面向隐私保护的查询处理主要是基于安全多方计算(secure multi-party computation,SMC)技术。下面分别介绍这几类技术的研究现状。

4.1 基于加密的隐私保护策略

在数据外包的隐私保护处理中,数据拥有者在服务器上的数据是以加密的方式存储的。查询用户的查询也用相同的方式加密,再发送给服务器进行查询处理。不可信的服务提供者为数据拥有者提供数据存储和查询服务,查询用户通常被认为是可信的。外包数据库面向隐私保护的查询处理主要是基于加密方法实现的。文献[21]针对一维数值数据提出了一种保持排序的加密模式(order preserving encryption schema,OPES)。SQL语句中诸如MAX、MIN、COUNT、GROUP BY和ORDER BY等操作可以在加密的数据上进行重写和处理。但是OPES不支持在加密数据上进行SUM和AVG操作,这两种操作必须在数据解密之后进行处理。文献[22]将空间转换技术扩充到二维空间数据上,提出了一种等级空间分割(hierarchical space division,HSD)方法。值得注意的是,在数据外包中,隐私保护主要针对不可信的服务提供方,而普通查询用户是可信的。为了同时保护数据隐私和查询隐私,文献[23]对最近邻查询的变换进行了研究。由于空间填充曲线能处理位置信息以及具有距离保持的属性,该文提出了一种利用空间填充曲线的转换进行隐私保护的方法。然而,在转换后的空间中,距离信息并没有被完全保持,转换结果仅是近似k-近邻查询(k-nearest neighbor,kNN)。此外,空间转换有潜在的暴露隐私的风险,为了能得到精确的查询结果,变换算法必须保存精确的距离信息,攻击者可能利用这些信息进行攻击。例如,文献[24]提出两种方法——线性代数方法和主成分分析方法将原始数据还原。文献[25]利用对象之间的交互距离恢复原始数据。

在数据外包环境中,用户不但同时拥有数据并产生查询,还可以设计一个加密模式支持在加密数据上的某些查询。但是,数据拥有者和查询用户不是同一方的应用中,很难甚至不可能找到一种加密模式可以支持在加密数据上的多种查询处理。比如,空间变换在数据外包中是一种常用的加密模式,然而因为这种方法不能保存在原始空间中的精确数据,所以该加密模式不支持一些需要精确距离的查询,比如最近邻查询。第二,即使可以找到一种加密模式支持多种查询处理,该加密模式必须在查询方和服务器方同时部署,一方可以使用加密参数把对

方的数据解密。为了防止这个漏洞，需要引入一个可信的第三方产生加密参数，这个加密参数必须分别存储在双方防止篡改的设备中。第三，许多加密模式在有安全攻击情况下是很脆弱的。比如空间变换方法在主成分分析方法下很容易被识破。数据库在外包应用中由于保护隐私的需求，需要服务提供商存储的数据是经过加密以后的数据，这样可以保证企业的机密信息不会泄露。但是数据在经过普通加密方法加密后，可用性大大下降，这样会给服务提供方，以及查询用户带来很多的额外开销。因此，在数据库外包的应用中需要能够有效支持数据操作的加密算法。

4.2 基于安全多方计算的隐私保护策略

安全多方计算是在一个分布式网络下，由多个参与方提供的输入来计算某个函数的值。在计算过程中，除了参与者的输入以及输出所暗含的信息之外，不会额外泄露参与方的任何信息。目前已有一些基于安全多方计算（SMC）的隐私保护方法。在SMC中最基本的问题是百万富翁问题。理论上讲，百万富翁问题和多方计算问题都可以用电路评估协议解决。在这个协议中，隐私保护函数用一个布尔电路来表示，每个部分在不暴露各自输入的情况下，联合起来对电路的输出进行评估。各个部分之间的通讯代价由电路大小、输入域大小以及函数的复杂程度决定。如果数据中的属性是由不同的参与方提供的，数据会被垂直划分。在垂直数据划分上也有一些研究工作。文献[26]研究了top-k查询，排序标准是各个参与方独立计算的分数之和的情况。文献[27]在解决隐私问题时引入了可信第三方对各方持有的数据进行连接操作，该连接操作是由可信第三方的一个安全协处理器完成的。文献[28]提出了一种安全k-means聚类方法，可以处理来自两个参与方的水平划分数据或者垂直划分数据。然而，基于SMC的解决方案产生的计算代价和通信代价过高，许多基于SMC的算法都是内存算法，要求数据全部驻留在内存中，因此这种方法不能被直接用于云计算中有几百万条记录的大数据集上。

