(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 101727465 A
(43)申请公布日 2010.06.09

(21)申请号 200810225486.3

(22)申请日 2008.11.03

(71)申请人 中国移动通信集团公司
地址 100032 北京市西城区金融大街 29 号

(72)发明人 徐萌 钱岭 罗治国 郭磊涛 赵鹏

(74)专利代理机构 北京同达信恒知识产权代理有限公司 11291
代理人 魏杉

(51)Int.Cl.
G06F 17/30(2006.01)

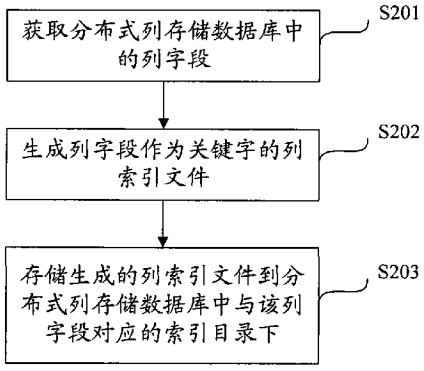
权利要求书 3 页 说明书 9 页 附图 4 页

(54)发明名称

分布式列存储数据库索引建立、查询方法及装置与系统

(57)摘要

本发明公开了一种分布式列存储数据库索引建立方法、查询方法及装置与系统。该分布式列存储数据库索引建立方法包括：获取分布式列存储数据库中的列字段，生成以所述列字段作为关键字的列索引文件，所述列索引文件中包含所述列字段在所述分布式列存储数据库中的值与对应的 Row 字段值的映射关系；存储所述列索引文件到所述分布式列存储数据库中与所述列字段对应的索引目录下。客户端发起以列字段作为查询条件和查询结果的查询请求时，通过匹配出相应的列索引文件，得到对应的 Row 字段值，从而实现索引查询。采用本发明能在现有分布式列存储数据库中，实现除 Row 字段外的其它列字段快速查询。



CN 101727465 A

1. 一种分布式列存储数据库索引建立方法，其特征在于，包括：

获取分布式列存储数据库中的列字段；

生成以所述列字段作为关键字的列索引文件，所述列索引文件中包含所述列字段在所述分布式列存储数据库中的值与对应的 Row 字段值的映射关系；

存储所述列索引文件到所述分布式列存储数据库中与所述列字段对应的索引目录下。

2. 如权利要求 1 所述的方法，其特征在于，在所述分布式列存储数据库的主服务器中，存储 Row 字段值与所述分布式列存储数据库的分片服务器的映射关系；

在所述分片服务器中，存储分配的分片数据中所述列字段对应的数据文件、以 Row 字段为关键字的索引文件和生成的对应列索引文件。

3. 如权利要求 2 所述的方法，其特征在于，所述分布式列存储数据库采用三级索引目录结构，具体包括：

在所述主服务器中存储第一级索引目录，在所述第一级索引目录中包含所述 Row 字段值与所述分片服务器的映射关系；

在所述分片服务器中存储第二级索引目录和第三级索引目录，在所述第二级索引目录中包含所述列字段与列存储文件的映射关系；在所述第三级索引目录中包含所述列存储文件对应列字段的所述数据文件、索引文件和列索引文件。

4. 如权利要求 3 所述的方法，其特征在于，当一个所述分片服务器中存储一片或以上的分片数据时，对每片分片数据分别建立所述第二级索引目录和第三级索引目录。

5. 如权利要求 1-4 任一所述的方法，其特征在于，当所述分布式列存储数据库中增加数据、删除数据或修改数据后，重新生成所述列索引文件或修改所述列索引文件中的对应数据。

6. 一种分布式列存储数据库查询方法，其特征在于，包括：

客户端向分布式列存储数据库的主服务器发起查询请求；

所述主服务器根据本地存储的 Row 字段值与所述分布式列存储数据库的分片服务器的映射关系，向所述客户端返回分片服务器信息；

所述客户端向所述分片服务器发起查询请求，在该查询请求中携带查询结果的列字段、查询条件的列字段及字段值信息；

所述分片服务器根据本地存储的列字段的索引目录，匹配出与所述查询条件的列字段对应的列索引文件，所述列索引文件中包含所述列字段在所述分布式列存储数据库中的值与对应的 Row 字段值的映射关系；

所述分片服务器根据匹配出的所述列索引文件及所述字段值信息获取对应 Row 字段值，并根据获取的 Row 字段值查询与所述查询结果的列字段对应的索引文件和数据文件，得到满足查询条件结果值，返回给所述客户端。

7. 如权利要求 6 所述的方法，其特征在于，当所述主服务器返回的分片服务器信息中包含多个分片服务器时，所述客户端并行地分别向每个分片服务器发起所述查询请求。

8. 如权利要求 6 或 7 所述的方法，其特征在于，当发送给所述分片服务器的所述查询请求中包含一个以上的查询条件时，所述分片服务器分别获取每个查询条件对应的 Row 字段值，再根据各查询条件之间的逻辑关系，确定出满足全部查询条件的最终 Row 字段值，并根据所述最终 Row 字段值查询所述查询结果的列字段对应的数据文件，得到满足查询条件结

