

行列混合存储的数据压缩策略思考

徐高省

北京合数信息技术有限公司 北京 100000

摘要: 随着当今信息时代和发展方式的变化逐渐推向深入, 现代人类社会每天无时无刻地在自动产出着大量新的生活数据, 并且由于大部分数据仍要继续保留使用一段时间, 数据压缩存储新技术亦应时而生, 但实际上随着数据种类的数量不断在增多, 数据可存储的容量自然也会达到了它一定程度的最低限度, 因此人们需要学会用一种数据压缩保存的有效方式来有效减少生活数据可存储容量, 降低查询处理对存储系统的消耗。目前数据压缩存储的方式主要为行存储和列存储, 本文采用行列存储的方式来对用户的数据按照行组织列压缩方式进行研究探讨, 并根据数据分布的特点选用合适的编码规则进行压缩, 提高数据压缩率。

关键词: 行列混合存储; 数据压缩; 动态编码; 字典压缩

Data compression strategy for column and column mixed storage

Xu Gaosheng

Beijing Composite Information Technology Co., LTD. Beijing 100000

Abstract: with the development of today's information age and changes in the way gradually to the deep, the modern human society all the time in automatic output with a large number of new data of life, and because most of the data are still want to use for a period of time to retain, data compression storage new technology should also be born from time to time, but in fact along with the continuously growing the number of data types, data can be stored has a capacity of nature Therefore, people need to learn to use an effective way of data compression and preservation to effectively reduce the storage capacity of daily data and reduce the consumption of query processing on the storage system. At present, the way of data compression storage is mainly row storage and column storage. In this paper, the way of row storage is used to study and discuss the compression mode of user's data according to row organization and column, and select appropriate coding rules for compression according to the characteristics of data distribution to improve the data compression rate.

Keywords: column and column mixed storage; Data compression; Dynamic coding; The dictionary compression

引言:

在当前的社会生活中, 大数据及其应用如今已经逐渐全面地融入了在人类了社会生活的其中和工作当中的各方方面面, 占据起了其主导地位, 数据资料库的基本数据及存储信息类型等也都将随之从原来的传统上的基于数据的关系型分析的数据库也将相应转变而为分析型数据库, 并在同时也为了能够进一步地减少了数据库的

本身的数据的存储容量, 采用数据压缩原理的控制方式可以对整个系统来进行实时控制。数据的压缩控制原理其主要功能作用就是指为了在达到尽可能多不应产生任何丢失问题的原始数据质量水平的前提下来能进一步有效缩减系统原始的数据量, 按照一种预先设计一定的程序的数据压缩控制算法对数据内容进行了适当地重新计算进行优化组织, 并期望由此能够提高对系统的实时传输、存储等质量管理水平和故障分析等处理工作效率。在我们整个乃至当前几乎所有领域的所谓现代信息计算机科学过程当中, 数据信息的编码压缩处理技术均强调必须编码是指信息按照基于其内部特定程序的编码压缩

作者简介: 徐高省, 男, 汉, 1979.9.18, 硕士研究生, 江苏, CTO, 从事的工作: 大数据智能营销, 大数据, 人工智能, 高并发处理等。

机制来用一个原始的数据位元编码形式来进行表示的信息，目前流行使用zip格式。在数据压缩当中需要对数据压缩机制的压缩能力、失真度和计算资源等因素进行考虑，确保数据的保真度。采用行列混合的方式对数据进行存储能够减少磁盘的读取量，提高查询效率。

