软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(1):145−166 [doi: 10.13328/j.cnki.jos.004688]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

云数据管理索引技术研究*

马友忠[1,2], 孟小峰[1]

[1](中国人民大学 信息学院,北京 100872)

[2](洛阳师范学院 信息技术学院,河南 洛阳 471022)

通信作者: 孟小峰, E-mail: xfmeng@ruc.edu.cn

摘 要: 数据的爆炸式增长给传统的关系型数据库带来了巨大的挑战,使其在扩展性、容错性等方面遇到了瓶颈.而云计算技术依靠其高扩展性、高可用性、容错性等特点,成为大规模数据管理的有效方案.然而现有的云数据管理系统也存在不足之处,其只能支持基于主键的快速查询,因缺乏索引、视图等机制,所以不能提供高效的多维查询、join 等操作,这限制了云计算在很多方面的应用.主要对云数据管理中的索引技术的相关工作进行了深入调研,并作了对比分析,指出了其各自的优点和不足;对在云计算环境下针对海量物联网数据的多维索引技术研究工作进行了简单介绍;最后指出了在云计算环境下针对大数据索引技术的若干挑战性问题.

关键词: 云数据管理;索引;Hadoop;大数据;多维查询

中图法分类号: TP311

中文引用格式: 马友忠,孟小峰.云数据管理索引技术研究.软件学报,2015,26(1):145−166. http://www.jos.org.cn/1000-9825/4688.htm

英文引用格式: Ma YZ, Meng XF. Research on indexing for cloud data management. Ruan Jian Xue Bao/Journal of Software, 2015,26(1):145−166 (in Chinese). http://www.jos.org.cn/1000-9825/4688.htm

Research on Indexing for Cloud Data Management

MA You-Zhong[1,2], MENG Xiao-Feng[1]

[1](School of Information, Renmin University of China, Beijing 100872, China)

[2](School of Information and Technology, Luoyang Normal University, Luoyang 471022, China)

Abstract: The explosive growth of the digital data brings great challenges to the relational database management systems in addressing issues in areas such as scalability and fault tolerance. The cloud computing techniques have been widely used in many applications and become the standard effective approach to manage large scale data because of their high scalability, high availability and fault tolerance. The existing cloud-based data management systems can't efficiently support complex queries such as multi-dimensional queries and join queries because of lacking of index or view techniques, limiting the application of cloud computing in many respects. This paper conducts an in-depth research on the index techniques for cloud data management to highlight their strengths and weaknesses. This paper also introduces its own preliminary work on the index for massive IOT data in cloud environment. Finally, it points out some challenges in the index techniques for big data in cloud environment.

Key words: cloud data management; index; Hadoop; big data; multi-dimensional query

随着信息技术的飞速发展,在电子商务、物联网、社交网络、计算机仿真、科学计算等众多应用领域,数据量正在以指数级的速度增加,人类已经进入了大数据时代.据统计,全世界的信息量每两年以超过翻倍的速度

* 基金项目: 国家自然科学基金(61379050, 91224008); 国家高技术研究发展计划(863)(2013AA013204); 高等学校博士学科点专项科研基金(20130004130001); 中国人民大学科学研究基金(11XNL010)

收稿时间: 2013-09-13; 修改时间: 2014-01-13; 定稿时间: 2014-07-01; jos 在线出版时间: 2014-08-19

CNKI 网络优先出版: 2014-08-19 14:23, http://www.cnki.net/kcms/doi/10.13328/j.cnki.jos.004688.html

© 中国科学院软件研究所 http://www.jos.org.cn

增长,2011 年将产生和复制 1.8ZB 的海量数据,其增长速度已经超过摩尔定律.传统的关系型数据库虽然能够提供十分成熟的数据存储、索引以及查询处理方案,然而面对不断增长的海量数据,关系型数据库在扩展性方面遇到了严重的瓶颈,无法实现高效、灵活的扩展.虽然一些专业的公司能够提供针对关系型数据库的扩展方案,但其部署、管理的代价非常大.

自 2004 年以来,Google 公司先后提出了 Google File System[1],BigTable[2]和 MapReduce[3]等技术.随着这三大关键性技术的提出,云计算作为一种新的海量数据存储、管理、分析模式应运而生,并得到业界众多大公司的广泛应用和深入研究,云计算已经成为海量数据处理的一种标准解决方案.同时也出现了很多优秀的分布式数据存储和管理系统,如雅虎的 PNUTS[4],Amazon 的 Dynamo[5]、开源的 HBase 等.基于 key-value 存储的云数据管理技术具有高可扩展性、高可用性和容错性等特点,能够实现对海量数据的高效存储和处理.

然而,现有的基于 key-value 存储的云数据管理系统在数据访问方面提供的功能比较简单.云数据管理系统大都是按照 rowkey 的顺序对数据进行组织,并在 rowkey 上建立类似 B+-tree 的索引结构,所以在 rowkey 上能够提供高效的点查询或范围查询.然而针对非 rowkey 的查询,它们只能通过全表扫描的方式来实现.虽然我们可以利用 MapReduce 技术来实现数据访问的并行化,在一定程度上提高查询速度,但是当数据量非常大的时候,对于时间延迟要求比较高的应用来说,全表扫描所需的时间仍然比较长,无法满足实际应用的需求.

在实际应用中,除了对 rowkey 的查询之外,还有很多针对非 rowkey 的多维查询需求.如在基于位置的服务中,我们经常需要针对某个对象的经度、纬度、时间等属性进行多维查询;在图片共享服务中,我们可以对图片的拍摄时间、拍摄地点、图片主题等属性进行查询;在电子商务网站中,商品的数量往往达到数十亿、甚至上百亿,并且每件商品都有几十个甚至上百个属性,如名称、类别、价格、上架时间等.用户往往需要从多个不同的角度对商品进行查询,从而对所要购买的商品有更加全面而深入的了解.

由于目前云数据管理系统在数据查询方面的局限性,限制了其在众多领域的广泛应用.索引是实现多维查询的一个有效方案,因此目前已有很多学者、公司针对云数据管理中的索引技术开展了大量研究工作,并提出了一系列有价值的解决方案.比如新加坡国立大学的 epiC 项目组创新性地提出了双层索引框架,并在此基础上给出了一系列解决方案;华为公司基于 HBase 的 Coprocessor 技术设计了新的二级索引方案,大大提高了查询的效率.本文主要对云数据管理索引技术的相关工作进行了深入调研,分析了各自的优点及不足;并对我们在物联网索引方面的研究工作进行了简单介绍;最后指出了在云计算环境下针对大数据索引技术的若干挑战性问题.

本文第 1 节对云数据管理中的索引技术进行分类.第 2 节和第 3 节分别介绍基于分布式文件系统的索引和基于 key-value 存储的索引.第 4 节介绍不同数据类型的索引技术.第 5 节指出云计算环境下针对大数据的索引技术中存在的若干挑战性问题.最后对本文进行总结.

1 云数据管理索引技术研究概述

云数据管理是指以云计算技术为基础,针对大规模数据的分布式、可扩展的数据管理技术.与传统的数据管理(以关系数据库管理技术为主)相比,在数据规模、数据对象、系统结构等方面都存在不同之处,并各有优劣,详细对比情况见表 1.

Table 1 Cloud data management vs. traditional data management

表 1 云数据管理与传统数据管理

对比项目	云数据管理	传统数据管理
数据模型	无明确的数据模型	关系模型为主
数据规模	超大规模数据	中小规模数据
数据类型	结构化、半结构化、非结构化	结构化为主
系统架构	分布式	集中式为主
优点	高可扩展性,高容错性,高可用性	功能丰富,容易使用,索引支持
缺点	不易使用,缺乏有效的索引支持,对 Join 等复杂查询支持不好	扩展性差,无法有效管理海量数据,无法有效管理半结构、非结构化数据

© 中国科学院软件研究所 http://www.jos.org.cn

为了丰富云数据管理系统的查询功能,提高数据查询和处理的效率,很多学者开展了云数据管理系统中索引技术的研究工作,并提出了很多有价值的实现方案.我们对各种索引方案的索引结构、实现方式、优缺点进行了深入分析,并按照数据的存储方式以及数据类型进行了划分,作了对比研究.按照数据存储方式的不同,可以分为基于分布式文件系统的索引和基于 key-value 存储的索引;按照索引对象数据类型的不同,可以分为轨迹数据索引、空间数据索引、图数据索引和物联网数据索引等.各类索引中根据索引实现方式的不同,又包含了多种不同的索引方案,具体见表 2,下面将分别进行详细的描述.

Table 2 Classification of index solution of cloud data management

表 2 云数据管理索引方案分类

分类标准	索引类别	索引方案
存储方式	基于分布式文件系统的索引	双层索引 全局分布式索引 位图索引 基于 Hadoop 框架的索引
	基于 key-value 存储的索引	二级索引 基于线性化技术的索引 HugeTable
数据类型	轨迹数据索引 空间数据索引 图数据索引 物联网数据索引	基于划分的多层索引 PMI 基于 R-tree 的索引;KR^{+}-index; Neo4j,Trinity UQE-index

2 基于分布式文件系统的索引

分布式文件系统(如 HDFS)能够对海量数据进行高效的存储和管理,其存储和管理数据的基本方式是:整个存储系统可以是由众多廉价的计算机组成的集群,其中有一个或多个 master 节点,众多 slave 节点,master 节点负责收集和管理全局信息,slave 节点负责数据的存储.数据一般以块的形式分散存储在不同的 slave 节点上.在这种情况下,如果要对数据进行索引,就必须考虑分布式文件系统的框架和数据组织方式.根据实现方式的不同,我们又可以把这些索引分为双层索引、全局分布式索引、位图索引和基于 Hadoop 框架的索引.表 3 给出了各种方案的优缺点、主要索引结构以及代表性技术,后面的章节将分别对各种索引方案进行详细介绍.

Table 3 Index solution based on distributed file system

表 3 基于分布式文件系统的索引方案对比

索引方案	优点	缺点	主要索引结构	代表技术
双层索引	可以通过全局索引来定位数据所在节点,避免额外查询开销	索引维护代价高 实现较复杂 维度较高时,查询性能急剧下降	B-tree	B-tree+BATON[6]
			R-tree	RT-CAN[7] A-Tree[8] EMINC[9]
			Quad tree	QT-Chord[10]
			多种索引结构	DBMS-like[11]
			Bitmap+R-tree	TLB-index[12]
全局分布式索引	实现简单 自动负载均衡	客户端负担较重 无法满足数据频繁插入	B-tree	Distributed B-tree[13]
位图索引	高并发 避免额外检索代价	属性值域划分规则复杂 范围查询时会出现很多 false positives	Bitmap	RBI[14]
基于 Hadoop 框架的索引	可以充分利用 MapReduce 的高效处理能力	需要对 Hadoop 框架进行修改	可以支持不同的索引结构	HadoopDB[15] HAIL[16]
		创建索引的代价比较大		Hadoop++[17]

2.1 双层索引

目前已有的云数据管理索引方案中关于双层索引的研究比较多,研究人员主要来自新加坡国立大学、中国

© 中国科学院软件研究所 http://www.jos.org.cn

人民大学、东北大学、希腊 Thessaly 大学.双层索引方案主要是为了满足数据的海量性以及系统的可扩展性而提出来的.伍赛等人[18]于 2009 年提出了一个云数据管理系统中的双层索引框架,后续的关于双层索引方案的研究大都是基于该框架进行的.该框架的实现方案如图 1 所示.

