文章编号:1673-9469(2014)04-0069-05 doi:10.3969/j.issn.1673-9469.2014.04.017

分布式系统下大数据存储结构优化研究

冯汉超,周凯东
(河北工程大学 信息与电气工程学院,河北 邯郸 056038)

摘要: 在分布式系统中,数据的存储结构直接影响了大数据的存储效率和处理性能。在行式存储结构下,数据从本地读取,加载速度快,但压缩效率低且存在数据冗余;在列式存储结构下,数据压缩效率高,但数据的跨节点访问增加了网络传输消耗。针对行式存储结构和列式存储结构的缺点,提出一种以行列结合的存储方式,对数据存储结构进行改进。实验结果表明,改进的数据存储结构在加载速度上略低于行式存储;在数据压缩上,比行式存储和列式存储的效率都高。行列结合的存储结构不仅避免行式存储的额外磁盘I/O开销,同时也减少了列式存储不必要的网络传输,极大地提高分布式系统对大数据存储效率及处理性能。

关键词: 大数据; 分布式; 行列存储
中图分类号: TP3 **文献标识码:** A

Research on optimizing big data storage structure in distributed system

FENG Han-chao, ZHOU Kai-dong
(School of Information & Electrical Engineering, Hebei University of Engineering, Hebei Handan 056038, China)

Abstract: In a distributed system, the data storage structure directly affects the storage efficiency and processing performance of big data. In the row store structure, the data is loaded locally and the speed is fast, but it also loads additional columns, and it's hard to compress. The column store structure has high compression efficiency, but it has additional network transferring overhead. To overcome their storages and improve the data storage structure, this paper presents a new data storage structure combining row and column. The experiment result shows that it's inferior a little in data loading to the row store structure, and it has high compression efficiency comparing with the row store structure and column store structure. It not only avoids additional disk I/O, but also cuts down the unnecessary network transfer time in column store. So, the row-column store can greatly improve big data storage and processing performance in distributed system.

Key words: big data; distributed system; row-column store

当前社会已经进入数据爆炸的时代,海量数据的处理与分析被称为"大数据"[1]。全世界亿万用户通过互联网相互联系,随之产生的数据也在高速增长。在互联网领域,与大数据有关的应用已变的非常重要。由于传统关系型数据库在管理大数据时遇到种种困难和阻碍,基于分布式的海量数据管理系统成了当前的研究热点。

Hadoop[2]是MapReduce[3]分布式计算框架的实现,为大数据处理提供可扩展、高容错性的大型分布式集群。基于MapReduce的数据仓库[4]并不能直接管理集群中的数据存储,而是由Hadoop分布式存储系统HDFS(Hadoop Distributed File System)[5-7]来管理海量数据。如何在HDFS中设计一个高效的数据存储结构来组织大数据遇到了一系列的困难,而影响数据仓库性能的关键因素是能够满足充分利用MapReduce计算特性来处理大数据的数据存储结构。本文在分析分布式存储系统HDFS局限性基础上,通过改变数据存储结构对大数据存储和处理作出了一定改进。

收稿日期:2014-06-10
作者简介:冯汉超(1990-),男,山东潍坊人,硕士,从事大数据存储与处理方面的研究。

1 大数据的特点及处理要求

1.1 大数据特点

大数据(big data),或称巨量资料,指的是所涉及的资料量规模无法通过目前主流软件工具,在合理时间内撷取、管理、处理。大数据主要特点有:数据量大、数据类型丰富、处理速度快、计算精确。

1.2 大数据处理要求

高速的数据加载。数据的高速加载是大数据处理中的一个关键问题,例如 Facebook 每天要有 20TB 的数据存放到数据仓库中。在正常的查询时,由于网络带宽、磁盘 I/O 在数据传输时的资源瓶颈,缩短数据加载时间变得非常关键。