果值，返回给所述客户端。

9. 一种分布式列存储数据库索引建立装置，其特征在于，包括：

获取单元，用于获取分布式列存储数据库中的列字段；

生成单元，用于生成以所述获取单元获取的所述列字段作为关键字的列索引文件，所述列索引文件中包含所述列字段在所述分布式列存储数据库中的值与对应的 Row 字段值的映射关系；

存储单元，用于存储所述列索引文件到所述分布式列存储数据库中与所述列字段对应的索引目录下。

10. 如权利要求 9 所述的装置，其特征在于，所述生成单元包括：

获取子单元，用于获取所述列字段在所述分布式列存储数据库中的值；

匹配子单元，用于在所述分布式列存储数据库中匹配出与所述列字段的值对应的 Row 字段值；

生成子单元，用于建立起所述列字段的值与对应的 Row 字段的值之间的映射关系，生成所述列索引文件。

11. 如权利要求 9 或 10 所述的装置，其特征在于，所述装置为软件模块，嵌入到存储所述分布式列存储数据库的分片数据的分片服务器中。

12. 一种分布式列存储数据库系统，包括主服务器和分片服务器，其特征在于：

所述主服务器包括：

第一存储单元，用于存储分布式列存储数据库的 Row 字段值与分片服务器的映射关系；以及

查询受理单元，用于接收客户端的查询请求，根据所述第一存储单元存储的所述映射关系向客户端返回分片服务器信息；

所述分片服务器包括：

列索引文件生成单元，用于获取分布式列存储数据库中的列字段，生成以所述列字段作为关键字的列索引文件，所述列索引文件中包含所述列字段在所述分布式列存储数据库中的值与对应的 Row 字段值的映射关系，并存储所述列索引文件到所述分布式列存储数据库中与所述列字段对应的索引目录下；

第二存储单元，用于存储分配的分片数据中的列字段对应的数据文件、以 Row 字段为关键字的索引文件和所述列字段的列索引文件；

分析单元，用于接收客户端发送的查询请求，分析所述查询请求中携带的查询结果的列字段、查询条件的列字段及字段值信息；

匹配单元，用于根据所述查询条件的列字段在所述第二存储单元中匹配出对应的列索引文件，并根据匹配出的所述列索引文件及所述字段值信息，获取对应 Row 字段值；

结果查询单元，用于用获取的 Row 字段值查询所述查询结果的列字段对应的索引文件和数据文件，得到满足查询条件的查询结果值；

结果返回单元，用于向发起查询请求的所述客户端返回所述查询结果值。

13. 如权利要求 12 所述的系统，其特征在于，在所述主服务器的第一存储单元中存储有第一级索引目录，在所述第一级索引目录中包含所述 Row 字段值与分片服务器的映射关系；

在所述分片服务器的第二存储单元中存储有第二级索引目录和第三级索引目录，在所述第二级索引目录中包含所述列字段与列存储文件的映射关系；在所述第三级索引目录中包含所述列存储文件对应列字段的所述数据文件、索引文件和列索引文件。

14. 如权利要求 12 或 13 所述的系统，其特征在于，所述分片服务器为多个。

分布式列存储数据库索引建立、查询方法及装置与系统

技术领域

[0001] 本发明涉及分布式列存储数据库，尤其涉及一种分布式列存储数据库的索引建立方法，数据查询方法及相应的装置与系统。

背景技术

[0002] 分布式列存储数据库是一种适合快速查询的、分布式的优良解决方案，它在提供海量数据存储的同时，还可以有效的提高对数据的查询速度。

[0003] 分布式列存储数据库的特点是：数据表中必须有 Row 字段，且 Row 字段为关键字，即不可重复，并排序。如果原表为 N 个列字段，则整个表在分布式列存储数据库中以 (N-1) 个表来进行存储；即除 Row 字段外，对其余的列字段分别存储一个对应表。

[0004] 举例说明如下：

[0005] 表一：GNTABLE

[0006]

Row	Time	UserID	SourceIP	ObjectIP	SingalType
1	20080909-12:00:00	13910001000	10.1.6.124	10.1.7.22	createPDP
2	20080909-12:00:00	13810001000	10.1.6.125	10.1.6.124	delPDP
3	20080909-12:00:01	13910001000	10.1.7.22	10.1.6.124	responsePDP
4	20080909-12:00:01	13910001000	10.1.7.22	10.1.6.124	createPDP

[0007] 上表一为分布式列存储数据库的一个原数据表 GNTABLE，其包含 Row 字段并排序，其余列字段包括：时间 (Time)、用户标识 (UserID)、源 IP 地址 (SourceIP)、目标 IP 地址 (ObjectIP) 和信号类型 (SingalType)。