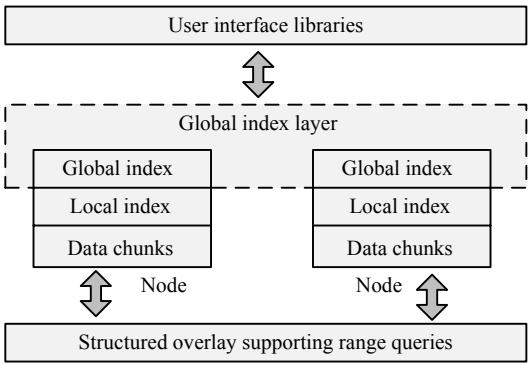


Fig.1 Framework of double-level index
图 1 双层索引基本框架

由图 1 可知,索引部分包括局部索引和全局索引.在云数据管理系统中,有大量的由廉价计算机组成的计算机集群可以为用户提供计算资源和存储资源.用户数据按照一定的规则被划分成数据块,这些数据块按照分布式文件系统的协议被分配到不同的计算机节点存储.在双层索引方案中,对每个计算机节点的数据建立一个本地的局部索引,该局部索引只负责本地节点上的数据.除局部索引外,每个计算节点还需要共享一部分存储空间用来存储全局索引.全局索引是由部分局部索引组成的,由于存储空间的限制和查询效率的要求,不可能把所有的局部索引节点全都发布到全局索引中,所以针对局部索引需要按照一定的规则选择其中的一部分索引节点来进行发布.对于被选中的索引节点,在全局索引中可以有不同的方案对其进行组织.

在双层索引方案中,根据局部索引的索引结构以及全局索引的组织方式的不同,又可以分为多种不同的种类,表 4 描述了各种方案的局部索引、全局索引以及节点选择策略.

Table 4 Double-Level index solutions

表 4 双层索引方案对比

方案名称	局部索引	全局索引	索引节点选择策略(代价模型)
Efficient B-tree[6]	B^+-tree	BATON	查询代价+索引维护代价
RT-CAN[7]	R-tree	CAN	查询代价+索引维护代价
QT-Chord[10]	IMX-CIF quad-tree	Chord	四叉树节点编码规则
TLB-Index[12]	B^+-tree for Bitmap	BATON for Bitmap	未明确描述
DBMS-Like[19]	多种索引结构	Cayley 图	根据历史查询进行选择
EMINC[9]	KD-tree	R-tree	索引节点的大小
A-Tree[8]	R-tree+Bloom filter	Array based	索引节点的大小

Efficient B-tree[6]是一个基于 B+-tree 和 BATON 网的双层索引方案,主要支持单个属性上的点查询和范围查询.在每个存储节点上,针对本地数据建立相应的 B+-tree 索引,然后按照特定的索引节点选择策略从每个局部索引中选择一部分节点发布到全局索引中.为了提高查询的速度,保证系统的可扩展性,全局索引采用了 P2P 网络 BATON[20]对全局索引节点进行组织.在进行索引节点选择时采用了基于代价模型的自适应调整策略,Wu 等人[6]提出了一个代价模型,每个节点的代价包括查询代价和索引维护代价,在进行节点选择时保证总的代价最小.具体调整过程是:最初选择每个局部 B+-tree 索引的根节点发布到全局索引中,然后定期地计算全局索引节点的子节点的代价之和,如果子节点的代价之和小于原有父节点的代价,那么就把原有的父节点替换成新的子节点.该方案能够对单个属性的查询提供高效的支持,但是无法实现多维查询.

RT-CAN[7]的基本实现思路与 Efficient B-tree[6]是一样的,不过 RT-CAN[7]是针对多维查询而设计的.为了支持多维查询,RT-CAN[7]在局部索引中采用了 R-tree 结构,在全局索引中采用了能够支持多维查询的覆盖网络,

© 中国科学院软件研究所 http://www.jos.org.cn

即 CAN 网络,节点选择策略与 Efficient B-tree[6]基本一致.R-tree 对于数据规模较大、插入比较频繁的应用来说,维护代价过高,会对系统的吞吐量带来一定的影响.因此,丁琳琳等人[10]提出了另外一种双层索引结构——QT-Chord.在 QT-Chord 中,局部索引采用的是改进的四叉树 IMX-CIF Quad-tree,全局索引采用的是 Chord 覆盖网络[21].在索引节点的选择上,QT-Chord 并没有利用基于代价的模型,而是基于四叉树的编码方法对每一个四叉树节点进行编码.由于父节点的编码是子节点编码的前缀,所以丁琳琳等人充分利用这一特点,设计了一个命名函数,把子节点递归地映射到上层父节点,然后把最后映射到的节点发布到全局索引中.

TLB-index[12]采取的索引方案与 Efficient B-tree[6]采用的方案基本一致,局部索引采用的是 B+-tree,全局索引采用 BATON 网络,不同之处在于,TLB-index[12]对每一个不同的值构造一个 bitmap,然后利用 B+-tree 对所有的 bitmap 进行索引,这样可以减少存储开销.

在一个数据管理系统中,往往需要多种不同的索引结构,如 B+-树、R-树等.文献[19]在文献[6,7]的基础上,通过进一步的分析和抽象,提出了一种可扩展的、简单的统一索引框架,该框架可以支持多种不同的索引结构,如哈希索引、B+-树、R-树等,并且可以用一个统一的 Cayley 图来管理多种不同的覆盖网络结构(overlay network).文献[19]主要通过两个映射函数来将不同类型的索引映射到 Cayley 图,数据映射函数负责将不同类型的数据值映射到 Cayley 图的键值空间(key space),覆盖映射函数由多个不同的路由算法操作符组成,不同的路由算法对应不同的 Cayley 图实例.由于文献[19]对不同的索引类型和覆盖网络提供了统一的管理框架,所以,可以大大降低索引的更新和维护代价,增强了系统的通用性.

前面 5 种方案在全局索引中均采用了基于 P2P 结构的覆盖网络,这种方式能够很方便地实现系统的可扩展性,使系统能够同时支持大规模的查询.但是这种方式也存在一些不足之处:首先,P2P 网络本身需要一定的维护代价,在查询时,往往也需要一定的网络传输代价,这在一定程度上会增加系统的延迟.另外,由于现有的云数据管理系统一般都是 master-slave 结构的,要在这些节点上重新构建一个 P2P 网络,会对原有的存储系统带来一定的负面影响.基于这些原因,EMINC[9]和 A-Tree[8]在全局索引中采用了集中式的索引方式.EMINC[9]中针对每一个局部节点建立一个 KD-tree[22],然后选择 KD-tree 的部分节点作为全局索引,Zhang 等人[9]把每一个局部索引节点看成是一个多个维度的立方体,然后在全局索引中用 R-tree 对这些立方体进行索引.由于 R-tree 树的各个节点之间会有一定的重叠,尤其是当维度较高,或者数据量较大时,重叠部分会比较多,这样在查询的时候,尤其是点查询,就会造成大量的 false positive 查询,从而影响查询的效率.为了解决这一问题,A-Tree[8]提出了另外一种方案:带 Bloom filter[23]的 R-tree,基本思路是这样的:针对每一个存储节点构建一棵 R-tree,同时创建一个 Bloom filter,在进行点查询时,首先通过 Bloom filter 进行验证,如果查询点不在其中,则不再进行 R-tree 查询,否则继续进行 R-tree 查询.这样就减少了 false positive 的机率,从而提高了查询效率.

上述各种索引技术方案具有较好的扩展性,但是总体来说实现过程比较复杂,索引更新维护的代价比较高,对于数据更新比较频繁的引用来说,会影响系统的性能.在全局索引中大都采用了基于 P2P 协议的覆盖网络来实现并行查询,但是 P2P 网络本身的维护比较复杂,查询时的网络开销也比较大,这会影响到系统的查询性能.

2.2 全局分布式索引

为了支持大规模数据存储,并保证系统具有较高的吞吐量,Distributed B-tree[13]提出了一种容错的、高可扩展的分布式 B-tree 结构.该方案除了具有传统 B-tree 的一般特点之外,还具有一些新的特性:自动负载均衡、操作的原子性和存储节点的动态增加或删除.其基本思想如图 2 所示.

所有数据以 B-tree 结构进行组织,B-tree 的节点(包括内部节点和叶子节点)分散存储在不同的节点上.为了实现数据的一致性,文献[13]的作者引入了 version table,用以记录节点的最新版本.为了提高查询的效率,B-tree 的内部节点都会缓存在客户端,并采用 lazy replica 的方式进行更新.这种方法主要有两个缺点:一是该方法对于简单的点查询具有较高的效率,但是对于复杂的范围查询和多维查询效率较低;二是服务器端的维护代价较高,客户端需要消耗大量的内存空间去缓存 B-tree 的内部节点.

© 中国科学院软件研究所 http://www.jos.org.cn

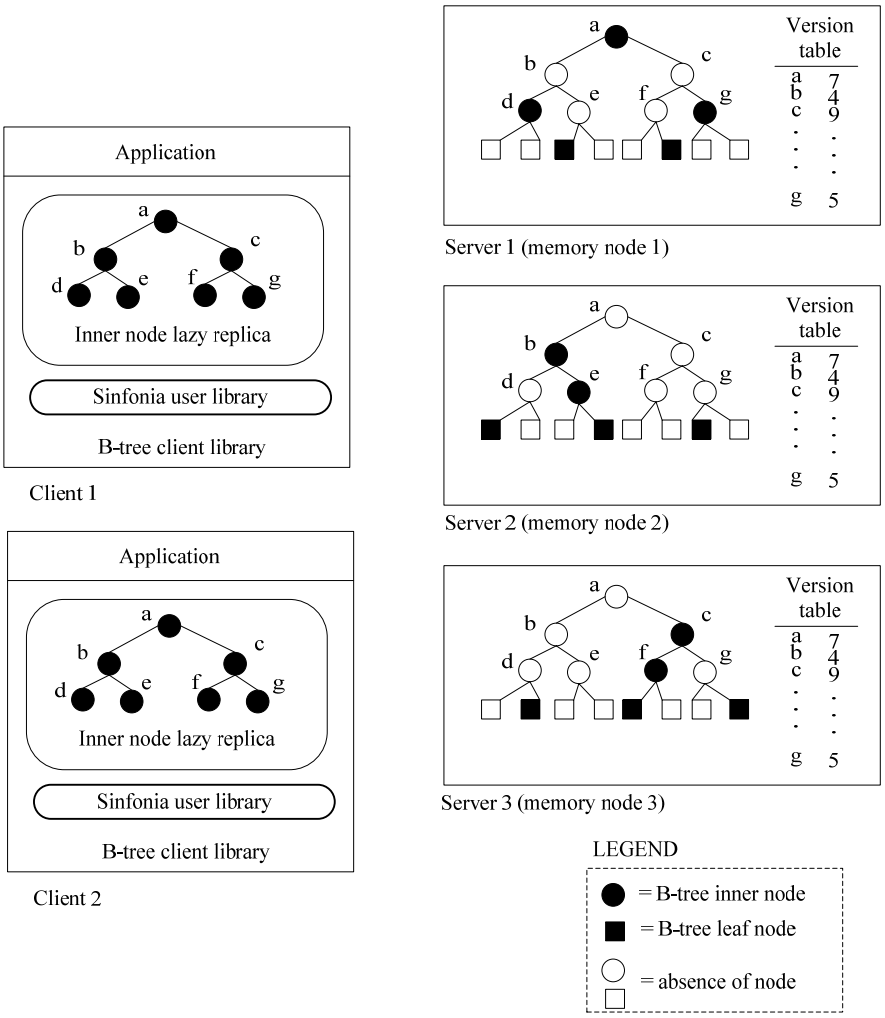


Fig.2 Distributed B-tree

图 2 分布式 B-tree

2.3 分片位图索引

分片位图索引[14]是在结合集中式索引和分布式索引各自优点的基础上提出的一种新的混合式索引方案.集中式索引的索引项一般由索引字段和原数据记录的主键组合而成,索引项按照索引字段值排序,同时索引项也采用 key-value store 来管理.集中式索引方案往往具有高可靠性、可扩展性和易管理性.但是由于查询以集中方式处理,因此无法通过并行化来充分利用分布式资源的优势.分布式方案是在各个节点建立局部索引,节点之间相互独立,因此局部索引的结构比较灵活,可以实现高度并发查询;但是由于每次查询都需要查询所有的节点,所以往往会带来额外的、不必要的查询开销,特别是对选择率很低的等值查询,会访问很多无用节点.扩展的分布式索引方案是在局部索引的基础上增加了一层全局索引,当一个查询来的时候,首先通过全局索引来确定对应的局部索引,然后再到各个局部节点去查询,从而避免了不必要节点的查询.但是全局索引往往采用一些 P2P 网络,如 BATON、CAN[24]网络来实现,结构比较复杂,网络开销比较大.分片位图索引[14]中也采用了全局索引和局部索引相结合的方式来实现,但是全局索引采用了一种比较新颖和有效的方式.与以往扩展分布式方案中使用复杂 P2P 网络所实现的全局索引相比,分片位图索引采用基于属性值的全局排序机制,通过一个全局位图来表示局部数据在全局值域中的分布情况,从而为各数据节点提供一定的全局信息,并且该全局位图直接存

© 中国科学院软件研究所 http://www.jos.org.cn

储在各个节点上,不需要特殊的全局索引结构来实现,因此可以避免不同数据节点之间的网络通信开销.除全局索引之外,文献[17]还引入了位图索引的分片机制,即各局部节点也都有各自的局部索引,从而使得各数据节点的查询任务相互独立,便于并发执行,以提高查询效率.但是该方案的问题在于,进行全局排序时,各属性值域的划分规则比较难以确定,因为不同的划分规则,对查询的性能会产生比较大的影响.除此之外,在实际应用当中,业务类型经常会发生改变,因此需要增加新的索引列,而要增加新的索引列,上述索引方案就需要对全部数据进行重新处理和索引,代价往往会非常大.