高速的查询处理。为满足大量用户同时向系统提交实时请求和高负载查询,要求底层数据存储结构在满足数据不断增长的同时,能够高效处理查询请求。

存储空间的高利用率。不断增长的互联网用户,导致全球数据急剧增长。这就要求系统在存储上有很好的扩展能力。如何存放数据使磁盘利用率达到最高是关键问题。

适应动态高负载模式。对于不同的应用程序,用户分析大数据集的方式不尽相同。虽然一些数据分析以静态模式周期执行,但大部分数据分析不遵循任何的常规模式。因此,需要系统在有限的存储空间下使用不可预测的数据分析请求,而不是在特定的模式下运行。

2 现有数据存储结构的优缺点分析

2.1 行式存储结构优缺点分析

行式存储结构[3]是传统关系型数据库存储结构,记录以行的形式存储在数据库关系表中。在添加行时,记录中所有的列都需要存储,且记录被连续的存储到磁盘的页块中。在分布式系统存储下,表按行水平分割,每行中所有数据存放在同一个 HDFS 块中。图 1 表示以行式存储的数据结构在 HDFS 块中的分布。

行式存储在 Hadoop 集群 DataNode[5] 节点间的数据分布分析。以行式结构存储时,每行中所有的列存放在同一个 HDFS 块中。在分布式系统 HDFS 中,大表中的数据按行水平分割,分割后每组数据可能分布在不同的 DataNode 节点上。行式存储在 Hadoop 集群 DataNode 节点间的数据分布结构图如图 2 所示。

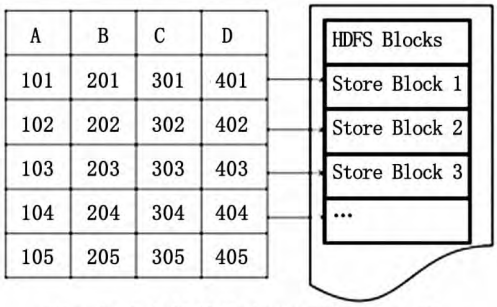

图1 行式存储结构在HDFS块的分布图

Fig. 1 The layout of row store structure among HDFS blocks

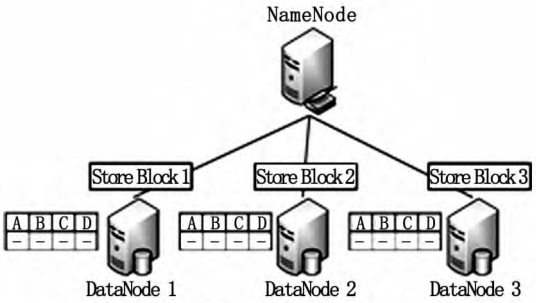

图2 行式存储结构在DataNode间的分布图

Fig. 2 The layout of row store structure among DataNodes

行式存储时数据读取操作分析。若读取行中 A 和 C 两列,首先读取本地 DataNode 节点上所有符合条件的行,然后从每行中选择 A 和 C 两列数据,过滤掉不需要的 B 和 D 列。行式数据读取操作示意图如图 3 所示。

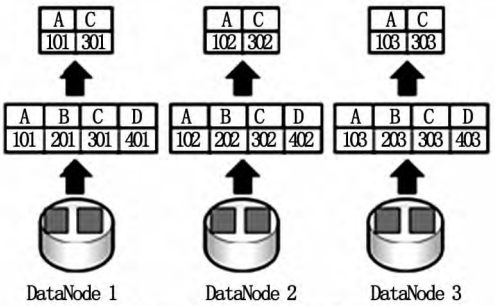

图3 行式存储结构在数据读取操作示意图

Fig. 3 The schematic diagram of read operation about row store

行式存储结构优点和缺点分析。数据加载速度快,所有数据都从本地读取,无需额外的网络带宽消耗。但是,每行中所有列都存放在相同的

HDFS 块中,在数据读取一行数据时,行中所有列都需从磁盘上读取,不需要的列也会被读取,这样会额外增加磁盘 I/O 开销。

每列存储的数据类型不能一样,在数据压缩时不同数据类型压缩效果很差,这样导致磁盘空间利用率低,同时也加大了磁盘 I/O 开销。

2.2 列式存储结构优缺点分析

列式存储结构[9-10]是将关系表按列垂直分割成多个子关系表,分割后的每组子关系表中的所有数据存放在同一个 HDFS 块中,每一列都独立存储。列式存储的数据结构在 HDFS 块中的分布如图 4 所示。