4.3 查询结果的完整性验证

数据库在外包给第三方服务提供商之后，需要提供额外机制保证外包数据库中的数据不会被未经授权的攻击者修改，服务提供商不能任意向数据库中增加元组，或者删除数据库中的元组。用户查询返回的结果应该是未经修改过的数据库中原始数据，且查询返回的结果是完全的，没有缺失任何有效解。文献[29-33]通过对外包数据中的元组使用密钥签名，让客户可以通过公钥验证数据是否来自原始数据库，并且保证服务提供商返回的用户查询结果是完全的，没有缺失任何有效解。但是，由于数字签名具有相当高昂的计算代价，使用这种方法验证数据的完整性会导致相当可观的客户端以及服务端开销。此外，该方法由于不够灵活，很难扩展用于复杂的查询类型，比如连接查询以及更新查询。另一种已经提出的解决方案[34]则采用了一种基于挑战-响应安全协议来保证服务提供商完整地执行用户的查询。但是该方法并不能保证返回的查询结果的正确性以及完全性，服务提供方完全可以通过修改返回的查询结果攻击该验证模式。而且该方法还需要用户在本地维护一份镜像数据信息，因此并不适合在实际场合使用。本文作者所在课题组在之前的研究工作中[35]提出了一种基于概率的外包数据库结果完整性验证方法，如果数据拥有者在将数据外包给第三方服务提供商时，在数据中混入了一组特别的监测元组，那么对于外包数据库上的所有查询，这些混在原始数据中的监测元组就会以一定概率包含在查询结果中，并返回给提交查询的用户。因此，用户可以通过监控这些额外插入的元组来监控外包数据库的完整性。如果一个满足查询条件的监控元组没有被返回，那么用户就可以断言完整性已经被攻击。反之，如果所有满足查询条件的监控元组都完整地返回，则以一定概率断定完整性没有受到攻击。

5 未来工作展望

综上所述，随着各领域数据规模的快速增长以及云计算技术的不断发展，云计算是未来数据管理

领域一个重要的发展方向,它对于解决海量数据的管理问题以及数据分析决策有着至关重要的意义。然而,云计算中的隐私保护技术是未来云计算发展中一个迫切需要解决的问题。国内外虽然在面向隐私保护的查询处理的某些方面提出了一些解决办法,但是由于上述种种局限,无法直接应用到云计算环境中,很多挑战性问题有待解决。

(1)云计算中查询处理的隐私保护框架研究

在云计算环境中有三个参与方交互,因此需要设计一个云计算环境下查询处理隐私保护框架,该框架能同时保护用户的查询隐私和数据提供者的数据隐私。在云计算环境中,数据量通常是庞大的,单独使用SMC技术进行数据隐私保护需要将所有数据读入内存,显然面临着内存不足的限制。目前,在数据库中构建索引是加快查询处理的常用技术。索引技术还可以减轻现存SMC技术中分布式存储和内存使用过多的问题,因此可以结合SMC技术和索引技术,来构建云数据库上高效的、面向隐私保护的查询。从本质上讲,在索引上进行查询处理可以看做在树结构上遍历结点。查询处理可以分为以下两个步骤:结点检索和结点获取。结点获取决定了下一个要遍历的结点,而结点检索又可以得到下一个结点存储的内容。为了保护查询隐私,在三方模型中,两个步骤都必须是安全的。也就是说,数据拥有者和云数据库系统都不能在用户查询的结点检索过程中识别出结点;也都不能在结点获取过程中获取结点的任何信息。同时,用户在结点获取阶段不能真正地取得真实结点的信息。