[0008] 在列存储数据库中，需要针对各列字段 (Time、UserID、SourceIP、ObjectIP 和 SingalType) 分别存储一个对应表。以 Time 和 UserID 列字段为例，其存储的对应表分别如下表二和表三所示：

[0009] 表二

[0010]

Row	Time
1Time	20080909-12:00:00
2Time	20080909-12:00:00

Row	Time
3Time	20080909-12:00:01
4Time	20080909-12:00:01

[0011] 表三

[0012]

Row	UserID
1UserID	13910001000
2UserID	13810001000
3UserID	13910001000
4UserID	13910001000

[0013] 在分布式列存储数据库中，包含有主服务器(Master)和分片服务器(TabletServer)。其中，在主服务器中保存Row字段值与各分片服务器之间的映射关系，在各分片服务器中分别保存分布式列存储数据库的分片数据。所谓分片数据，是指将一个原数据表按照行分为几个分片（一个分片包含若干行），每个分片包括各行的全部数据。每个分片数据可以存储于一个分片服务器中（当然，一个分片服务器可以存储多个分片数据），各分片数据中按Row排序。每个分片数据中第一行的Row值为开始(begin)值，最后一行的Row值为结束(end)值，根据分片规则，则下一个分片数据的begin值＞上一个分片数据的end值。其存储架构示意图如图1所示，包括：

[0014] 在主服务器(Master)中包含有元数据(Metadata)模块，存储Row字段值与各分片服务器(TabletServer)的映射关系。在各分片服务器中包含数据片模块(HRegion)，在该模块中存储列字段（或列家族，在分布式列存储数据库中，将经常被同时访问的几个列定义为列家族，同一个列家族存储于一个列文件中）与对应列存储文件(HStoreFile)之间的映射关系，一个或多个HStoreFile存储在一个列模块(HStore)下。每个HStoreFile保存了两个文件，即数据(Data)文件和索引(Index)文件，并建立两者之间的映射。Data文件保存数据，其格式为<Key，value>，Index文件保存Key的索引，通过Key的索引，可以直接定位到Data文件中的某行数据。

[0015] 仍以上表一中的UserID列字段为例，在对应的HStoreFile中，其对应的Data文件和Index文件分别如下表四、表五所示。

[0016] 表四：

[0017]

	Row	Value
0	1 UserID	13910001000
2	2 UserID	13810001000
4	3 UserID	13910001000
6	4 UserID	13910001000

[0018] 表五：

[0019]

Row	Offset
1	0
2	2
3	4
4	6

[0020] 根据上述现有技术的存储架构，对于分布式列存储数所库，整体的索引机制形成树的形式，可以通过三层快速对 Row 进行定位。

[0021] 但由于现有技术中数据是根据主关键字 Row 排序并存储的，对于 Time、UserID 等非主关键字的列则不是排序的，因此以这些列为条件的访问就必须根据 Row 遍历整个数据表才能实现。在没有索引情况下的遍历数据，即便是分布式数据库，可以并发处理遍历请求，但其面对海量数据时，性能也无法忍受。而对于传统的数据库应用来说，使用非主关键字查询的场合非常多，因此需要有一种针对非主关键字的列的索引机制以满足使用需求。

发明内容

[0022] 本发明提供一种分布式列存储数据库索引建立方法、查询方法及装置与系统，用以解决现有分布式列存储数据库中不能够按照除 Row 字段外的其它列字段进行快速高效查询的问题。

[0023] 本发明提供的分布式列存储数据库索引建立方法，包括：

[0024] 获取分布式列存储数据库中的列字段；

[0025] 生成以所述列字段作为关键字的列索引文件，所述列索引文件中包含所述列字段在所述分布式列存储数据库中的值与相对应的 Row 字段值的映射关系；

[0026] 存储所述列索引文件到所述分布式列存储数据库中与所述列字段对应的索引目录下。

[0027] 本发明还提供一种分布式列存储数据库查询方法，包括：

[0028] 客户端向分布式列存储数据库的主服务器发起查询请求；

[0029] 所述主服务器根据本地存储的 Row 字段值与所述分布式列存储数据库的分片服务器的映射关系，向所述客户端返回分片服务器信息；

[0030] 所述客户端向所述分片服务器发起查询请求，在该查询请求中携带查询结果的列字段、查询条件的列字段及字段值信息；

[0031] 所述分片服务器根据本地存储的列字段的索引目录，匹配出与所述查询条件的列字段对应的列索引文件，所述列索引文件中包含所述列字段在所述分布式列存储数据库中的值与相对应的 Row 字段值的映射关系；

[0032] 所述分片服务器根据匹配出的所述列索引文件及所述字段值信息获取对应 Row 字段值，并根据获取的 Row 字段值查询与所述查询结果的列字段对应的索引文件和数据文件，得到满足查询条件结果值，返回给所述客户端。