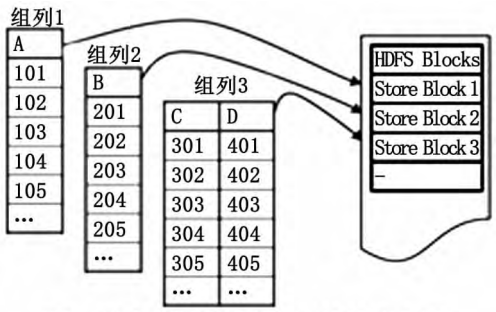

图4 列式存储结构在HDFS块的分布图

Fig. 4 The layout of column store structure among HDFS blocks

列式存储在 Hadoop 集群 DataNode 节点间的数据分布分析。以列式结构存储时,每列中所有数据存放在同一个 HDFS 块中。在分布式系统 HDFS 中,大表中的数据按列垂直分割,分割后每组数据可能分布在不同的 DataNode 节点上。列式存储在 Hadoop 集群 DataNode 节点间的数据分布结构图如图 5 所示。

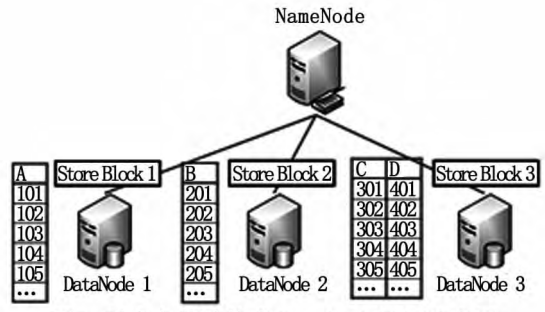

图5 列式存储结构在DataNode间的分布图

Fig. 5 The layout of column store structure among DataNodes

列式存储时数据读取操作分析。若读取行中 A 和 C 两列,A 和 C 列分别存放在两个不同的节点上,首先从 DataNode1 上读取 A 列所有数据,再从 DataNode3 上读取 C 列数据,最后通过网络将数据传输到同一个机器上。列式数据读取操作示意图如图 6 所示。

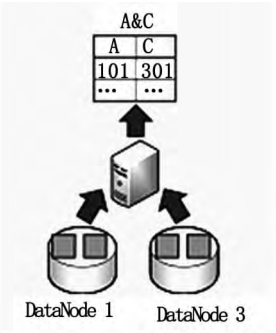

图6 列式存储结构在数据读取操作示意图

Fig. 6 The schematic diagram of read operation about column store

列式存储结构优点和缺点分析。列式存储只读取有用的列,能够避免额外的磁盘 I/O 开销。同时,同一列中的数据类型相同,在数据压缩时有很好的压缩比,提高磁盘的空间利用率。但是,列式存储是按列垂直分割,不同的列可能分布在不同的数据节点上,读取不同列的数据会跨节点访问,增加了网络传输所消耗的时间。

2.3 分布式系统中数据存储结构的最优化分析

通过以上行式存储和列式存储结构的优缺点分析,行式存储主要是在读取不必要的列消耗了额外的磁盘 I/O,列式存储主要是在跨节点查询数据时消耗额外的网络传输时间。最理想的情况是能够避免额外的磁盘 I/O 消耗和网络传输消耗。设从 n 行数据中读取 i 列有效数据,其行列存储结构在数据读取和网络消耗上比较如表 1 所示。