(2)安全的结点获取协议

为了实现感知隐私的查询处理技术,并使上述架构支持更多的查询种类,需要为查询用户和云数据库系统的数据交互设计一种安全结点获取协议。为了保护查询隐私和数据隐私,查询和数据项均被加密,给查询处理带来了极大的挑战。针对键值的查询和选择操作可以采用现有的SMC方法处理。然而,针对复杂查询(比如范围查询、kNN查询和连接查询),尤其是基于距离的查询时,SMC方法并不适用。然而,可以从基本的距离安全计算入手,通过对不同的查询类型设计安全结点获取协议,实现在加密数据上的查询处理。目前,有两种可行的解决方案。

第一种解决方案:在加密模式下,在客户端完成查询和索引项间距离的计算。该方案基于的理论基础是秘密同态技术(privacy homomorphism, PH)。PH是一种特殊的加密方法,它能够在密码域保存在明文域里进行的操作。例如,对于支持加法操作的PH函数 $E(\cdot)$,在明文 a 和 b 上进行操作后再加密得到 $E(a+b)$,等同于在密码域内先加密再相加,即 $E(a)+E(b)$。使用PH加密方式,用户可以在加密的用户查询和数据上计算距离。对于距离度量(比如欧式距离),必须设计一种同时支持加法、减法、乘法的PH加密算法,且PH算法必须能抵御密码攻击。此外,计算得出的距离也是由同一个PH算法加密,这就增加了向云数据库系统暴露距离数据的风险。为了防止云数据库系统得到计算出的距离,安全获取协议必须保证用户计算得出的距离信息不会暴露给云数据库系统。使用PH算法的另一属性可以解决这个问题:将加密后的距离分成几个数据片发送给云数据库系统解密。

第二种解决方案:在真实的结点获取之前添加一个距离校准过程。校准过程可以重新存储明文查询和数据项之间的真实距离。比如,对于明文 x 和 y,如果查询用户持有 $Ec(x)$,云数据库系统持有 $Ec(y)$,校准过程将把 $Ec(x)$ 修正为 $cal(Ec(x))$,把 $Ec(y)$ 修正为 $cal(Ec(y))$,而使 $cal(Ec(x))-cal(Ec(y))=x-y$。为了保护隐私,可以在两方计算协议下设计校准过程。以基本加密算法为例:$Ec(\cdot)$ 是一个求幂加密算法,要对这个加密函数进行校准,首先要计算 $(y^{\varepsilon})^{1/\varepsilon}$,其中 y^{ε} 是由云数据库系统保存的,ε 是由查询用户保存的。两边取对数之后,可以将已有的安全两方乘法协议应用到求幂算法上。上述运算的结果是 y,然后将 y 分成两部分 a 和 b,分别由查询用户和云数据库系统保存。云数据库系统用数据 b 得到校准结果 $cal(Ec(y))$,查询用户用数据 $x-a$ 得到校准结果 $cal(Ec(x))$。很容易从上述计算中通过 $cal(Ec(x))-$

$cal(Ec(y))=x-a-b=x-y$ 验证正确性。上面仅用简单的方式举例说明校验方法的可行性，有一些复杂的加密模式的校验过程需要重新设计，这也是该项研究的技术难题之一。

(3)查询结果完整性验证技术

云计算模型的三个参与方构成了一种半信任的安全模型。也就是说，每一部分根据自己的协议独立运行，但是每一部分可以记录中间结果，并且试图通过推理得出关于其他部分的内容。因此，如何设计一套面向隐私保护的查询处理框架，使得三方之间既可以交互信息，又不能通过交互的数据推导出更多的信息是最具挑战的问题。正常情况下，云计算中的三方均独自正确执行各自的协议，此模型是安全的。但是如果某两个参与方用中间结果推导其他参与方的数据或者查询内容，半信任的安全模型就不再安全。更严重的是，如果云数据库系统有恶意行为，那么查询者和数据提供方必须用必要的手段证实查询结果的正确性。目前，可以采用概率方法和确定性方法两种方式。