一、行列混合存储概述

1.1 行存储

在人们最常用查询到数据库的各种数据库及管理类系统当中，mysql数据库与数据库及oracle数据库等都是属于关系型数据库类型的串行数据信息存储类型的数据库，是中国目前国内一种相对的较为传统的串行数据库信息的存储形式，这类存储结构对于传统的数据工作较为有效。行存储，顾名思义，是按行来进行存储，其布局与表格的数据表示方式相似，每一行都具有相同字段的集合，能够有效存储用户条目，适用于由多个字段组成且由某个键唯一表示的数据记录。在传统数据库当中按照大数据行式结构来进行存储与访问的大数据行当中尤其是，行结构的数据库存储和访问起来最为便捷快速而有效，且这样也的确能够显著大幅地提高了数据的存储及访问空间结构中存在的数据库局部性，但是在由于大多数采用的传统数据库结构的大数据行式存储数据库的结构体系当中，将这样一个数据相对独立完整的大数根据行存储在同一个大数据页面当中，在对用户数据进行大行数据量大的数据查询的操作时比较为容易的就已经出现在了在数据库并没有完全建立索引视图的情况下在进行查询处理时需要耗费了大量的数据IO时间的情况这种特殊情况以及数据库需要在建立了索引视图和物化视图的特殊情况下数据库才能更加快速正确地完成定位任务行列情况但又由于要花费数据的时间过长。不仅如此，目前移动互联网及物联网等业务中的数据主要分为稠密数据于稀疏数据，按照行存储结构是数据压缩比较低，容易造成较大存储空间的浪费，并且行存储无法按照数据特征来对数据进行针对性的存储选择，无法时限混合高效存储。虽然并行存储能够极大提高输入数据存储的效率，对数据可进行快速加载以及高负载，但如果对于多行的存储属性进行数据查询加载时则必须同时读取一整行条目的数据记录，当存储属性数量较多时，查询与读取数据时的时间代价系数也必然会相对随着增大，智能满足读优化或写优化的其中一个，而不能两者兼备，查询效率不高。

1.2 列存储

列存储与行存储的区别就在于列存储中数据时按照

列来进行连续存储的，与行存储相比其压缩算法更加高效简单，压缩率也较高，其磁盘存储会被更加高效地利用。在SQL Server数据库中，列存储索引的这种方式还能够大幅改进查询输入的响应时间，这仅仅是因为用户在查询数据库中输入索引时只需一次读取所需要输入的索引列，从数据库内存中转移存储到数据库处理器的重缓存中的数据量也便能因此而得到减少，并且索引列本身经过进行了一次对数据的高度的压缩，能够显著减少数据读取次数与数据移动中的字节数，减少总IO。列存储式数据库原理是指数据库会将其中包含在同一个数据列上信息中的全部各个数据值都存放在一起，在读数据或是查询其它数据时也就都不需要去一次性地读取查询到其中所有数据相同值的任意一个数据列，而行存储于数据库内的即是指数据库会在读取与查询信息时所需的一次性读取其中的所有全部值相同数据的各个数据列，大大提高了读取效率。但数据列存储过程本身也有这样一些其的本身存在缺陷，当你只需要单独地读取数据到其中某的一个数据行时，需要由你先从不同数据的同一存放的地方开始同时分别对在其中的各个数据列内存放的一个数据值同时地进行一次单独的读取之后再分别读取合并数据为行，在这种情况下，需要整行的数据时，列存储便不适用了。列存储的方式更加适合计算聚合的分析型工作负载。在存储访问的数据方面，列存储比行式存储方式更简单加快捷，并且还利于对数据进行压缩，但是由于写数据进度较慢，在写数据文件的过程同时仍需数据进行在解压处理和数据在压缩，代价就较大。同时，采用属性列分开存储操作时还需再进行实物画框操作，将已分开存储好的属性列再重新整理组织，按照操作模式来进行物化，其中延迟物化能够获得较好的查询效率^[1]。

1.3 行列混合存储

在当前的数字爆炸时代下，单纯的运用行存储或列存储对数据建立数据库的形式已经不能满足数据处理的需求，并且在大量数据存储的情况下，查询效率无法达到工作要求，因此国内逐渐开始对行列混合存储的方式进行研究探讨。行列混合模式不仅能够实现实体的行式存储与列式存储，还能够是在以内部对不同数据的疏密程度等特征来选择合适的存储格式，实现动态存储，从而获得行列混合存储模式。主要方式在于通过分层、可配置的存储结构在经过大数据筛选处理之后，将存储结构的层次进行划分与构建，基于Key-Value键值模式来对数据定义层进行定义。在元数据当中，将其在逻辑层

面表现为表格的形式,进行数据约束,按照键值模型来以分层的形式对数据结构进行描述,融入模式定义将数据可视化,在Key-Value的基础之上将各种混合数据中的每一个键、值来对应其实体属性值,将其组织为不同属性的数据子集。基于此,将数据子集存储分别地定义为数据按行顺序存储数据的格式与信息按行列顺序存储数据的存储格式,稠密的数据应采用行顺序存储,稀疏的数据应该采用列存储^[2]。行列混合存储的方式能够充分发挥行存储与列存储的优势,实现对数据的多层管理模式,拥有较好的压缩性与可配置能力,对数据进行大范围多方向的索引管理与存储。