2.4 基于Hadoop框架的索引技术

目前 Hadoop 已经逐渐成为海量数据存储与处理的标准平台,通过 HDFS 可以实现海量数据的高效存储和管理,通过 MapReduce 可以实现海量数据的并行查询与处理.但是由于缺乏索引技术的支持,使得某些复杂查询(如针对非 rowkey 的单维或多维查询)的执行效率比较低.为了提高海量数据的复杂查询处理的效率,很多学者在已有的 Hadoop 相关技术的基础上,进行了索引技术的研究,期待基于这些索引技术,能够提高利用 MapReduce 对海量数据的查询处理效率.表 5 给出了各种索引方案的详细信息.

Table 5 Index solution based on Hadoop framework

表 5 基于 Hadoop 框架的索引技术

方案名称	优点	缺点
HadoopDB	将 Hadoop 框架与并行数据库进行有机结合; 可以利用并行数据库提供的索引功能及高效的查询性能	需要对 Hadoop 框架进行大量修改
Trojan Index	对 Hadoop 进行了非入侵式的优化,便于和新的 Hadoop 框架兼容	创建索引的代价比较大
HAIL	针对每一个数据块的物理副本建立不同的索引; 索引是在数据上传时建立的,所以创建的代价比较小	需要修改 Hadoop 框架中的部分接口

HadoopDB[14]首次尝试将 Hadoop 框架与并行数据库结合起来:Hadoop 中的 HDFS 能够实现大规模数据的分布式存储,MapReduce 可以支持大规模数据的高度并行处理;并行数据库则支持索引、视图等功能,具有丰富的查询功能.HadoopDB 将两者的优势有机地结合在了一起,大大提高了复杂查询的性能.但是为了能够与并行数据库有机结合,需要对 Hadoop 框架进行相应的修改,如果有新版本的 Hadoop 出现,则无法兼容,需要对新版本的 Hadoop 框架重新进行修改.

Hadoop++[17]则对 Hadoop MapReduce 进行了非入侵式优化,即不需要对 Hadoop 框架本身作任何修改,这主要是通过自定义函数(UDF)来实现的,如 split,map,reduce 等.Hadoop++对 Hadoop 的优化主要包括 Trojan Index, Trojan Join 和 Trojan Layout 这 3 个方面.其中 Trojan Index 的核心是针对 HDFS 中的每一个逻辑块(logical block)创建一个块级别的索引.通过用户自定义函数,将每一个逻辑块中的数据进行重新组织,组成一个新的依次由数据、索引、Header 和 Footer 这 4 部分构成的 block,其中 Footer 是 split 的分界符,最后一个 Footer 一定位于文件末尾.索引构建时由 MapReduce 完成排序.查询时 split 函数从文件末尾开始根据 Footer 信息解析出各个 split,itemize 函数根据搜索范围条件快速定位满足条件的内容,这样就可以大大提高查询的性能.但是 Hadoop++创建索引的代价比较大,原因是,当数据上传到 HDFS 以后,需要用一个新的 MapReduce Job 来对每一个逻辑块建立索引,其开销甚至比对数据进行全表扫描的代价还要高.

为了降低创建索引的额外开销,HAIL[16]提出了一种入侵式的索引库方案(Hadoop aggressive indexing library).其主要思想是针对 HDFS 中每一个 block 的每一个物理备份分别建立不同的聚簇索引,这样,每一个备份就可以在不同的属性上建立索引,从而可以支持多个属性的查询.为了降低建立索引的开销,HAIL 是在上传数据的过程中来建立索引的,主要是对原有的 HDFS client 进行修改得到 HAIL client,通过 HAIL client 来上传数据,在上传数据的过程中,对每一个副本分别建立聚簇索引,并且索引建立的过程是在内存中完成的,索引开销比较小.同时,HAIL 对 MapReduce JobClient 进行了修改,使得在执行 MapReduce Job 之前首先对已经建立的索引进行选择,然后再选择相应的物理备份进行处理,从而就可以使用已有的索引,大大提高查询的效率.

© 中国科学院软件研究所 http://www.jos.org.cn

除了针对 HDFS 和 MapReduce 的通用索引方案之外,在某些具体问题的处理中也用到了一些特定的索引技术.如李飞飞等人[25]在 MapReduce 环境下,利用基于 z-order 划分的方法来解决 KNN Join 的问题,其中在 reduce 任务中执行连接操作时,利用 R-tree 对局部数据进行了索引,从而提高了连接的速度.Vernica[26]等人利用 MapReduce 技术来解决集合相似性连接的问题,他们利用 prefix-filtering[27]技术对所有的集合建立倒排索引,然后在倒排索引的基础上进行比较,从而大大减少了比较的次数,提高了执行效率.

3 基于 key-value 存储的索引

随着云计算技术的不断升温,云数据库也得到迅速发展,目前已经出现了很多以 key-value 形式存储的云数据库管理系统,如 HBase,Cassandra,MongoDB 以及 HugeTable[28]等.这些云数据库系统都具有较好的可扩展性,为海量数据的存储和管理提供了一个行之有效的方案.但是现有的云数据库管理系统只能提供比较简单的基于 rowkey 的查询,缺乏索引支持,无法提供复杂的多维查询、join 等操作,从而限制了其在某些方面的应用.针对现有系统存在的上述问题,一些企业、科研人员以及各开源社区团队都进行了一些研究,设计了不同的索引方案,表 6 列出了现有的几种基于 key-value 存储的索引方案,主要包括二级索引、基于线性化技术的索引和 HugeTable.

Table 6 Index solutions based on key-value storage

表 6 基于 key-value 存储的索引方案对比

索引方案	优点	缺点	主要索引结构	代表技术
二级索引	实现简单; 维护代价低	多维查询效率较低	索引表	THbase IHbase Asynchronous View[29]
		需要访问所有的 region	索引表	基于 Coprocessor 的二级索引
		空间冗余大	索引表	CCIndex[30]
基于线性化技术的索引	较高的写入吞吐量; 维护代价较低	一致性维护复杂	KD-tree	MD-HBase[31]
HugeTable	能够实现可伸缩、高性能的海量结构化信息存储	需要重新进行开发,工作量较大	稠密索引 稀疏索引 块索引	HugeTable

3.1 二级索引

二级索引(secondary index)方案类似于关系数据库中的二级索引,该方案主要应用于 key-value 存储的云数据库管理系统中,如 BigTable,HBase 等,目前大多数的二级索引方案都是基于 HBase 实现的.在 HBase 中,数据是以表的形式来存放的,为了实现对非 rowkey 的查询,需要为每一个索引列构建一个索引表,以索引列的值(其实是索引列与 rowkey 的组合)作为新的 rowkey,实现索引列与原有 rowkey 的映射.查询时,首先通过在索引表中进行查询,找到相应的原有 rowkey 的列表,然后再根据这些 rowkey 信息到原来的数据表中去查找所需要的原始数据信息.

目前基于二级索引的实现方案主要有 ITHBase,IHBase,CCIndex[30],Asynchronous views[29]和华为公司基于 Coprocessor 技术的索引.其中 ITHBase 和 IHBase 是两个开源的实现方案,ITHBase 主要关注于数据一致性,事务性是其重要特性.IHBase 与 ITHBase 比较相似,从 HBase 源码级别进行了扩展,重新定义和实现了一些 Server,Client 端处理逻辑,所以,它是具备强侵入性的.不过,IHBase 在修改完 HBase 0.20.5 版兼容 bug 以后没有再更新.

除了 ITHBase 和 IHBase 之外,中国科学院提出了另外一种二级索引方案:互补聚簇式索引(complemental clustering index,简称 CCIndex)[30].前面的二级索引方案中,索引表中仅仅存放索引列与原来的 rowkey 信息,在查询时,通过查询索引表得到 rowkey 以后,还需要根据 rowkey 到原来的表中去查找,然而由于得到的 rowkey 大都是随机的,所以需要进行大量的随机读操作才能最终得到所需要的数据,效率相对比较低.为了减少随机查询带来的开销,CCIndex 提出了一个新的方案,把数据的详细信息也存放在索引表中,这样在查询的时候就可以直接

© 中国科学院软件研究所 http://www.jos.org.cn

在索引表中通过顺序扫描找到相应的数据,即把随机读变成了顺序读,可以大大缩短查询时间.然而,把详细信息存储在索引表中会造成存储空间的增加,为了尽可能地减少存储空间的开销,Zou 等人[30]把 HDFS 文件块备份数设为 1,这样就可以保证存储空间不会增加太多.但是,备份数设为 1 之后,数据的容错性又成了新的问题,为了解决这一问题,他们又提出了新的容错方案,通过创建 clustering check table,再配合原有的索引表,能够实现快速恢复.同时,CCIndex 还提出了一种查询优化机制,用以支持多维查询.该优化机制主要是利用 HBase 中的一些元数据信息(region-to-server information)来估算每个子查询结果的大小,根据查询结果来生成合适的查询计划,从而降低查询时间.CCIndex 方案实现起来相对比较简单,但是也存在一些不足之处,如存储开销比较大,尤其是当索引列比较多的时候,空间开销会更大;索引更新代价比较高,会影响系统的吞吐量;索引创建以后,不能够动态增加或修改.

Asynchronous View[29]提出了一种异步视图的方式来实现非 rowkey 的查询,视图的方式类似于二级索引,因为视图物化以后就相当于索引.文献[29]提出了两种视图方案:远端视图表(remote view tables,简称 RVTs)和局部视图表(local view tables,简称 LVTs).RVTs 是指针对某个查询列建立一个视图,视图与原始表一样,可能会被分散存储到各个不同的节点上,这样,视图中索引的数据与视图所在节点的数据可能不一致,当节点中的数据发生改变的时候,就需要对其他节点上的视图进行更新维护,代价较高.这种方式对于查询来说效率比较高,同时维护代价也比较大.LVTs 是针对每个节点的数据分别建立一个局部的视图,这样,节点数据发生改变的时候就可以直接在本地对局部视图进行更新,维护代价较小.但是在查询时,就需要对所有节点的视图进行查询,然后再进行汇总,所以查询效率比较低.一般来说,对于聚集查询,可以用 LVTs,其他查询可以用 RVTs,而对一些特殊的查询,还可以采用两者相结合的方式.