概率方法：该方法是基于本文作者所在课题组在VLDB07上发表的关于完整性验证机制方法的改进。在云数据库系统中，数据提供方使用公开密钥系统在数据记录上进行数字签名，云数据库系统中的少量假数据记录也被签名。这样，向云数据库系统发出查询时，有一些假数据记录有可能被作为结果返回。假设查询用户事先知道所有的假数据记录，那么查询用户对结果进行分析之后，很容易得到正确的结果。但是如果有一些满足查询条件的假数据记录没有作为结果返回，那么该查询的结果就是不完整的。如果所有满足查询条件的假数据记录都作为结果返回，该查询在一定的概率下是正确的。基于上述思想，可以将该方法扩展到云计算环境中。首先，数据隐私最大的威胁是查询用户，因此产生假记录的函数要重新设计，以免用户从假记录中推导出云数据库系统中真实数据的信息。其次，除了范围查询的授权验证，还将支持基于距离查询的授权验证。最后，研究动态服务质量技术，满足用户的个性化服务的需求。

确定性方法：概率方法新颖、简单，但并不能保证错误结果完全被检测。为了提高检测错误结果的能力，可以设计一种确定性方法。该方法在每个加密后的记录上添加一个核实验证数据结构(verification data structure，VDS)。VDS和加密记录一起作为查询结果返回给用户。如此一来，面临的关键问题是如何设计VDS以满足监测错误结果的要求。例如，对于一个简单的范围查询$[ql,qu]$，为了验证查询结果的正确性，VDS必须能保证以下两点：第一点，返回的结果是邻近的；第二点，包含若干查询范围之外的结果。针对第一点，可以将记录分类，使VDS包含一个由数据拥有者签名的序列号；针对第二点，查询用户可以要求云数据库系统返回额外的几个查询结果。

6 结束语

随着计算机技术及通讯技术的发展，互联网、医疗卫生、通信等行业所产生的数据量都在以指数级别增长，如何以经济实用的方式实现高效存储和管理海量数据是人们面临的极具挑战性的问题。云数据管理系统为人们提供了一种性价比很高的管理海量数据的方式。然而，当云计算提供各种便利服务的同时，也存在着严重的隐私泄露的风险。本文对最近几年国际上在该领域的主要研究成果进行了回顾与总结，综述了云计算中面向隐私保护的查询处理技术的研究现状，指出仍然存在的问题和将来可能的解决办法。总体来说，云计算中隐私保护的研究仍然处于起步阶段，仍然有大量关键的问题需要进行深入细致的研究。

References:

[1] IDC's new IT cloud services forecast: 2009-2013[EB/OL]. (2009-10)[2011-12]. http://blogs.idc.com/ie/?p=543.

[2] Privacy in the clouds: risks to privacy and confidentially from cloud computing[R]. World Privacy Forum Report, 2009.

[3] Hu Haibo, Xu Jianliang, Ren Chushi, et al. Processing private queries over untrusted data cloud through privacy homomorphism[C]//Proceedings of the 27th IEEE International Conference on Data Engineering (ICDE ' 11), Hannover, Germany, Apr 11-16, 2011. Washington, DC, USA: IEEE Computer Society, 2011: 601-612.

[4] Abouzeid A, BajdaPawlikowski K, Abadi D, et al. HadoopDB: an architectural hybrid of MapReduce and DBMS technologies for analytical workloads[J]. Proceedings of the VLDB Endowment, 2009, 2(1): 922-933.

[5] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters[C]//Proceedings of the 6th Symposium on Operating System Design and Implementation, San Francisco, California, USA, Dec 6-8, 2004: 137-150.

[6] Cooper B F, Ramakrishnan R, Srivastava U, et al. PNUTS: Yahoo!' s hosted data serving platform[J]. Proceedings of the VLDB Endowment, 2008, 1(2): 1277-1288.

[7] Aguilera M K, Golab W, Shah M A. A practical scalable distributed b-tree[J]. Proceedings of the VLDB Endowment, 2008, 1(1): 598-609.

[8] Wu Sai, Wu Kunlung. An indexing framework for efficient retrieval on the cloud[J]. IEEE Data Engineering Bulletin, 2009, 32(1): 77-84.

[9] Wang Jinbao, Wu Sai, Gao Hong, et al. Indexing multidimensional data in a cloud system[C]//Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD '10), Indianapolis, Indiana, USA, June 6-10, 2010. New York, NY, USA: ACM, 2010: 591-602.