二、数据压缩概述

数据压缩技术目前应用在大数据信息存储等问题解决中也已越来越被业界广泛所讨论关注与深入研究,是另一项用于解决复杂大数据的存储分析与数据查询需求的最有效实现方式,一般指用尽量的少的数码信息来表示数据信源时所需发出去的信号,减少数据集合。被压缩的对象通常为数据集合所占用的时域、空域与频域。为了能方便用户将数据文件进行数据压缩,最基础简单最为主要的几个工作步骤中之一的就是将数据压缩进行信源编码,LZ的数据无损压缩解码处理方法同样也是当前国内市场目前最流行与使用最好的一种数据的无损存储算法之一,DEFLATE作为LZ算法中的另一种变体对数据文件中的数据无损解压的编码速度与数据解压缩率上分别也进行的做的了优化。LZ的数据压缩的编码的方法同样也是采用的由用户将最初所能输入接收到的数据动态生成的表格做为数据进行压缩编码的模型并进行了霍夫曼编码。在数据过程的压缩可视化工作模式当中,需要分别进行的一个是压缩的数据过程本身与另外一个就是解压数据的操作过程,对于这样一个数据压缩可视化的操作过程的本身来说,可选择直接进行的离线的压缩数据或并行压缩的可视化操作,而对于数据解压可视化这个数据过程自身来说则可以只做需要的速度的较快的一些的压缩可视化数据,能够的做到一个直接进行的数据基于压。

三、行列混合存储结构的数据压缩

3.1 叠加式混合存储动态编码数据压缩

在考虑运用行列的混合存储模型结构方法进行数据压缩设计时,首先我们需考虑明确行列的混合存储结构模型方法中使用的存储引擎与数据查询引擎。存储引擎指在一个数据库系统当中功能主要的在于用来完成对数据资料的输入或储存,查询引擎则主要指能够在这个数

据库系统当中直接执行数据查询工作行动的一个部件,通过直接访问这些存储查询引擎以自动获得相关数据信息来方便进行数据库查询相关工作。在通过对行存储、列存储及数据压缩进行一定的了解之后,明确了行存储与列存储在数据存储与查询等方面的优缺点,于是本文提出应用两种存储方式的优点,并对缺点进行弥补,从而提出叠加式混合存储方式。

3.2 基于行列混合存储应用字典编码深度压缩

字典编码压缩方式技术是一个目前普遍应用于在数据压缩工作中使用的一种基础编码技术,能够用于对数据文本进行字典快速地定位并进行编码,实现文本快速地压缩,在采用行列混合存储方式中应用字典编码的方式能够高效地位为上层数据库进行快速压缩,并在一定压缩率下加快解压速度。在进行字典编码实现数据压缩之前,首先需要对数据进行行列混合存储。对于需要进行压缩的数据,建立数据库系统,根据数据库中的表来选择按照行或者列来进行存储,每个表在数据库中存储一份,通过分层、可配置结构来进行行列混合存储。在通过行列混合的方式导入数据之后,应用压缩缓冲区来导入数据,将一定行数据作为独立的数据包,然后在数据包内结合数据表的形式来将数据按照属性以列的方式进行抽取分割并存储。由此便根据属性的不同在行列混合存储库内进行了分类。建立数据包之后,对数据按照字典编码来构建结构与权重表。对于每个属性列的数据,按照字典规则来辅助建立数据结构,其中字典表采用静态哈希结构,在属性列当中根据属性值来配对字典表中的哈希值,由此来找出对应的权重值。对于数据行,则将数据行中的每个数据属性按照属性列进行分配与对应,设置入字典编码结构当中,获取属性值。然后压缩缓冲区便可将属性值替换为字典表条目的编号,存储在属性列的引用表中。在得到属性值之后可预先在无压缩的状态根据字典表中的条目来预估数据进行字典编码压缩之后的存储大小。在完成字典编码结构之后,对字典表内按照属性排列的数据类型进行升序排序重新填入,同时更新引用表。对属性列按照升序排序后,为属性列建立列间规则来发掘辅助结构。辅助结构可通过阅读每个不同属性列中的引用表,来根据引用表内的行号及字典项重构出字典表。将列间规则作为编码条件对数据进行判断,按照权重来检查当前的字典表,当发现属性列超过了列内规则编码缩小后的一半,则重新进行辅助结构的发掘。当所数据都能够经过规则编码之后,便可进行压缩工作。对每个属性列使用LAOP方法进行压缩