Coprocessor 技术出现以后,一些研究机构基于 Coprocessor 机制对 HBase 的二级索引进行了创新性的修改,如华为公司,其改进的重点在于减少由随机查询带来的大量网络开销.下面以 scan 操作为例来描述其具体的实现机制.

如图 3 所示,R1 表示原用户表的一个 region,R2 是索引表的一个 region,并且 R2 中索引的数据与 R1 中的数据是相互对应的,即 R2 中只索引 R1 中的数据,这是与原有二级索引相比的一个主要区别.当客户端发起一个 scan 操作时,Coprocessor 会去查询各个 region 对应的索引表的 region,通过索引表可以查到原有数据的 rowkey 列表,然后再根据得到的 rowkey 列表去 R1 中查询原始数据的详细信息,最后再把结果返回到客户端.

由图 3 可以看出,在最终结果返回之前,所有的操作都是在服务端完成的,所以就减少了大量的网络开销,从而提高了查询效率.

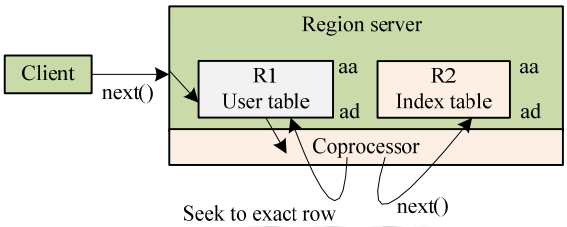


Fig.3 Procedure of scan operation

图 3 scan 操作流程

为了达到上述目的,在建立索引表时,必须采取特殊的机制.由于要保证索引表的 region 与原数据表的 region 一一对应,所以索引表不能自动分裂,其分裂必须由原数据表来控制.如图 4 所示.当原数据表的一个 region 分裂成两个子 region 的时候,其对应的索引表的 region 也分成两个新的子 region,并且与数据表的两个子 region 一一对应.

© 中国科学院软件研究所 http://www.jos.org.cn

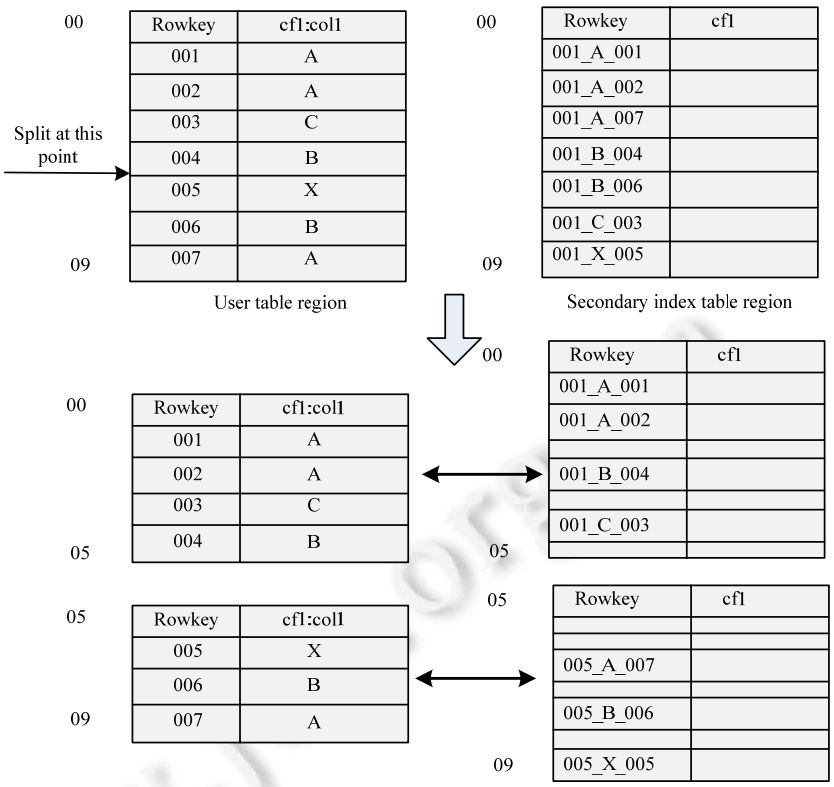


Fig.4 Region splitting scheme

图 4 Region 分裂机制

上述实现方案具有很多优点,如可以在多个列上建立索引,并且可以动态地增加或减少索引,同时不需要对 HBase 的核心代码作太多的修改,便于 HBase 的升级和维护.该索引实现方案应该是目前为止扩展性和实用性都比较好的一种方案.

3.2 基于线性化技术的索引

上述几种索引方案均需要维护一种特定的索引结构,如 R-tree 或 B-tree,当数据更新十分频繁的时候,索引更新维护的代价就非常高,所以为了降低索引更新维护的代价,同时又能保证系统的性能,MD-HBase[31]提出了一种基于线性化技术的索引方案.MD-HBase[31]的基本思想是:按照一定的规则将整个空间范围划分为大小相等的格子,并给每一网格分配一个编号,用这些编号为空间目标生成一组具有代表意义的数字.其实质是通过某种特定的方式将 k 维空间的实体映射到一维空间,从而可以利用现有数据库管理系统中比较成熟的一维索引技术来组织数据.用一维的数值对多维的空间目标进行排序的方法有很多,常见的方法有 Z-order 排序、Hilbert 曲线等.这些技术的思路基本相同,都是利用一个线性序列来填充空间,构造一种空间填充曲线,把多维空间的数据转换成一维数据.这种转换关系必须能够较好地保持原多维空间目标之间的临近关系,从而提供较好的多维查询性能.

MD-HBase[31]以 HBase 作为数据存储方案,用 Z-ordering 技术对数据进行排序,并以 Z-value 作为每条记录的 rowkey 值.但是单纯的 Z-ordering 排序方法在搜索的过程中会带来一些额外的不必要的搜索空间(false positive search),所以,文献[31]在此基础上利用 KD-tree 或四叉树对多维数据空间进行划分,根据最长公共前缀的方式计算得到每个子空间的名称,并以此名称作为索引项对数据各个子空间的数据进行索引,从而减少不必要的检索,大大提高了查询效率.基本思路如图 5 所示.但是,该方法在进行空间划分的过程中,会带来数据的一

© 中国科学院软件研究所 http://www.jos.org.cn

致性问题,虽然目前有相应的解决方案,但是实现起来仍比较复杂,也会带来一定的额外负担.

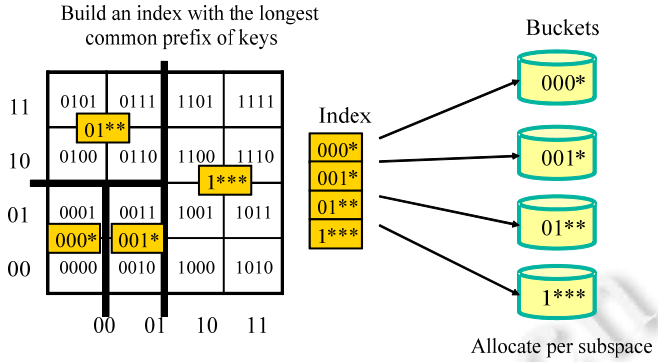


Fig.5 Index based on longest common prefix

图 5 基于最长公共前缀的索引

3.3 HugeTable

“大云”计划是中国移动研究院为打造中国移动云计算基础设施实施的关键技术,目前可实现分布式文件系统、分布式海量数据仓库、分布式计算框架、集群管理、云存储系统、弹性计算系统、并行数据挖掘工具等关键功能.其中,HugeTable 是一种高可靠、可伸缩、高性能的海量结构化信息存储系统,能够实现分布式海量数据仓库.

图 6 是 HugeTable 的系统架构,其中 Index Store 主要用来提供索引支持,基于不同的存储引擎和查询要求,HugeTable 设计了密集索引、稀疏索引和块索引 3 种索引方案.在密集索引中,针对每条数据记录都会建立一条索引项,B+-tree 树是一种常用的密集索引结构.密集索引的主要优势是能够在索引列上提供高效的点查询和范围查询.为了减少索引空间开销,HugeTable 又提供了另外一种索引结构-稀疏索引,稀疏索引用来记录每个数据块中所包含的索引列的最大值和最小值,在查询的时候,通过将待查询值与索引项中的最大值和最小值进行比较来确定候选索引项.如果索引列的值分布比较均匀且在顺序存储的情况下,稀疏索引的性能较好;但是,如果索引列的值随机分布,则稀疏索引效率较低.针对电信应用的具体特点,HugeTable 又提出了块索引方案.块索引的基本格式为⟨key,fileID,blockID⟩,表示在 blockID 块中包含某个 key.与密集索引相比,块索引具有索引速度快、查询性能高和装载速度快等优点.

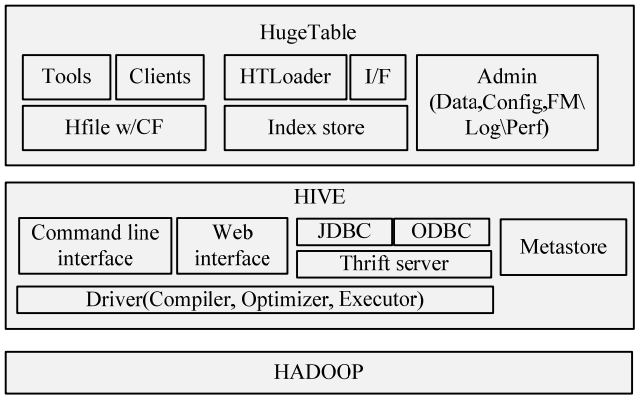


Fig.6 System framework of HugeTable

图 6 HugeTable 系统架构

针对不同索引方案的特点,HugeTable 设计了如下的查询框架,如图 7 所示.当提交一个新的查询时,HugeTable 会首先分析查询列的索引情况:如果查询列有索引,则可以直接通过索引来查找所需数据;如果查询

© 中国科学院软件研究所 http://www.jos.org.cn

列上没有建索引,则根据应用和数据量本身的特点来确定不同的查询方式.如果数据量比较少,则可以直接采取顺序扫描的方式来查询;如果数据量比较大,则可以利用 MapReduce 方式来并行查询,从而提高查询性能.

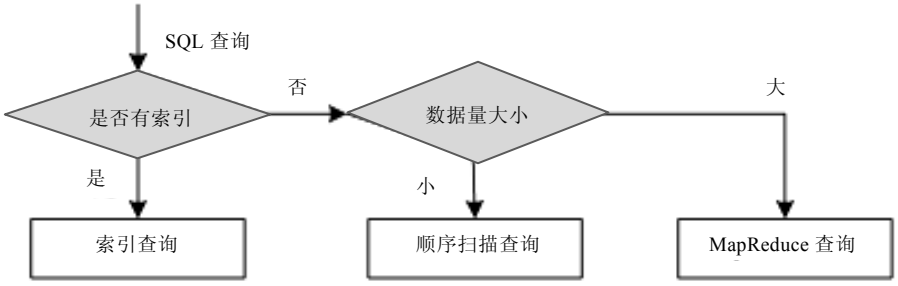


Fig.7 Query framework of HugeTable

图 7 HugeTable 查询框架

4 针对不同数据类型的索引技术

随着云计算技术的不断发展和广泛应用,云计算和云数据管理技术在不同应用领域中都得到了尝试和应用,如海量轨迹数据管理、空间数据、图数据管理以及物联网数据管理.该节主要对上述几种不同应用领域中所涉及到的索引相关技术进行分析和总结.

4.1 轨迹数据

随着物联网技术、GPS 定位技术和基于位置的服务等相关技术的快速发展和广泛应用,产生了越来越多的移动数据,传统的集中式处理技术已经无法高效处理如此大规模的移动数据,因此有部分学者开始研究基于云计算的海量移动数据处理技术,希望通过对海量轨迹数据的高效存储和处理,为智能交通提供决策支持.