[10] Zaharia M, Konwinski A, Joseph A D, et al. Improving MapReduce performance in heterogeneous environments[C]// Proceedings of the 8th USENIX Symposium on Operating Systems Design and Implementation (OSDI '08), San Diego, California, USA, Dec 8-10, 2008. Berkeley, CA, USA: USENIX Association, 2008: 29-42.

[11] Nykiel T, Potamias M, Mishra C, et al. MRShare: sharing across multiple queries in MapReduce[J]. Proceedings of the VLDB Endowment, 2010, 3(1): 494-505.

[12] Dittrich J, Quiané-Ruiz J A, Jindal A, et al. Hadoop: making a yellow elephant run like a cheetah (without it even noticing)[J]. Proceedings of the VLDB Endowment, 2010, 3(1): 518-529.

[13] Jiang Dawei, Ooi B C, Shi Lei, et al. The performance of MapReduce: an in-depth study[J]. Proceedings of the VLDB Endowment, 2010, 3(1): 472 - 483.

[14] Zhang Xiangyu, Ai Jing, Wang Zhongyuan, et al. An efficient multi-dimensional index for cloud data management[C]// Proceedings of the 1st International CIKM Workshop on Cloud Data Management (CloudDB '09), Hong Kong, China, Nov 2, 2009. New York, NY, USA: ACM, 2009: 17-24.

[15] Wang Hui. Privacy-preserving data publishing in cloud computing[J]. Journal of Computer Science and Technology, 2010, 25(3): 401-414.

[16] Giannotti F, Lakshmanan L V S, Monreale A, et al. Privacy-preserving mining of association rules from outsourced transaction databases[C]//Proceedings of the Workshop on Security and Privacy in Cloud Computing, Brussels, Belgium, 2010.

[17] Singh M D, Krishna P R, Saxena A. A cryptography based privacy preserving solution to mine cloud data[C]//Proceedings of the 3rd Bangalore Annual Compute Conference, Bangalore, India, Jan 22-23, 2010. New York, NY, USA: ACM, 2010.

[18] Liu Qin, Wang Guojun, Wu Jie. An efficient privacy preserving keyword search scheme in cloud computing[C]// Proceedings of the 2009 International Conference on Computational Science and Engineering (CSE '09). Washington, DC, USA: IEEE Computer Society, 2009: 715-720.

[19] Wang Cong, Cao Ning, Li Jin, et al. Secure ranked keyword search over encrypted cloud data[C]//Proceedings of the 2010 IEEE 30th International Conference on Distributed Computing Systems (ICDCS '10), Genova, Italy, June 21-25, 2010. Washington, DC, USA: IEEE Computer Society, 2010: 253-262.

[20] Gao Jun, Yu J X, Jin Ruoming, et al. Neighborhood-privacy protected shortest distance computing in cloud[C]//Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD '11), Athens, Greece, June 12-16, 2011. New York, NY, USA: ACM, 2011: 409-420.

[21] Agrawal R, Kiernan J, Srikant R, et al. Order-preserving en-

cryption for numeric data[C]//Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD '04), Paris, France, June 13-18, 2004. New York, NY, USA: ACM, 2004: 563-574.

[22] Yiu M L, Kalnis P, Ghinita G, et al. Outsourcing search services on private spatial data[C]//Proceedings of the 25th International Conference on Data Engineering (ICDE ' 09), Shanghai, China, Mar 29-Apr 2, 2009. Washington, DC, USA: IEEE Computer Society, 2009: 1140-1143.

[23] Khoshgozaran A, Shahabi C. Blind evaluation of nearest neighbor queries using space transformation to preserve location privacy[C]//Proceedings of the 10th International Symposium on Large Spatio-Temporal Databases (SSTD '07), Boston, MA, USA, July 16-18, 2007. Berlin, Heidelberg: Springer-Verlag, 2007: 239-257.

[24] Liu Kun, Giannella C, Kargupta H. An attacker's view of distance preserving maps for privacy preserving data mining[C]//Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD ' 06), Berlin, Germany, Sep 18-22, 2006. Berlin, Heidelberg: Springer-Verlag, 2006: 297-308.