[0033] 本发明再提供一种分布式列存储数据库索引建立装置，包括：

[0034] 获取单元，用于获取分布式列存储数据库中的列字段；

[0035] 生成单元，用于生成以所述获取单元获取的所述列字段作为关键字的列索引文件，所述列索引文件中包含所述列字段在所述分布式列存储数据库中的值与相对应的 Row 字段值的映射关系；

[0036] 存储单元，用于存储所述列索引文件到所述分布式列存储数据库中与所述列字段对应的索引目录下。

[0037] 本发明再提供一种分布式列存储数据库系统，包括主服务器和分片服务器，所述主服务器包括：

[0038] 第一存储单元，用于存储分布式列存储数据库的 Row 字段值与分片服务器的映射关系；以及

[0039] 查询受理单元，用于接收客户端的查询请求，根据所述第一存储单元存储的所述映射关系向客户端返回分片服务器信息；

[0040] 所述分片服务器包括：

[0041] 列索引文件生成单元，用于获取分布式列存储数据库中的列字段，生成以所述列字段作为关键字的列索引文件，所述列索引文件中包含所述列字段在所述分布式列存储数据库中的值与对应的 Row 字段值的映射关系，并存储所述列索引文件到所述分布式列存储数据库中与所述列字段对应的索引目录下；

[0042] 第二存储单元，用于存储分配的分片数据中的列字段对应的数据文件、以 Row 字段为关键字的索引文件和所述列字段的列索引文件；

[0043] 分析单元，用于接收客户端发送的查询请求，分析所述查询请求中携带的查询结果的列字段、查询条件的列字段及字段值信息；

[0044] 匹配单元，用于根据所述查询条件的列字段在所述第二存储单元中匹配出对应的列索引文件，并根据匹配出的所述列索引文件及所述字段值信息，获取对应 Row 字段值；

[0045] 结果查询单元，用于用获取的 Row 字段值查询所述查询结果的列字段对应的索引文件和数据文件，得到满足查询条件的查询结果值；

[0046] 结果返回单元，用于向发起查询请求的所述客户端返回所述查询结果值。

[0047] 本发明通过获取分布式列存储数据库中除 Row 字段外的列字段，生成以列字段作为关键字的列索引文件，在该列索引文件中包含列字段在分布式列存储数据库中的值与相对应的 Row 字段值的映射关系；并将生成的列索引文件存储到与列字段对应的索引目录下。使得客户端可以向分布式列存储数据库的主服务器发起携带查询结果的列字段、查询

条件的列字段及字段值信息的查询请求，通过主服务器、分片服务器根据存储的列字段的索引目录，匹配出与查询条件的列字段对应的列索引文件，根据列索引文件获取对应 Row 字段值，并根据获取的 Row 字段值查询所述查询结果的列字段对应的数据文件，得到满足查询条件结果值，返回给客户端。从而实现客户端可以方便地针对分布式列存储数据库采用非 Row 字段的其余列字段进行快速高效的索引查询。

附图说明

[0048] 图 1 为现有技术中分布式列存储数据库存储架构示意图；

[0049] 图 2 为本发明实施例提供的分布式列存储数据库索引建立方法流程图；

[0050] 图 3 为本发明实施例提供的 HStoreFile 下的文件结构示意图；

[0051] 图 4 为本发明实施例提供的分布式列存储数据库查询方法流程图；

[0052] 图 5 为本发明实施例提供的分布式列存储数据库索引建立装置结构示意图；

[0053] 图 6 为本发明实施例提供的分布式列存储数据库索引建立装置中生成单元的内部结构示意图；

[0054] 图 7 为本发明实施例提供的分布式列存储数据库系统结构示意图。

具体实施方式

[0055] 本发明实施例提供一种分布式列存储数据库索引建立方法，其实现流程如图 2 所示，包括：

[0056] 步骤 S201、获取分布式列存储数据库中的列字段。

[0057] 步骤 S202、生成以获取的列字段作为关键字的列索引文件，在列索引文件中包含该列字段在分布式列存储数据库中的值与相对应的 Row 字段值的映射关系。

[0058] 在该步骤 S202 中，可以针对获取的每一个列字段（或列家族），分别生成一个对应的列索引文件。

[0059] 实际应用中，为方便用户查询，理论上可以对分布式列存储数据库中除 Row 字段外的每一个列字段，都生成一个对应的列索引文件。当然，如果某些列字段基本没有查询的价值，实际中几乎不会采用该字段进行查询，则不必生成对应的列索引文件，以节省数据库占用的存储资源。

[0060] 步骤 S203、存储生成的列索引文件到分布式列存储数据库中与该列字段对应的索引目录下。

[0061] 根据上述流程描述可知，本发明在现有技术的基础上，为分布式列存储数据库中除 Row 字段外的其余列字段分别生成了一个对应的列索引文件，并存储到与列字段对应的索引目录下。

[0062] 仍沿用上述表一为例，针对列字段 UserID 生成的列索引文件如下表六所示：

[0063] 表六：

[0064]