Ma 等人[32]主要对海量轨迹数据的存储与检索进行了研究,采用 HDFS 来对轨迹数据进行存储,利用 MapReduce 提供查询.他们以线段模型来表示轨迹数据,并设计了一种静态-动态相结合的轨迹数据划分方案,实现了轨迹数据的高效存储,如图 8 所示.在划分的过程中,首先利用四叉树技术对轨迹数据在空间上进行划分,然后在每个划分区间内再根据时间维度进行划分,在时间区间的选择上是可以动态改变的,如果某个空间区域内的轨迹数据比较稠密,则时间间隔可以小一些,否则时间间隔可以大一些,从而实现轨迹数据的有效划分,以尽可能地保证在空间、时间上相邻的轨迹数据存放在相同的分块中.每一块中的轨迹存储在 HDFS 中的一个 block 中.为了实现轨迹数据的快速查询,Ma 等人提出了两种不同的索引方案.针对空间-时间的范围查询,他们提出了 Partition based Multilevel Index(PMI);针对轨迹查询,提出了 Object Inverted Index(OII).图 9 描述了相应的处理框架.

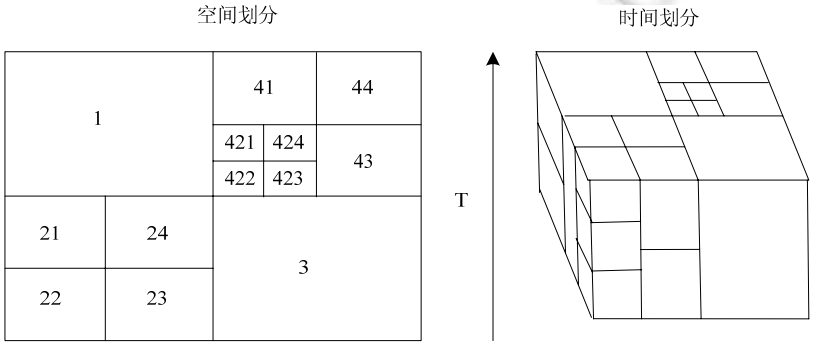


Fig.8 Data dividing solution

图 8 数据划分策略

PMI 主要是依据前面描述的划分方案,把处于同一划分中的轨迹线段存储在同一个数据块中,并以

© 中国科学院软件研究所 http://www.jos.org.cn

〈PartitionID,Sk〉的形式记录下来,这样在进行空间-时间范围查询时,就可以首先根据划分方案,计算出需要查询的轨迹可能位于哪些块中,然后直接到相应的块中去查询,这样就可以大大节省查询的开销.OII 主要是为了查询某个车辆的轨迹而设计的,通过搜集某个车辆的所有轨迹线段,并以〈OID,{PartitionID,T}〉的形式记录下来,实际上,相当于是建立了一个倒排索引表.当需要查询某个特定车辆轨迹的时候,就可以直接利用 OII 找到相应的车辆,从中即可查出其对应的轨迹.

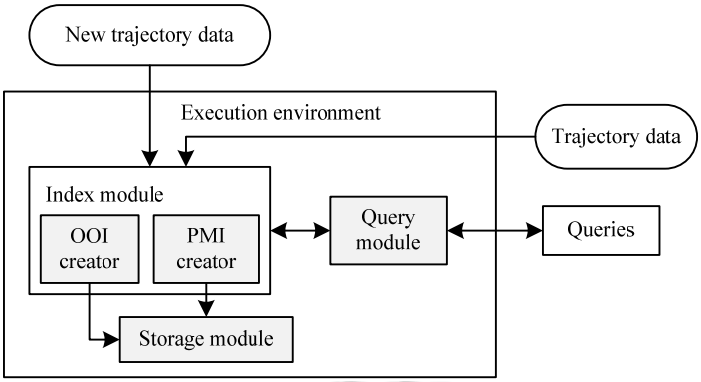


Fig.9 Processing framework of trajectory data

图 9 轨迹数据处理框架

为了提高大规模轨迹查询的效率,Chu 等人[33]在 HBase 的基础上提出了一种可扩展的轨迹索引方案.根据轨迹数据的特点,他们主要设计了两种表结构来实现轨迹数据的存储和索引:主表(main table)主要是用来存储原始轨迹数据,包括 GPS 坐标及其他轨迹相关数据;轨迹索引表(trail indices table)主要是对空间维度进行不同粒度的划分,然后在不同的粒度上对空间点分别建立索引,这样,查询的时候就可以从粗粒度开始进行过滤,大大提高了查询效率.

张彤等人[34]结合浮动车数据的实际特点,设计了基于 Bigtable 的存储模型,并针对路网数据和浮动车数据分别建立了相应的索引,以提高查询处理的效率.存储模型如图 10 和图 11 所示.

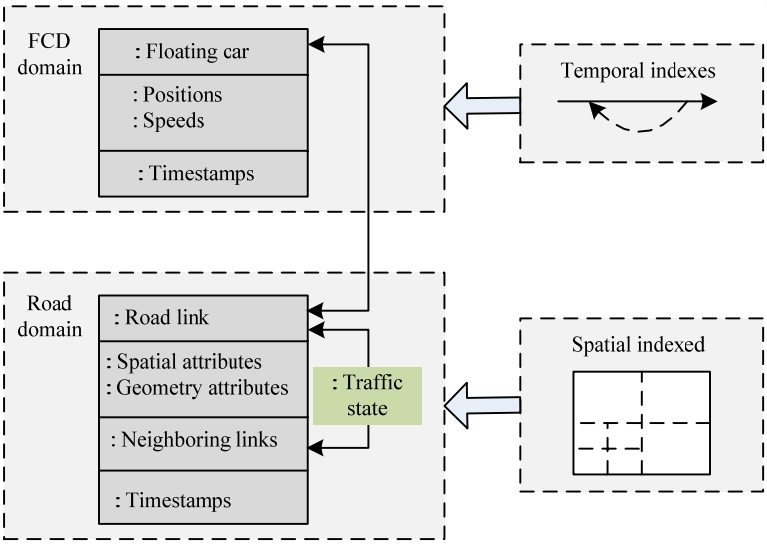


Fig.10 Concept view

图 10 概念视图

为了查询出租车在某段时间内的运行轨迹,张彤等人[34]针对浮动车数据建立了时间索引,实际上是利用时

© 中国科学院软件研究所 http://www.jos.org.cn

间戳和车辆 ID 组合形成新的 rowkey,建立了二级索引.同时,他们还针对路网数据建立了四叉树索引,方便进行地图匹配和车速计算.地图匹配和车速计算是浮动车数据处理常见任务,可以为智能交通提供决策支持.为了提高地图匹配和车速计算的效率,他们利用 MapReduce 技术进行了优化.

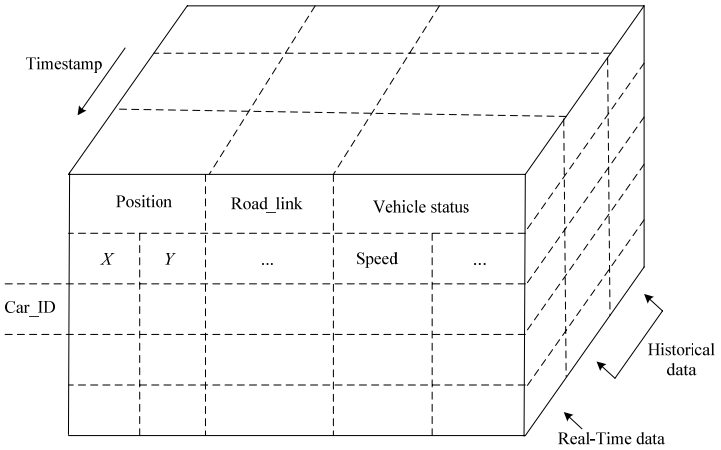


Fig.11 Floating data table structure

图 11 浮动车数据表结构

文献[34]提出的存储模型和索引方案,虽然可以提高某些查询的效率,但对其他一些复杂查询,仍然无法实现较好的支持,如多维度的范围查询(某时间段内、某区域范围内的车辆).所以,设计一种可支持多种不同类型查询的存储与索引方案仍然是一件极具挑战性的工作.

4.2 空间数据

随着空间数据对象的不断增加,如何在海量空间数据对象中实现高效、快速的空间查询成为一个日益重要的问题.Cary 等人[35]重点描述了海量空间数据库中的空间查询问题.他们首先分析了传统的基于 R-tree 的索引结构的一些缺点和不足,然后基于云计算技术提出了一个新的可扩展的解决方案,基本思路如图 12 所示.方案主要包括 3 个阶段:第 1 阶段是数据划分.为了找到一个比较好的数据划分方案,Cary 等人通过采样的方法,得到相应的数据划分方案,并以此作为整体数据的划分方案.第 2 阶段是生成 R-tree,根据第 1 阶段得到的划分方案,在空间上将所有数据划分为互不相交的若干部分,对于每一部分的数据分别建立 R-tree 索引.由于各部分数据是不重叠的,索引的建立可以采用并行的方式来进行,这样可以大大提高索引的构建速度.他们采用 MapReduce 的方式来进行索引的建立,Map 阶段的主要任务是根据前一阶段的划分方案把每一个空间对象划分到相应的分区中,Reduce 阶段分别对每一个分区中的数据建立 R-tree.第 3 个阶段的任务是对第 2 阶段所产生的各个子 R-tree 进行合并,生成一棵大的 R-tree.查询时,首先根据整体的索引树找出待查找对象所在的子 R-tree,然后分别到对应的子 R-tree 上去进行查询,最后再对查找到的结果进行合并.

Hsu 等人[36]针对海量空间数据提出了一种基于 R+-tree 和 key-value 键值生成规则的索引结构——KR+-index,其基本结构如图 13 所示.KR+-index 首先利用 R+-tree 对数据进行划分,把数据划分成若干个互不相交的矩形,然后再利用网格对空间进行划分,并利用 Hilbert-curve 对这些网格单元进行编码,根据网格单元和矩形的交叉情况,记录下网格编号与矩形单元的对应关系.查询时,首先根据查询范围在 KeyTable 中找出数据所在的矩形单元,然后再根据所找到的矩形单元去查找相应的数据.除上述空间数据索引技术之外,崔纪锋等人[37]对云存储环境下海量空间数据的索引技术进行了全面总结,在国内尚属首次.他们按照不同空间数据的特点和空间索引结构对云存储特征的要求有所不同进行了分类,主要包括适应单节点的海量空间索引结构、适应集群元数据服务的层次索引结构以及基于空间语义推理的空间索引算法,并指出了云存储环境下,海量空间数据索引技术的发展趋势及若干挑战性问题.