表 1 行式存储与列式存储读取效率比较

Tab. 1 The efficiency of row store and column store

	行式存储	列式存储	最优结构
数据读取	n	i/n	i/n
网络消耗	0	β (0% ≤β≤100%)	0

3 对现有数据存储结构的改进

3.1 行列结合的方式存储结构分析

分别结合行式存储和列式存储节点的优点,将关系表中的数据先按水平划分成多个行组(Row Group),每一个行组存放在同一个 HDFS 块中。在每一个行组内,将表按列垂直划分多个子关系表,每一列在进行数据压缩后独立存储在同一个 HDFS 块中。行组在 Hadoop 集群 DataNode 节点间的数据分布如图 7 所示。

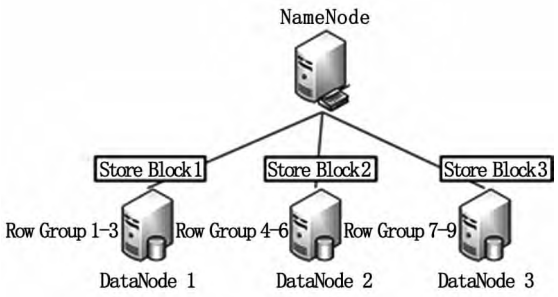

图7 行组在DataNode间的分布图

Fig. 7 The layout of row-column store structure among DataNodes

行组内按列垂直划分多个列,每个列在数据压缩后独立存储在 HDFS 块中,每行中所有列都存放在同一个 HDFS 块中,行组内数据按列垂直分割后的列存储方式在 HDFS 块中分布结构如图 8 所示。

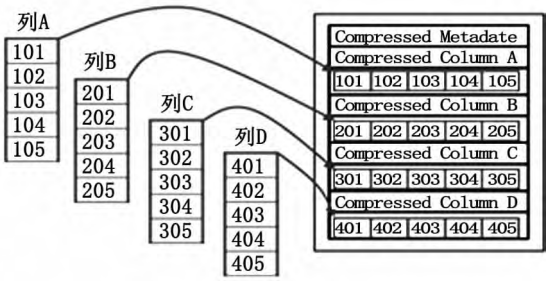

图8 行组内列式存储结构在HDFS块中分布图

Fig. 8 The layout of row-column store structure among HDFS

3.2 行列结合存储结构优点分析

数据读取操作分析。关系表被水平分割成多个行组,行组内按列垂直分割,存储在同一个 HDFS 块中。若读取列 A 和 C,首先读取本地行组,然后在行组中选择需要的列 A 和 C。这样既避免了行式存储结构中读取不必要的列而消耗额外的磁盘 I/O,也避免了列式存储结构中因跨节点访问而消耗额外的网络传输时间。

数据压缩分析。由于行组内是按列垂直分割,每一列具有相同的数据类型,因此在数据压缩时有很好的压缩比,从而提高了磁盘空间的利用率。

3.3 行组分割大小分析

Hadoop 分布式存储系统 HDFS 的块大小默认为 64MB,该值可以在配置文件中修改。若行组的存储空间比 HDFS 块要大,则一个行组被存储在多个 HDFS 块中,带来的问题主要有:

1) 由分布式存储系统特点,不同的块有可能分配在不同的数据节点中。因此,在数据访问时,有可能跨节点访问,从而消耗了不必要的网络传输时间。

2) HDFS 的高容错性是默认将每份数据在集群中有三处备份。若数据节点发生故障,因行列结合存储结构的约束,需要更长的恢复时间,同时也需要额外维护。

所以每个行组分割的大小最大不能超过 HDFS 块的大小,可以通过改变 HDFS 块的大小来改变行组大小。

4 性能评估

对行式存储、列式存储、行列结合的数据存储结构在数据加载时间和数据压缩效率上进行性能比较。比较结果如图 9 所示。

通过测试比较,行式存储在数据加载上,列式存储消耗时间最长,行式存储略优于行列存储。而在数据压缩效果上,行式存储压缩效果最差,行列存储压缩效果最好。

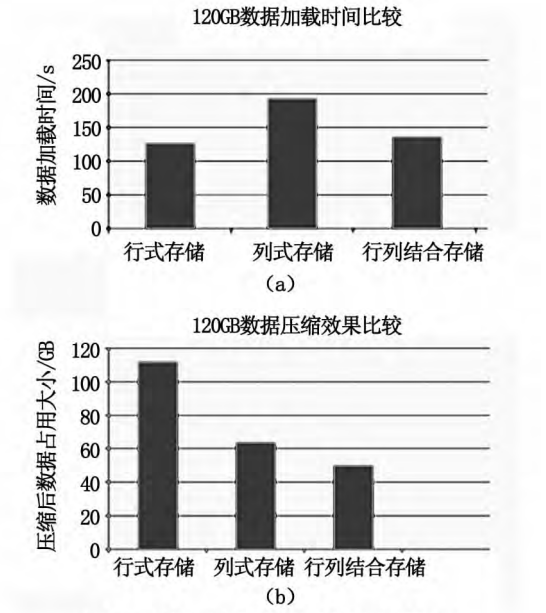

图9 三种数据存储结构性能比较图

Fig. 9 The comparison result of three data storage structures

5 结束语

行式数据存储结构虽然在数据加载上有较好的效果,但由于列的数据类型不同,在数据压缩上效率很低,不能有效地提高磁盘利用率。列式存

储结构虽然在数据压缩上有很好的压缩效果,但由于跨数据节点访问使得数据加载时间较长,从而降低了系统吞吐量。行列相结合的数据存储结构不仅有行式结构存储的高效数据加载,同时有很好的数据压缩效果。在分布式系统中,行列数据存储结构极大地提高了大数据存储及处理性能。

参考文献:

[1] MANYIKA J, CHUI M, BROWN B, et al. Big data: The next frontier for innovation, competition, and productivity [J]. Comm unications of the ACM, 2011, 56(2): 100 – 105.

[2] WHITE T. Hadoop: the definitive guide [M]. O´ Reilly, 2012.

[3] DEAN J, GHEMAWAT S. MapReduce: simplified data processing on large clusters [J]. Communications of the ACM, 2008, 51(1): 107 – 113.

[4] THUSOO A, SARMA J S, JAIN N, et al. Hive: a warehousing solution over a map – reduce framework [J]. Proceedings of the VLDB Endowment, 2009, 2(2): 1626 – 1629.

[5] BORTHAKUR D. The hadoop distributed file system: Architecture and design [J]. Hadoop Project Website, 2007 (11): 21.

[6] KAUSHIK R T, BHANDARKAR M, NAHRSTEDT K. Evaluation and analysis of greenhdfs: A self – adaptive, energy – conserving variant of the hadoop distributed file system [C]//Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on. IEEE, 2010: 274 – 287.

[7] SHVACHKO K, KUANG H, RADIA S, et al. The hadoop distributed file system [C]//Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on. IEEE, 2010: 1 – 10.

[8] LI N, RAO J, SHEKITA E, et al. Leveraging a scalable row store to build a distributed text index [C]//Proceedings of the first international workshop on Cloud data management. ACM, 2009: 29 – 36.

[9] IVANOVA M G, KERSTEN M L, NES N J, et al. An architecture for recycling intermediates in a column – store [J]. ACM Transactions on Database Systems (TODS), 2010, 35(4): 24.

[10] ABADI D J, BONCZ P A, HARIZOPOULOS S. Column – oriented database systems [J]. Proceedings of the VLDB Endowment, 2009, 2(2): 1664 – 1665.

(责任编辑　刘存英)