UserID	Row
13910001000	1
	3
	4
13810001000	2

[0065] 表六中，左边一栏为 UserID 在原分布式列存储数据库中的值，根据表三可知，其字段值只有两个，其一为 13910001000 和 13810001000；右边一栏为 Row 字段值，即与 UserID 的每个值分别对应的 Row 字段值，由表三可知，与 13910001000 对应的 Row 字段值分别为 1、3、4，与 13810001000 对应的 Row 字段值为 2。

[0066] 下面结合分布式列存储数据库的存储架构，进行具体说明：

[0067] 在分布式列存储数据库的主服务器中存储第一级索引目录，在第一级索引目录中包含 Row 字段值与各分片服务器的映射关系；例如，在主服务器的元数据模块中存储第一级索引目录。根据第一级索引目录，主服务器可以查找到全部的分片服务器。

[0068] 在每个分片服务器中存储第二级索引目录和第三级索引目录，在第二级索引目录中包含列字段与列存储文件的映射关系；例如，在分片服务器的数据片模块中存储第二级索引目录。在第三级索引目录下，存储列存储文件对应列字段的数据文件、索引文件和本发明生成的列索引文件。第三级索引目录相当于现有现有技术中的 HStoreFile，所不同的是，本发明在现有技术的 HStoreFile 下增加了一个与该列字段对应的列索引文件，其层级关系示意图如图 3 所示：

[0069] 在列存储文件 (HStoreFile) 下，存储有三个文件，分别为：

[0070] 在对应的分片服务器分配的分片数据中，该列字段对应的数据 (Data) 文件（为描述方便，后续统一称为 Data 文件）、以 Row 字段为关键字的索引 (Index) 文件（为描述方便，后续统一称为 Index 文件）和本发明生成的对应列索引 (ColIndex) 文件（为描述方便，后续统一称为 ColIndex 文件）。

[0071] 在分片服务器中，对列字段建立对应的列索引文件，可由用户指定。即在分片服务器向用户提供创建索引、删除索引的接口，用户可以根据自己的使用需要，建立全部或部分列字段对应的列索引文件。

[0072] 根据本发明上述实施例提供的方法，当一个分片服务器中存储一片及一片以上的分片数据时，在该分片服务器中针对每片分片数据分别建立第二级索引目录和第三级索引目录。

[0073] 当分布式列存储数据库中增加数据、删除数据或修改数据后，需要重新生成列索引文件，或者修改已生成的列索引文件中的对应数据，以保证列索引文件中的数据与当前数据库中的相关数据相一致，以避免后续查询时出现错误的查询结果。

[0074] 基于同一发明构思，根据本发明提供的上述分布式列存储数据库索引建立方法，

本发明还提供一种分布式列存储数据库查询方法，其具体实现流程如图 4 所示，包括：

[0075] 步骤 S401、客户端向分布式列存储数据库的主服务器发起查询请求；

[0076] 步骤 S402、主服务器根据本地存储的 Row 字段值与分片服务器的映射关系，向客户端返回分片服务器信息；

[0077] 步骤 S403、客户端向分片服务器发起查询请求，在该查询请求中携带查询结果的列字段、查询条件的列字段及字段值信息；

[0078] 步骤 S404、分片服务器根据本地存储的列字段的索引目录，匹配出与查询条件的列字段对应的 ColIndex 文件；

[0079] 步骤 S405、分片服务器根据匹配出的 ColIndex 文件及查询条件中携带的列字段的字段值信息，获取对应 Row 字段值；

[0080] 步骤 S406、分片服务器根据获取的 Row 字段值，以及查询结果的列字段对应的 Index 文件和 Data 文件，得到满足查询条件结果值；

[0081] 步骤 S407、分片服务器返回符合查询条件的结果值给发起查询请求的客户端。

[0082] 仍以上表一为例，假设查询请求为“Select SignalType from GNTABLEwhere UserID ＝’13910001000’”，即从 GNTABLE 数据表中选择列字段 UserID 为“13910001000”的用户对应使用的信号类型。该查询请求中，携带的查询条件的列字段为“UserID”字段，字段值为“13910001000”，查询结果列字段为“SignalType”字段。

[0083] 根据本发明提供的上述流程，客户端先向主服务器发起查询请求，主服务器将各分片服务器信息返回给客户端；客户端再分别向各分片服务器发起查询，当有多个分片服务器时，客户端并行地分别向每个分片服务器发起查询请求，实现分布式查询；每个分片服务器根据本地存储的分片数据，查询出满足查询条件的结果值后返回给客户端，客户端接收各分片服务器返回的查询结果，即得到最终的查询数据。