© 中国科学院软件研究所 http://www.jos.org.cn

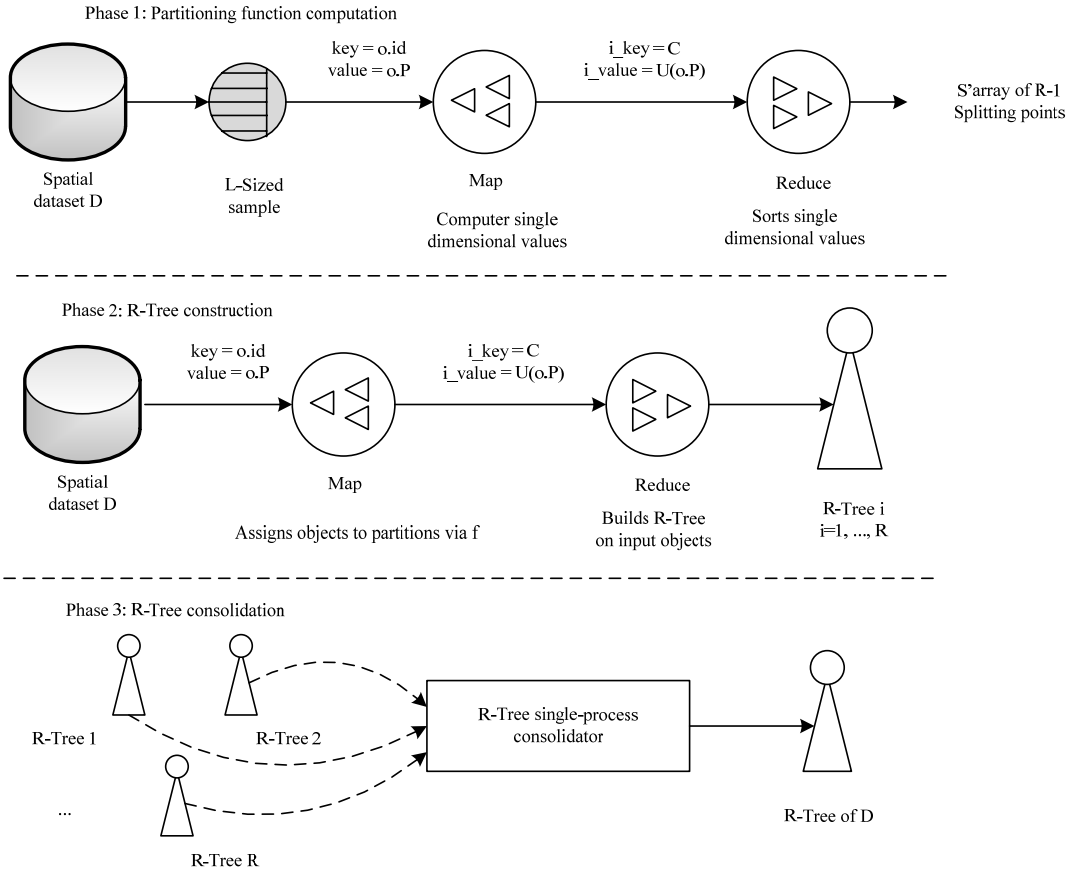


Fig.12 Index procedure

图 12 索引过程

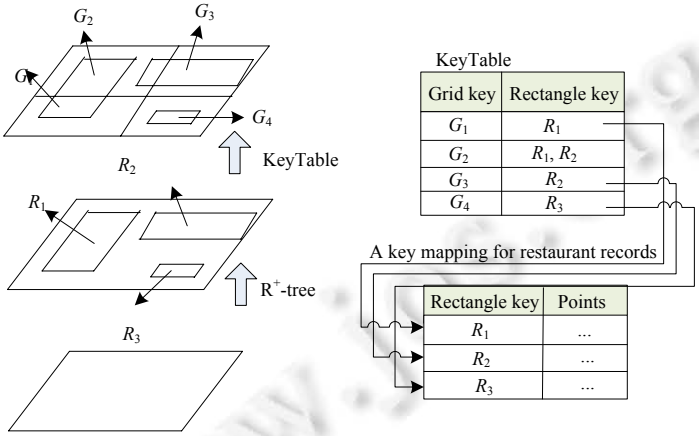


Fig.13 Basic structure of KR^+-index

图 13 KR^+-index 基本结构

4.3 图数据

随着社交网络分析、语义 Web 分析、生物信息网络分析等新兴应用的快速增长,对亿万个顶点级别大规模图的处理能力的需求愈加迫切,于戈等人[38]结合云计算的特点,从图数据管理与图数据处理机制两个方面,综

© 中国科学院软件研究所 http://www.jos.org.cn

述了云计算环境下进行大规模图数据处理的关键问题,并对大规模图数据的索引技术进行了总结.他们首先对目前云计算环境下的通用索引技术进行了简单分析,并指出这些索引技术虽然不是专门针对图数据进行设计的,但大都可以被大规模图数据管理所借鉴.在文献[38]中,他们重点对两种分布式图数据库中所采用的索引技术进行了分析.

Neo4j 的索引分为两类:数据库本身就是一个树形结构的索引,可用于提高查询效率.此外,还可以使用独立的 Lucene 索引,提供全文索引和索引命中率排序功能.Neo4j 可以对图顶点和边分别建立索引,通过对索引的缓存,进一步加快查找速度.索引的维护操作(如删除、更新)则必须在事务管理机制下进行.对于更新,必须先删除旧的索引值,然后才能添加新索引值,代价较高.Trinity 数据库提供 Trie 树和 Hash 两种索引结构来访问图顶点、边的名字以及相关的其他信息,从而可以减少有公共前缀的字符串的匹配次数.

4.4 物联网数据

在对云数据管理系统中现有索引技术进行深入调研分析的基础上,我们针对云计算环境下海量物联网数据的多维索引做了一些研究工作[39],提出一个能同时支持频繁更新和高效多维查询的索引结构 UQE-index.为了同时支持高效的频繁更新和快速的多维查询,主要采取两个措施:一是对当前数据和历史数据在不同的粒度级别上进行索引,在数据写入的过程中,尽可能地减少索引更新的次数,从而降低索引维护的代价;二是尽量把索引和数据分开,减少索引创建对系统写入性能的影响.图 14 和图 15 分别描述了系统架构和多层索引框架.

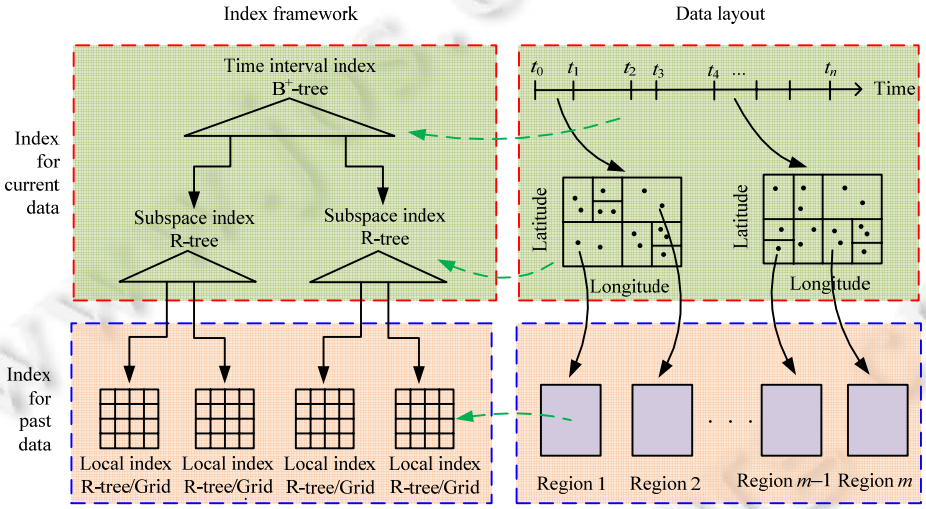


Fig.14 Multi-Level index framework
图 14 多层索引框架

图 14 展示了索引的框架,主要包括 3 个层次,其中时间段索引和子空间索引主要针对当前数据进行索引,网格索引主要针对历史数据进行索引.由于物联网数据具有时空连续性,并且不同时刻的数据在空间上的分布往往会发生变化,所以我们把时间维度和空间维度分开进行考虑.首先在时间维度上分成若干个时间段,针对每个时间段内的数据,在二维空间中进行划分,划分成若干个子空间,每个子空间内的数据存储到 HBase 中的同一个 region 中,这样就能够保证在时间和空间上临近的数据尽量存储在相同的区域中,减少了查询过程中需要扫描的 region 的数量,从而提高了查询效率.

在索引时,我们在时间维度上把数据分成当前时段的数据和历史时段的数据.对于当前时段的数据,仅对其所在的时间段和所在的子空间进行索引,而不对数据记录本身进行索引,这样,在数据插入的时候就可以大大减少索引更新的次数.其中时间段可以采用 B+-tree 进行索引,由于每个子空间都是互不相交的矩形区域,所以采用 R-tree 对每个子空间进行索引.当前时段结束以后,数据开始从新的时间段插入,历史时段的数据不再发生改变,此时可以批量地对历史时段中每个 region 内部的数据建立记录级别的索引,可以采用 R-tree 索引或者网格

© 中国科学院软件研究所 http://www.jos.org.cn

索引.该索引方案更新维护的代价比较低,对数据插入性能的影响比较小,从而保证系统能够支持大规模的频繁更新.最后,我们结合 HBase 开发了原型系统,并做了大量实验,实验结果表明,我们所提出的索引方案既能支持高效的多维查询,又能保证较高的系统吞吐率,达到了预期目的.

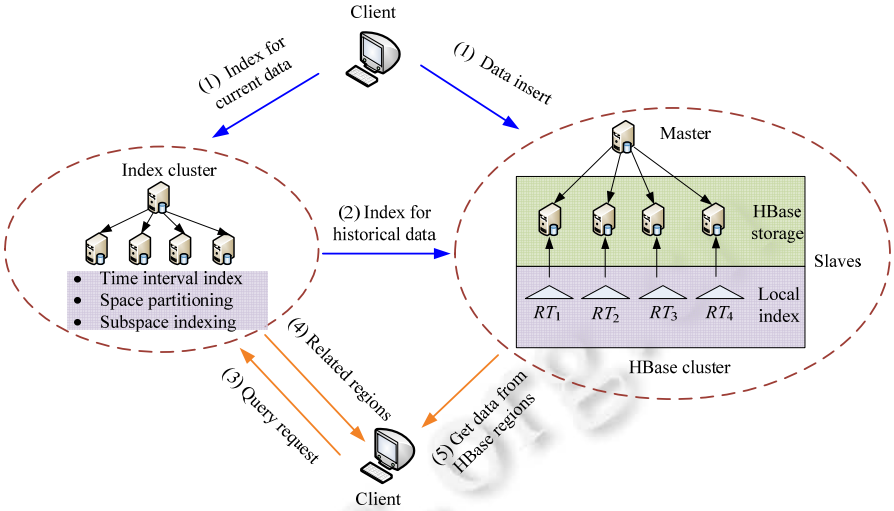


Fig.15 System architecture

图 15 系统架构

4.5 图像数据

图像数据一般根据某种特征提取算法把图像的视觉特征提取出来,然后把图像表示成高维向量,最后在高维向量的基础上建立相应的索引,基于此提供图像检索.然而,图像的索引构建计算代价和索引存储开销都比较大.文献[19]针对上述问题,提出了一种高维分布式局部敏感哈希索引(LSH)方法.该文首先对现有的局部敏感哈希索引方法进行改进,变成松耦合的索引结构,然后在 Hadoop 框架下提出了索引的并行化构建算法以及分布式索引存储方案.为了减少 LSH 索引的 I/O 操作,该文还提出了一种新的分布式内存索引管理方法,从而大大提高了图像的检索效率.文献[40]针对海量图像数据提出了一种基于 MapReduce 的高维索引构建算法,该算法是在 ePC 算法(extended cluster pruning algorithm)的基础上进行扩展和改进而来,可以快速索引大规模图像数据.在索引的基础上,提出了一种基于 MapReduce 的快速批量查询处理机制,能够提供较高的吞吐量.该文对包含有 1 亿张图片的数据集进行了实验,结果表明了所提方案的有效性.

4.6 文本数据

Google 公司提出 MapReduce 框架的初衷就是为了提高海量网页数据的索引构建及更新效率,采用的基本索引结构是倒排索引(inverted index),但是 Google 公司并没有披露详细的实现细节,有部分其他学者和研究机构对海量文本数据的分布式索引技术进行了研究.

文献[41]在 Hadoop 框架的基础上,提出了一个分布式的文本索引,称为分布式 Lucene.Lucene 本身是一个成熟的、集中式的开源文本索引框架,分布式 Lucene 的主要目的就是结合 Hadoop 框架的特征,对原有的 Lucene 进行重构,实现分布式的 Lucene,从而能够索引大规模的文本数据.为了提高索引的速度和索引模块的独立性,文献[41]分析了 HDFS 本身的一些局限性,设计并实现了一些新的 API.