[25] Turgay E O, Pedersen T B, Saygin Y, et al. Disclosure risks of distance preserving data transformations[C]//Proceedings of the 20th International Conference on Scientific and Statistical Database Management (SSDBM ' 08), Hong Kong, China, July 9-11, 2008. Berlin, Heidelberg: Springer-Verlag, 2008: 79-94.

[26] Vaidya J, Clifton C. Privacy-preserving top-*k* queries[C]// Proceedings of the 21st International Conference on Data Engineering (ICDE '05), Tokyo, Japan, Apr 5-8, 2005. Washington, DC, USA: IEEE Computer Society, 2005: 545-546.

[27] Li Yaping, Chen Minghua. Privacy preserving joins[C]//Proceedings of the 24th International Conference on Data Engineering (ICDE '08), April 7-12, Cancún, México, 2008. Washington, DC, USA: IEEE Computer Society, 2008: 1352-1354.

[28] Jagannathan G, Wright R N. Privacy-preserving distributed *k*-means clustering over arbitrarily partitioned data[C]//Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '05), Chicago, USA, Aug 21-24, 2005. New York, NY, USA: ACM, 2005: 593-599.

[29] Devanbu P T, Gertz M, Martel C U, et al. Authentic third-party data publication[C]//Proceedings of the 14th Annual Working Conference on Database Security, Schoorl, The Netherlands, Aug 21-23, 2000: 101-112.

[30] Li Feifei, Hadjieleftheriou M, Kollios G, et al. Dynamic authenticated index structures for outsourced databases[C]// Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD '06), Chicago, Illinois, USA, June 27-29, 2006. New York, NY, USA: ACM, 2006: 121-132.

[31] Mykletun E, Narasimha M, Tsudik G. Authentication and integrity in outsourced databases[C]//Proceedings of the Network and Distributed System Security Symposium, San Diego, California, USA, 2004.

[32] Pang H, Jain A, Ramamritham K, et al. Verifying completeness of relational query results in data publishing[C]//Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD '05), Baltimore, Maryland, USA, June 14-16, 2005. New York, NY, USA: ACM, 2005: 407-418.

[33] Pang H, Tan K. Authenticating query results in edge computing[C]//Proceedings of the 20th International Conference on Data Engineering (ICDE '04), Mar 30-Apr 2, Boston, MA, USA, 2004. Washington, DC, USA: IEEE Computer Society, 2004: 560-571.

[34] Sion R. Query execution assurance for outsourced databases[C]//Proceedings of the 31st International Conference on Very Large Data Bases (VLDB '05), Trondheim, Norway, Aug 30-Sep 2, 2005: 601-612.

[35] Xie Min, Wang Haixun, Yin Jian, et al. Integrity auditing of outsourced data[C]//Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB '07), Vienna, Austria, Sep 23-27, 2007: 782-793.

HUO Zheng was born in 1982. She is a Ph.D. candidate at Renmin University of China, and the member of CCF. Her research interests include location privacy-preserving and trajectory privacy-preserving, etc.
霍峥(1982—),女,河北邯郸人,中国人民大学博士研究生,CCF会员,主要研究领域为位置隐私保护,轨迹隐私保护等。

MENG Xiaofeng was born in 1964. He received his Ph.D. degree in computer science from Institute of Computing Technology, Chinese Academy of Sciences in 1999. Now he is a professor and Ph.D. supervisor at Renmin University of China, and the executive director of CCF. His research interests include cloud data management, Web data management, native XML databases, flash-based databases and privacy-preserving, etc.
孟小峰(1964—),男,河北邯郸人,1999年于中国科学院计算技术研究所获得博士学位,现为中国人民大学教授、博士生导师,CCF常务理事,主要研究领域为云数据管理,Web数据管理,XML数据库,闪存数据库,隐私保护技术等。

XU Jianliang was born in 1976. He received his Ph.D. degree in computer science from Hong Kong University of Science and Technology in 2002. Now he is an associate professor and Ph.D. supervisor at Hong Kong Baptist University. His research interests include data management, mobile/pervasive computing and distributed systems, etc.
徐建良(1976—),男,2002年于香港科技大学获得博士学位,现为香港浸会大学副教授、博士生导师,主要研究领域为数据管理,移动及普适计算,分布式系统等。