[0084] 具体地，分片服务器接收到上述查询请求后，在本地存储的列字段的索引目录中匹配出与查询条件的列字段“UserID”字段对应的列索引文件，即如表六所示，分片服务器根据匹配出的列索引文件，获取 UserID 字段值为“13910001000”对应 Row 字段值为“1、3、4”；得到 Row 字段值后，再采用现有技术中分布式列存储数据库的查询方式，得到查询结果；即：再根据本次查询结果对应的列字段（“SignalType”字段）的 Index 文件和 Data 文件，即可获得满足查询要求的对应 SignalType 字段值。

[0085] 当查询请求中携带有多个查询条件时，分片服务器分别获取每个查询条件对应的 Row 字段值，再根据各查询条件之间的逻辑关系（逻辑“或”，逻辑“与”或其组合），确定出满足全部查询条件的最终 Row 字段值，再根据确定出的最终 Row 字段值，查询得到满足查询条件的结果值返回给客户端。

[0086] 采用本发明提供的分布式列存储数据库查询方法，客户端可以并行地同时向各分片服务器发起查询请求，使得对数据的多条件查询处理在各分片服务器同时进行，从而实现了快速高效的查询。而如果不采用分布式查询方式，由主服务器进行集中式的多条件查询处理，当进行海量数据查询时，会出现海量数据单节点无法处理的情况。

[0087] 其次，采用本发明提供的分布式列存储数据库查询方法，分片服务器直接在本地进行数据查询处理，即各分片服务器只需处理本地存储的数据就能获得查询结果，没有网络交互，减少了网络的开销，进一步提高了查询速度及效率。

[0088] 基于同一发明构思，根据本发明上述实施例提供的分布式列存储数据库索引建立方法，本发明还提供一种分布式列存储数据库索引建立装置，其结构示意图如图 5 所示，包括：

[0089] 获取单元 71，用于获取分布式列存储数据库中的列字段；

[0090] 生成单元 72，用于生成以获取单元 71 获取的列字段作为关键字的列索引文件，在该列索引文件中包含列字段在分布式列存储数据库中的值与相对应的 Row 字段值的映射关系；

[0091] 存储单元 73，用于存储生成单元 72 生成的列索引文件到分布式列存储数据库中与该列字段对应的索引目录下。

[0092] 其中，生成单元 72 的内部结构如图 6 所示，可以进一步包括：

[0093] 获取子单元 721，用于获取列字段在所述分布式列存储数据库中的值；

[0094] 匹配子单元 722，用于在分布式列存储数据库中匹配出与列字段的值相对应的 Row 字段值；

[0095] 生成子单元 723，用于建立起列字段的值与相对应的 Row 字段的值之间的映射关系，生成列索引文件。

[0096] 在实际应用中，本发明提供的分布式列存储数据库索引建立装置可以是软件模块，嵌入到存储分布式列存储数据库的分片数据的分片服务器中。

[0097] 基于同一发明构思，本发明再提供一种分布式列存储数据库系统，其结构示意图如图 7 所示，包括主服务器和分片服务器，其中：

[0098] 所述主服务器包括：

[0099] 第一存储单元 81，用于存储分布式列存储数据库的 Row 字段值与分片服务器的映射关系；以及

[0100] 查询受理单元 82，用于接收客户端的查询请求，根据第一存储单元 81 存储的所述映射关系向客户端返回分片服务器信息；

[0101] 所述分片服务器包括：

[0102] 列索引文件生成单元 91，用于获取分布式列存储数据库中的列字段，生成以列字段作为关键字的列索引文件，在该列索引文件中包含列字段在分布式列存储数据库中的值与相对应的 Row 字段值的映射关系，并存储生成的列索引文件到分布式列存储数据库中与该列字段对应的索引目录下；

[0103] 第二存储单元 92，用于存储分配的分片数据中的列字段对应的数据文件、以 Row 字段为关键字的索引文件和列字段的列索引文件；

[0104] 分析单元 93，用于接收客户端发送的查询请求，分析所述查询请求中携带的查询结果的列字段、查询条件的列字段及字段值信息；

[0105] 匹配单元 94，用于根据查询请求中携带的查询条件的列字段在第二存储单元 92 中匹配出对应的列索引文件，并根据匹配出的列索引文件以及字段值信息，获取与查询条件列字段的字段值对应 Row 字段值；

[0106] 结果查询单元 95，用于用获取的 Row 字段值查询所述查询结果的列字段对应的索引文件和数据文件，得到满足查询条件的查询结果值；

[0107] 结果返回单元 96，用于向发起查询请求的所述客户端返回查询结果值。

[0108] 主服务器用于存储分布式列存储数据库的 Row 字段值与分片服务器的映射关系；在分片服务器中，除了存储分配的分片数据中的列字段对应的 Data 文件、以 Row 字段为关键字的 Index 文件外，还存储该列字段的 ColIndex 文件；该 ColIndex 文件和 Data 文件以及 Index 文件一起保存在列字段对应的索引目录下。所述列索引文件，采用本发明上述实施例提供的方法建立，在其中包含列字段在分布式列存储数据库中的值与相对应的 Row 字段值的映射关系。