文献[42]提出了一种基于云存储的文档索引分类存储模型.其核心是两层的 MapReduce 分类存储算法,第 1 层 MapReduce 主要负责构建倒排索引表,第 2 层 MapReduce 主要负责计算每个索引词在不同文档中的权重,并根据给定的权重阈值,对文档进行分类存储,这样就可以得到具有权重阈值范围的倒排索引表.从而在不降低查全率的基础上提高查询速度.

文献[43]针对大规模数据的选择查询提出了一种基于全文索引的优化方法.在 HDFS 中,数据是以块(block)

© 中国科学院软件研究所 http://www.jos.org.cn

为单位进行组织和存储的,为了节省存储空间,一般都采用了特定的压缩算法.在进行查询时,需要遍历所有的数据块,先解压,再比较.虽然可以采用 MapReduce 框架来提高并发性,但是扫描的代价仍然非常高,因为很多数据块都不包含所要的数据.为了减少访问的数据块数量,该文提出了一个全文倒排索引方案,其思想非常简单,与传统的倒排索引一样.不同之处在于,每一个索引词对应的不再是包含该词的文档列表,而是包含该词的数据块的列表.这样,在处理查询时,就可以首先通过倒排索引确定相关的数据块,然后再针对相关数据块进行解压、对比,从而可以大大提高查询效率.

4.7 RDF数据

随着 RDF 数据规模的不断扩大,基于云平台的分布式 RDF 存储与查询方案变得日益重要.文献[44]结合 Cassandra 的数据组织与存储特点,提出了一种分布式的倒排索引机制.该文把具有相同主语(subject)的所有三元组的集合称为一个 RDF 文档,然后提出了一个基本的倒排索引结构,以每个单词作为主键(row key),以 RDF 文档 ID 作为列名(column name).基本倒排索引结构具有很好的扩展性,但却没有考虑到 RDF 数据模型本身的特点,也没有包含相应的三元组信息.该文还提出了一种改进的倒排索引结构,该索引结构中,仍是以单词作为主键,以文档 ID 作为超列名(super column name),以资源(resource)ID 作为列名(column),列的值表示资源的位置信息.在上述索引结构的基础上,文献[44]提出了基于 MapReduce 的索引构建算法和查询算法.文献[45]为了实现对海量 RDF 数据的有效管理和快速检索,提出了一种基于分布式哈希表(distributed Hash table)的 RDF 数据存储和索引方案,并提出了基于代价的分布式查询优化算法.文献[46]针对不同的 RDF 查询类型和查询需求,提出针对 RDF 数据的水平索引和垂直索引两种索引结构,并提出基于 MapReduce 的分布式索引构建方案.

5 未来工作展望及挑战

虽然目前针对云数据管理中的索引技术已经有了很多研究工作,但是目前的研究成果还不是十分成熟,还未能得到广泛的应用.并且,随着 Web 2.0、物联网、移动互联网的快速发展和广泛应用,数据呈现爆炸式增长,人类已经进入了大数据时代.云计算是解决大数据的一种有效方式,而随着数据量的不断增长以及数据复杂性的不断增加,在云计算环境下,针对大数据的索引技术也面临着一系列新的挑战.

5.1 复杂数据类型的索引

如前所述,随着社交网络、物联网技术、移动互联网、智能手机等相关技术的快速发展,一些复杂类型的数据量在呈指数级增长,给传统的数据管理方法及索引技术带了巨大挑战,如图数据、轨迹数据、序列数据、空间数据等.针对大规模图数据而言,虽然基于云计算环境下的并行处理机制可以提高其管理和处理效率,但是由于图数据本身的复杂性、特殊性,使得现有的云计算处理方式无法很好地适应大规模图数据的管理.在大规模图数据处理中,只有很少的处理应用了云计算环境下的索引技术,如最短路径计算等[38],大部分的图处理问题还没有考虑到索引的应用,有待于进一步深入研究.轨迹数据也是越来越重要的一种数据类型,主要是得益于智能手机的普遍应用以及移动互联网的快速发展,如基于位置的服务(LBS)、基于位置的社交网络(LBSN)、FourSquare、车辆的 GPS 数据等.虽然针对轨迹数据的索引研究已经有很多相对比较成熟的技术,但在现有的应用环境下,数据量已远非过去的应用所能比,再加上轨迹数据往往更新比较频繁,同时也具有多维特性,使得已有的轨迹索引技术无法在云计算环境下有效地管理海量轨迹数据.

5.2 针对复杂查询的索引

复杂查询主要包括多维查询、最近邻查询(KNN)、Top-K 查询、Join 以及 KNN-Join 等.上述复杂查询类型在数据规模比较小的情况下,都有比较成熟的索引技术和查询方案,可以实现高效查询.但在大数据情况下,原有的索引技术和查询处理方法已经无法满足性能要求,并且目前的云计算技术,如 MapReduce,还不能很好地支持多维查询、连接查询等复杂操作.因此,在云计算环境下,设计特定的索引方案,从而提高针对大数据的、复杂查询的处理效率,是一个重要的研究问题.

© 中国科学院软件研究所 http://www.jos.org.cn

5.3 面向数据隐私保护的索引技术研究

随着数据规模的不断扩大,越来越多的企业把自己的数据部署到公有云或者私有云中.然而企业和个人在享有云计算带来的好处的同时,也面临着隐私泄露的风险.在云计算环境下,为了提高查询效率,往往需要创建相应的索引,然而在提供高效查询功能的同时,如何保护数据的隐私成为一个亟待解决的问题.目前,面向数据隐私保护的索引技术研究工作还处于起步阶段,需要进一步深入研究,为云计算技术的广泛应用提供技术保障.

5.4 基于负载均衡的索引维护机制

在云计算环境下,数据往往分散存储在各个不同的节点上,为了保证整个系统的性能,云存储系统一般都具有自动负载均衡的功能.在负载均衡的过程中,数据会进行迁移,即从一个节点迁移到另一个节点.数据发生迁移以后,对应的索引也需要作相应的调整,如果索引保持不变的话,那么索引和对应的数据就会处于不同的数据节点上,查询的代价会比较高;如果索引随着数据一起迁移,负载均衡的代价又会比较高,也会影响整个系统的性能.所以,如何设计一种有效的索引维护机制,既能够保证原有系统的负载均衡功能,又能够提供高效的查询,是一个重要的研究课题.

5.5 基于索引的查询优化

云计算环境下,基于 MapReduce 的查询优化技术已经有一些相关的研究工作[47],如 MapReduce 在多核硬件与 GPU 上的改进[48]、基于调度技术的 MapReduce 性能优化[49];还有一些工作对现有的 MapReduce 框架进行了修改[50],从而提高了查询效率.在云计算环境下,也有一些研究工作基于索引进行查询优化,如 CCIndex[30]、Hadoop++[17].但是目前的查询优化技术还比较简单,无法很好地满足实际应用的需求,因此在这方面还需要进一步深入加以研究.基于代价的查询优化在传统的关系数据库系统中已经有比较深入的研究,并得到了广泛的应用.但是基于代价的查询优化需要一系列统计信息,如数据表的总记录数、每个属性中不同值的个数、数据的分布直方图等等.但是目前这些统计信息在云计算环境下还无法方便地获得.所以通过对原有云计算框架的修改或扩充,使其能够支持统计信息的收集和更新,并与索引技术相结合,进而支持查询优化,是一个可以尝试的方案.

6 结束语

随着数据的爆炸式增长,我们已经进入了大数据时代.由于云计算技术本身所具有的一些优势,如高扩展性、可用性、容错性等,使得云计算技术得到了广泛应用,因此,云计算也就成为大数据存储、管理与处理的一种有效方式.然而云计算本身也存在一些不足之处,如缺乏对索引的支持就是其中之一,这限制了云计算在很多方面的应用.本文对云数据管理的索引技术的相关问题进行了深入研究,并作了分析、对比和总结,指出了其中各自的优点及不足;同时对我们在物联网数据多维索引方面的研究工作进行了介绍;最后指出了在云计算环境下针对大数据索引技术的若干挑战性问题,并进行了简单分析.该问题的研究还处于起步阶段,还没有比较成熟的、可以进行大规模实际应用的成果和方案,因此,具有重要的研究价值和实际意义.

致 谢 感谢慈祥、张榆等同学对本文提出的宝贵建议.

References:

[1] Ghemawat S, Gobioff H, Leung ST. The google file system. In: Proc. of the 19th ACM Symp. on Operating Systems Principles. New York: ACM, 2003. 29−43. [doi: 10.1145/945445.945450]

[2] Chang F, Dean J, Ghemawat S, Hsieh WC, Wallach DA, Burrows M, Chandra T, Fikes A, Gruber RE. Bigtable: A distributed storage system for structured data. ACM Trans. on Computer Systems, 2008,26(2):1−26. [doi: 10.1145/1365815.1365816]

[3] Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. In: Brewer EA, Chen P, eds. Proc. of the 6th Conf. on Symp. on Opearting Systems Design & Implementation. New York: ACM, 2004. 137−149.

© 中国科学院软件研究所 http://www.jos.org.cn

[4] Cooper BF, Ramakrishnan R, Srivastava U, Silberstein A, Bohannon P, Jacobsen HA, Puz N, Weaver D, Yerneni R. PNUTS: Yahoo!'s hosted data serving platform. Proc. of the VLDB Endowment, 2008,1(2):1277−1288. [doi: 10.14778/1454159.1454167]

[5] DeCandia G, Hastorun D, Jampani M, Kakulapati G, Lakshman A, Pilchin A, Sivasubramanian S, Vosshall P, Vogels W. Dynamo: Amazon's highly available key-value store. In: Proc. of the 21st ACM Symp. on Operating Systems Principles. New York: ACM, 2007. 205−220. [doi: 10.1145/1294261.1294281]

[6] Wu S, Jiang DW, Ooi BC, Wu KL. Efficient B-tree based indexing for cloud data processing. Proc. of the VLDB Endowment, 2010,3(1):1207−1218. [doi: 10.14778/1920841.1920991]

[7] Wang JB, Wu S, Gao H, Li JZ, Ooi BC. Indexing multi-dimensional data in a cloud system. In: Proc. of the 2010 Int'l Conf. on Management of Data. New York: ACM, 2010. 591−602. [doi: 10.1145/1807167.1807232]

[8] Papadopoulos A, Katsaros D. A-Tree: Distributed indexing of multi-dimensional data for cloud computing environments. In: Proc. of the 3rd IEEE Int'l Conf. on Cloud Computing Technology and Science. Washington: IEEE Computer Society, 2011. 407−414. [doi: 10.1109/CloudCom.2011.61]

[9] Zhang XY, Ai J, Wang ZY, Lu JH, Meng XF. An efficient multidimensional index for cloud data management. In: Proc. of the 1st Int'l Workshop on Cloud Data Management. New York: ACM, 2009. 17−24. [doi: 10.1145/1651263.1651267]

[10] Ding LL, Qiao BY, Wang GR, Chen C. An efficient quad-tree based index structure for cloud data management. In: Proc. of the 12th Int'l Conf. on Web-Age Information Management. Berlin: Springer-Verlag, 2010. 238−250. [doi: 10.1007/978-3-642-23535-1_22]

[11] Chen G, Tam VH, Wu S, Ooi BC, Özsu MT. A framework for supporting DBMS-like indexes in the cloud. Proc. of the VLDB Endowment, 2011,4(11):702−713.