^[5], 并且用过压缩算法在规则编码压缩数据的基础上进行深度压缩, 进一步降低数据的存储空间。

3.3 行列存储数据查询

在运用行列混合存储的数据压缩方法中, 对于数据的查询执行流程, 由于在对行列存储模型进行建立时, 便将元数据输入了数据字典, 因此在查询阶段, 首先还需对数据进行解压缩, 获取数据的存储模式, 再根据查询所得信息生成数据的访问参数。在行列的混合存储结构当中, 为数据文件的解压缩过程提供出了一个单独的处理区域, 当进行查询时, 区域便会被激活, 对所要求压缩后的控制数据进行加工处理, 将这些控制参数信息直接转换输出为实际用户需求中的实际数据。一般解压缩的过程为数据抽值与列值扫描^[4]。根据查询所生成的访问参数会被传入存储区域内, 然后再进行读取与选择数据之后输出, 主要通过SELECT语句查询分析算法, 获取目标属性列表与选择条件列表, 再按照行数据库来执行查询计划。在查询计划中可采用协同定位的方式, 对行数据与列数据进行交叉同步查询, 将结果及时反馈至数据分析查询层, 缩短数据查询时间。

四、运行效果检测

在对行列混合存储的数据压缩方式进行探讨后, 对其运行效果进行了检测。从数据存储量方面, 将其与单一行数据压缩或列存储数据压缩及应用行列混合存储的大数据存储的运行效果进行比较。采用单一行数据压缩或单一列数据压缩方式的存储方式在数据存储量方面比应用行列混合存储的大数据存储方式要少, 比应用行列混合存储的数据压缩存储方式需要的数据存储量也更少。不仅如此, 采用行列混合存储技术的数据压缩方式的数据压缩比均在80%以上至大于100%以上之间, 压缩比效率高。在数据的压缩和时间方面, 行列和混合并行数据的存储处理过程中进行的数据的数据压缩和处理的时间往往都比对于单一行的数据或对多行列数据并行数据存储过程时进行的数据处理或压缩计算的处理时间也通常是要得比较短。在对于单一行混合数据所存储时的所有数据信息进行数据压缩时的方式当中, 数据记录被压

缩输入处理完成后一般是需要先再对数据存储时的所有的混合数据记录信息再进行一个比较全面的压缩获取的过程分析及压缩结果再判定, 处理这个过程时间往往也都较长, 而对同一行列的混合数据所存储时的所有的混合数据记录在进行压缩获取处理这个方式过程分析中则一般需要对数据记录信息的整个压缩获取处理整个过程时间相对要都较为短, 并且同一行、列信息也均可同时进行压缩和对数据记录的处理。在提高数据解压缩比方面, 行列混合存储的数据压缩的方式由于为数据解压缩比提供了出了一个可以单独压缩的数据处理的空间, 因此处理时间较短, 但列存储的方式按照列结构对数据进行属性分类, 解压缩时间更短, 行存储则可能由于需要对所压缩的数据内容进行相当长时间的压缩处理, 因此其解压缩处理时间通常较长。

五、结语

综上所述, 在对行列混合存储的数据压缩方式进行思考研究的过程中, 对行存储、列存储及数据压缩等概念充分理解之后, 结合行了数据存储方式与列数据存储形式的比较优势来合理设计数据存储压缩的方式, 提出采用动态编码进行数据压缩, 运用字典编码的方式来进行深度压缩, 并在实验过程中对数据压缩效果进行测验, 发现采用行列混合存储的数据压缩查询方式能够帮助在有效提高查询数据压缩率的同时快速提高数据查询效率, 可对大量且复杂的数据进行处理分析, 很好地解决了大数据的存储问题, 具有较高的可行性。

参考文献:

- [1]孙林超, 陈群, 肖玉泽, 等.行列混合存储数据库系统的研究[J].计算机应用研究, 2013, 30(2): 480-482, 486.
- [2]安洋, 赵洪松.基于行列混合存储的大数据存储方法研究与实现[J].通信管理技术, 2014(1): 24-27.
- [3]吴向阳, 周洋全, 计忠平.体数据压缩技术综述[J].计算机辅助设计与图形学学报, 2015(1): 26-35.
- [4]魏玲, 郭新朋.行列混合存储的数据压缩策略研究[J].小型微型计算机系统, 2017, 38(6): 1267-1272.