[0109] 如前所述，在主服务器中可以存储有第一级索引目录，在第一级索引目录中包含 Row 字段值与分片服务器的映射关系；在分片服务器中可以存储有第二级索引目录和第三级索引目录，在第二级索引目录中包含列字段与列索引文件的映射关系；在第三级索引目录下，存储列索引文件对应列字段的 Data 文件、Index 文件和本发明建立的 ColIndex 文件。

[0110] 本发明提供的分布式列存储数据库系统中，分片服务器可以是一个或多个。

[0111] 综上所述，本发明通过获取分布式列存储数据库中除 Row 字段外的列字段，生成以列字段作为关键字的列索引文件，在该列索引文件中包含该列字段在分布式列存储数据库中的值与相对应的 Row 字段值的映射关系；并将生成的列索引文件存储到与列字段对应的索引目录下。从而使得客户端可以向分布式列存储数据库的主服务器发起携带查询结果的列字段、查询条件的列字段及字段值信息的查询请求，通过匹配出与查询条件的列字段对应的列索引文件，获取对应 Row 字段值，再利用现有技术的查询方式根据 Row 字段值获得查询结果，实现了在分布式列存储数据库中采用非 Row 字段的其余列字段进行索引查询，极大地满足用户的使用需求。

[0112] 采用本发明提供的分布式列存储数据库查询方法，由客户端并行地同时向各分片服务器发起查询请求，使得对数据的多条件查询处理在各分片服务器同时进行，从而实现了快速高效的查询。而如果不采用本发明提供的分布式列存储数据库查询方式，而采用现有数据库常用的索引方法，即在主服务器中建立一个索引表，进行集中式的多条件查询处理，在索引表中存储列字段中列数据到其存储位置的映射，这种常规索引方法在处理所有的条件数据判断时，主服务器极有可能内存溢出，导致无法处理；且在获取数据的存储位置时，需要经过三次索引定位，增加网络开销。

[0113] 其次，采用本发明提供的分布式列存储数据库查询方法，分片服务器直接在本地进行数据查询处理，即各分片服务器只需处理本地存储的数据就能获得查询结果，没有网络交互，减少了网络的开销，进一步提高了查询速度及效率。

[0114] 再次，采用本发明提供的分布式列存储数据库查询方法，每次查询是针对列索引文件进行的，相对于采用遍历方式查询所需要的时间复杂度 N 而言，其时间复杂度仅为 $\log_2 N$。

[0115] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分步骤是可以通过程序来指令相关的硬件来完成，该程序可以存储于一计算机可读取存储介质中，如：ROM/RAM、磁碟、光盘等。

[0116] 显然，本领域的技术人员可以对本发明进行各种改动和变型而不脱离本发明的精神和范围。这样，倘若本发明的这些修改和变型属于本发明权利要求及其等同技术的范围之内，则本发明也意图包含这些改动和变型在内。