[12] Huang B, Peng YX. An efficient two-level bitmap index for cloud data management. In: Proc. of the 3rd IEEE Int'l Conf. on Communication Software and Networks. Washington: IEEE Computer Society, 2011. 509−513. [doi: 10.1109/ICCSN.2011.6014776]

[13] Aguilera MK, Golab WM, Shah MA. A practical scalable distributed B-tree. Proc. of the VLDB Endowment, 2008,1(1):598−609. [doi: 10.14778/1453856.1453922]

[14] Meng BP, Wang TJ, Li HY, Yang DQ. Regional bitmap index: A secondary index for data management in cloud computing environment. Chinese Journal of Computers, 2012,35(11):2306−2316 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2012.02306]

[15] Abouzeid A, Pawlikowski KB, Abadi DJ, Silberschatz A, Rasin A. HadoopDB: An architectural hybrid of MapReduce and DBMS technologies for analytical workloads. Proc. of the VLDB Endowment, 2009,2(1):922−933. [doi: 10.14778/1687627.1687731]

[16] Dittrich J, Quiané-Ruiz JA, Richter S, Schuh S, Jindal A, Schad J. Only aggressive elephants are fast elephants. Proc. of the VLDB Endowment, 2012,5(11):1591−1602. [doi: 10.14778/2350229.2350272]

[17] Dittrich J, Quiané-Ruiz JA, Jindal A, Kargin Y, Setty V, Schad J. Hadoop++: Making a yellow elephant run like a cheetah. Proc. of the VLDB Endowment, 2010,3(1):518−529. [doi: 10.14778/1920841.1920908]

[18] Wu S, Wu KL. An indexing framework for efficient retrieval on the cloud. IEEE Data Engineering Bulletin, 2009,32(1):75−82.

[19] Lin CH, Yu JQ, He YF, Guan T, Ai LF. High-Dimensional distributed indexing based on locality-sensitive hashing. Journal of Frontiers of Computer Science and Technology, 2013,7(9):811−818 (in Chinese with English abstract). [doi: 10.3778/j.issn.1673-9418.1303047]

[20] Jagadish HV, Ooi BC, Quang Hieu Vu. BATON: A balanced tree structure for peer-to-peer networks. In: Böhm K, Jensen CS, Haas LM, Kersten ML, Larson P, Ooi BC, eds. Proc. of the 31st Int'l Conf. on Very Large Data Bases. New York: ACM, 2005. 661−672.

[21] Stoica I, Morris R, Karger DR, Kaashoek MF, Balakrishnan H. Chord: A scalable peer-to-peer lookup service for Internet applications. In: Proc. of the ACM SIGCOMM. New York: ACM, 2001. 149−160. [doi: 10.1145/383059.383071]

[22] Bentley JL. Multidimensional binary search trees in database applications. IEEE Trans. on Software Engineering, 1979,5(4):333−340. [doi: 10.1109/TSE.1979.234200]

[23] Cai HL, Ge P, Wang J. Applications of Bloom filters in peer-to-peer systems: Issues and questions. In: Proc. of the Int'l Conf. on Networking, Architecture, and Storage. Washington: IEEE Computer Society, 2008. 97−103. [doi: 10.1109/NAS.2008.52]

© 中国科学院软件研究所 http://www.jos.org.cn

[24] Ratnasamy S, Francis P, Handley M, Karp R, Shenker S. A scalable content-addressable network. In: Proc. of the 2001 Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communications. New York: ACM, 2001. 161−172. [doi: 10.1145/383059.383072]

[25] Zhang C, Li FF, Jestes J. Efficient parallel *k*NN joins for large data in MapReduce. In: Proc. of the 15th Int'l Conf. on Extending Database Technology. New York: ACM Press, 2012. 38−49. [doi: 10.1145/2247596.2247602]

[26] Vernica R, Carey MJ, Li C. Efficient parallel set-similarity joins using MapReduce. In: Proc. of the ACM SIGMOD/PODS Conf. New York: ACM Press, 2010. 495−506. [doi: 10.1145/1807167.1807222]

[27] Chaudhuri S, Ganti V, Kaushik R. A primitive operator for similarity joins in data cleaning. In: Proc. of the 22nd Int'l Conf. on Data Engineering. Washington: IEEE Computer Society, 2006. 5 [doi: 10.1109/ICDE.2006.9]

[28] Sun SL, Zhou D, Qian L. High performance query technique of cloud data warehouse. Design of Posts and Telecommunications Technology, 2011,428(10):23−26 (in Chinese with English abstract).

[29] Agrawal P, Silberstein A, Cooper BF. Asynchronous view maintenance for VLSD databases. In: Proc. of the 2009 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM, 2009. 179−192. [doi: 10.1145/1559845.1559866]

[30] Zou YQ, Liu J, Wang SC, Zha L, Xu ZW. CCIndex: A complemental clustering index on distributed ordered tables for multi-dimensional range queries. In: Proc. of the 7th IFIP Int'l Conf. on Network and Parallel Computing. Berlin: Springer-Verlag, 2010. 247−261. [doi: 10.1007/978-3-642-15672-4_22]

[31] Nishimura S, Das S, Agrawal D, Abbadi AE. MD-HBase: A scalable multi-dimensional data infrastructure for location aware services. In: Proc. of the 11th Int'l Conf. on Mobile Data Management. Washington: IEEE, 2011. 7−16. [doi: 10.1109/mdm.2011.41]

[32] Ma Q, Yang B, Qian WN, Zhou AY. Query processing of massive trajectory data based on MapReduce. In: Proc. of the 1st Int'l Workshop on Cloud Data Management. New York: ACM, 2009. 9−16. [doi: 10.1145/1651263.1651266]

[33] Chu ST, Yeh CC, Huang CL. A cloud-based trajectory index scheme. In: Proc. of the IEEE Int'l Conf. on E-Business Engineering. Washington: IEEE Computer Society, 2009. 602−607. [doi: 10.1109/ICEBE.2009.95]

[34] Li QQ, Zhang T, Yu Y. Using cloud computing to process intensive floating car data for urban traffic surveillance. Int'l Journal of Geographical Information Science, 2011,25(8):1303−1322. [doi: 10.1080/13658816.2011.577746]

[35] Cary A, Sun ZG, Hristidis V, Rish N. Experiences on processing spatial data with MapReduce. In: Proc. of the 21st Int'l Conf. on Scientific and Statistical Database Management. Berlin: Springer-Verlag, 2009. 302−319. [doi: 10.1007/978-3-642-02279-1_24]

[36] Hsu YT, Pan YC, Wei LY, Peng WC, Lee WC. Key formulation schemes for spatial index in cloud data managements. In: Proc. of the 13th IEEE Conf. on Mobile Data Management. Washington: IEEE Computer Society, 2012. 21−26. [doi: 10.1109/MDM.2012.67]

[37] Cui JF, Zhang Y, Li C, Xing CX. Overview of spatial index in the cloud storage. Journal on Communications, 2011, 32(9A):34−41 (in Chinese with English abstract). [doi: 1000-436X(2011)9A-0034-08]

[38] Yu G, Gu Y, Bao YB, Wang ZG. Large scale graph data processing on cloud computing environment. Chinese Journal of Computers, 2011,34(10):1753−1767 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2011.01753]

[39] Ma YZ, Rao J, Hu WS, Meng XF, Han X, Zhang Y, Chai YP, Liu CQ. An efficient index for massive IOT data in cloud environment. In: Proc. of the 21st ACM Conf. on Information and Knowledge Management. New York: ACM, 2012. 2129−2133. [doi: 10.1145/2396761.2398587]

[40] Diana M, Denis S, Gylfi G, Laurent A. Indexing and searching 100M images with map-reduce. In: Proc. of the 3rd ACM Conf. on Int'l Conf. on Multimedia Retrieval. New York: ACM, 2013. 17−24. [doi: 10.1145/2461466.2461470]

[41] Mark HB, James R. Distributed Lucene: A distributed free text index for Hadoop. HP Laboratories, HPL-2008-64, 2008, http://www.hpl.hp.com/techreports/2008/HPL-2008-64.html

[42] Lu XL, He JM. Study on cloud storage model of Map/Reduce-based index data. Journal of Ningbo University, 2011,24(3):29−33 (in Chinese with English abstract). [doi: 10.3969/j.issn.1001-5132.2011.03.007]

[43] Jimmy L, Dmitriy R, Kevin W. Full-Text indexing for optimizing selection operations in large-scale data analytics. In: Proc. of the 2nd Int'l Workshop on MapReduce and Its Applications. New York: ACM, 2011. 59−66. [doi: 10.1145/1996092.1996105]

© 中国科学院软件研究所 http://www.jos.org.cn

[44] Li X, Wang X, Shi H, Sheng ZH, Feng ZY. A distributed inverted indexing scheme for large-scale RDF data. In: Proc. of the 13th Int'l Conf. on Web-Age Information Management Workshops. Berlin: Springer-Verlag, 2012. 151−161. [doi: 10.1007/978-3-642-33050-6_16]

[45] Marcel K, Kai-Uwe S, Manfred H. Scalable distributed indexing and query processing over linked data. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 2012,10:3−32. [doi: 10.1016/j.websem.2011.11.010]

[46] Peter M, Barcelona S. Distributed indexing for semantic search. In: Proc. of the 3rd Int'l Semantic Search Workshop. New York: ACM, 2010. 1−4. [doi: 10.1145/1863879.1863882]

[47] Qin XP, Wang HJ, Du XY, Wang S. Big data analysis-competition between RDBMS and MapReduce, their inter-fusing and symbiosis. Ruan Jian Xue Bao/Journal of Software, 2012,23(1):32−45 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4091.htm [doi: 10.3724/SP.J.1001.2012.04091]

[48] He BS, Fang WB, Luo Q, Govindaraju NK, Wang TY. Mars: A MapReduce framework on graphics processors. In: Proc. of the 7th Int'l Conf. on Parallel Architectures and Compilation Techniques. New York: ACM, 2008. 260−269. [doi: 10.1145/1454115.1454152]

[49] Sandholm T, Lai K. MapReduce optimization using regulated dynamic prioritization. In: Proc. of the 11th Int'l Joint Conf. on Measurement and Modeling of Computer Systems. New York: ACM, 2009. 299−310. [doi: 10.1145/2492101.1555384]

[50] Yang HC, Dasdan A, Hsiao RL, Parker DS. Map-Reduce-Merge: Simplified relational data processing on large clusters. In: Proc. of the ACM SIGMOD/PODS Conf. New York: ACM, 2007. 1029−1040. [doi: 10.1145/1247480.1247602]

附中文参考文献:

[14] 孟必平,王腾蛟,李红燕,杨冬青.分片位图索引:一种适用于云数据管理的辅助索引机制.计算机学报,2012,35(11):2306−2316.

[19] 林朝晖,于俊清,何云峰,管涛,艾列富.高维分布式局部敏感哈希索引方法.计算机科学与探索,2013,7(9):811−818.

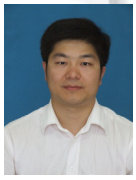
[28] 孙少陵,周大,钱岭.云数据仓库高性能查询技术研究.邮电设计技术,2011,428(10):23−26.

[37] 崔纪锋,张勇,李超,邢春晓.面向云存储的空间索引技术研究.通信学报,2011,31(9A):34−41.

[38] 于戈,谷峪,鲍玉斌,王志刚.云计算环境下的大规模图数据处理技术.计算机学报,2011,34(10):1753−1767.

[42] 陆小丽,何加铭.基于 Map/Reduce 的索引数据云存储模型研究.宁波大学学报(理工版),2011,24(3):29−33.

[47] 覃雄派,王会举,杜小勇,王珊.大数据分析——RDBMS 与 MapReduce 的竞争与共生.软件学报,2012,23(1):32−45. http://www.jos.org.cn/1000-9825/4091.htm [doi: 10.3724/SP.J.1001.2012.04091]



马友忠(1981－),男,博士生,CCF 会员,主要研究领域为 Web 数据管理,云数据管理.

孟小峰(1964－),男,博士,教授,博士生导师,CCF 会士,主要研究领域为网络数据管理,云数据管理,移动数据管理,XML 数据管理,闪存数据库系统,隐私保护.

© 中国科学院软件研究所 http://www.jos.org.cn