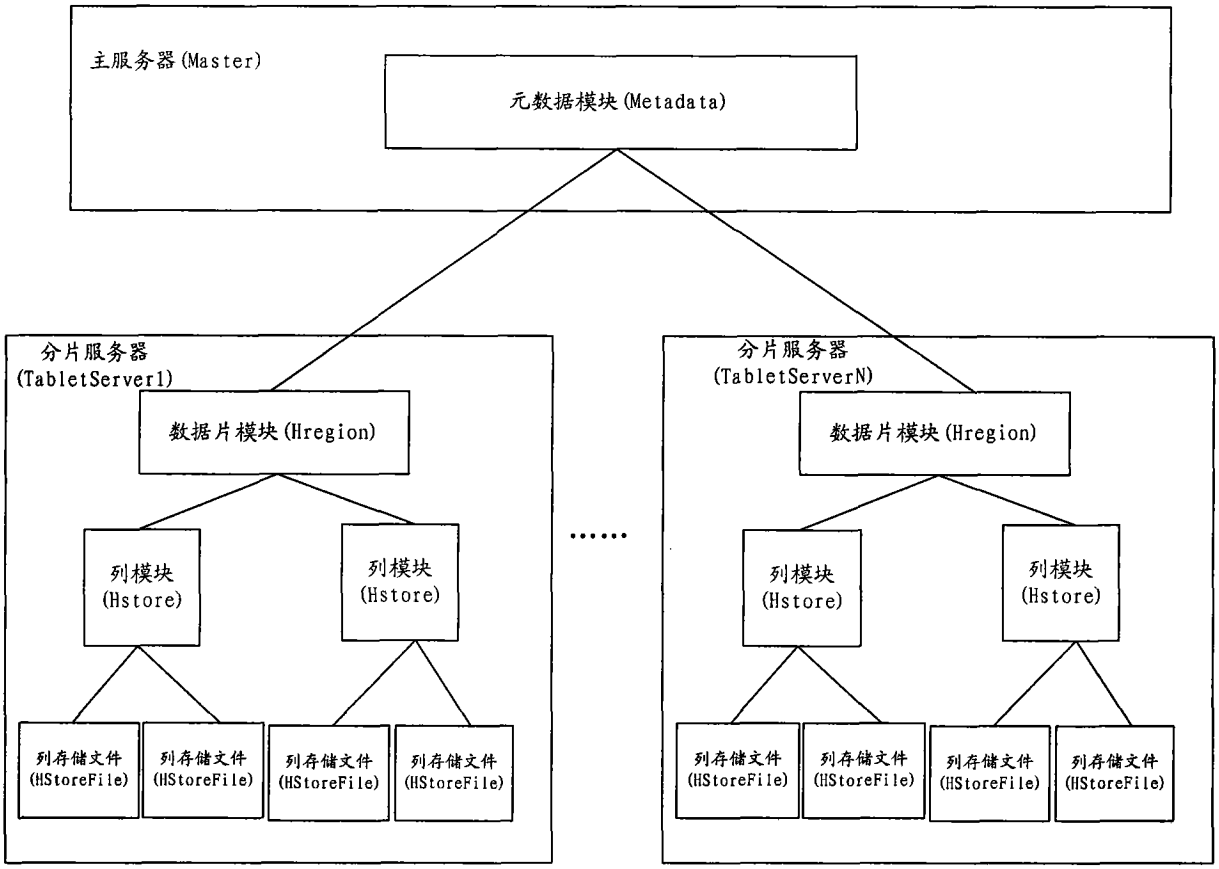

主服务器(Master)
元数据模块(Metadata)
分片服务器
(TabletServer1)
数据片模块(Hregion)
列模块
(Hstore)
列模块
(Hstore)
列存储文件
(HStoreFile)
列存储文件
(HStoreFile)
列存储文件
(HStoreFile)
列存储文件
(HStoreFile)
.....
分片服务器
(TabletServerN)
数据片模块(Hregion)
列模块
(Hstore)
列模块
(Hstore)
列存储文件
(HStoreFile)
列存储文件
(HStoreFile)
列存储文件
(HStoreFile)
列存储文件
(HStoreFile)

图 1

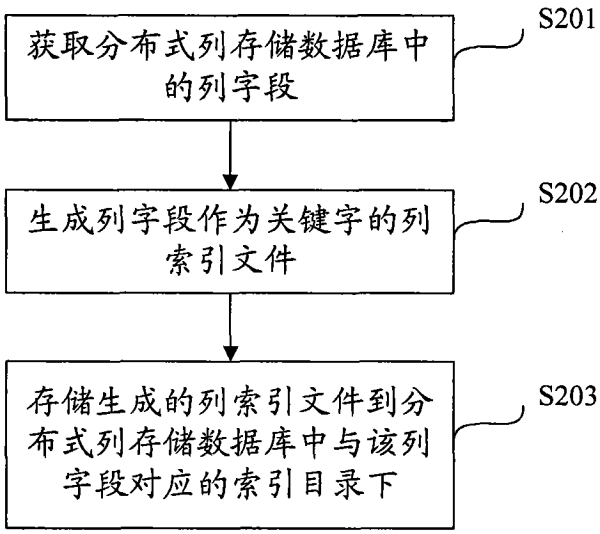

获取分布式列存储数据库中的列字段
S201
生成列字段作为关键字的列索引文件
S202
存储生成的列索引文件到分布式列存储数据库中与该列字段对应的索引目录下
S203

图 2

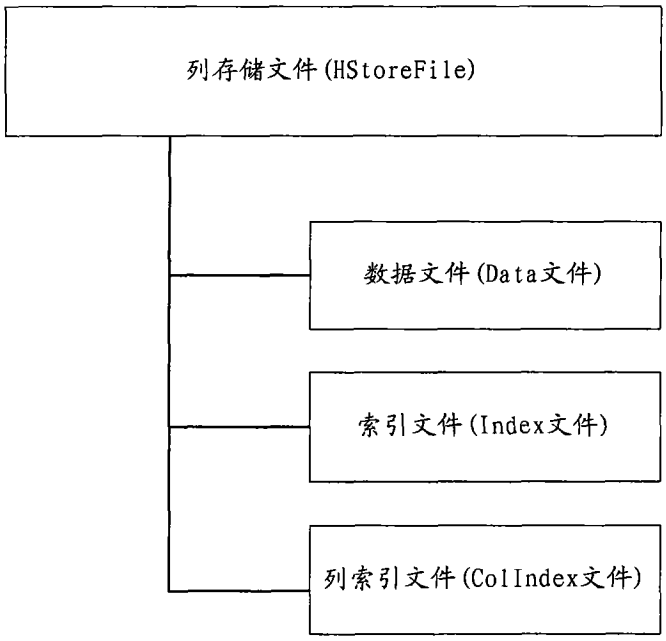

列存储文件(HStoreFile)
数据文件(Data文件)
索引文件(Index文件)
列索引文件(ColIndex文件)

图 3

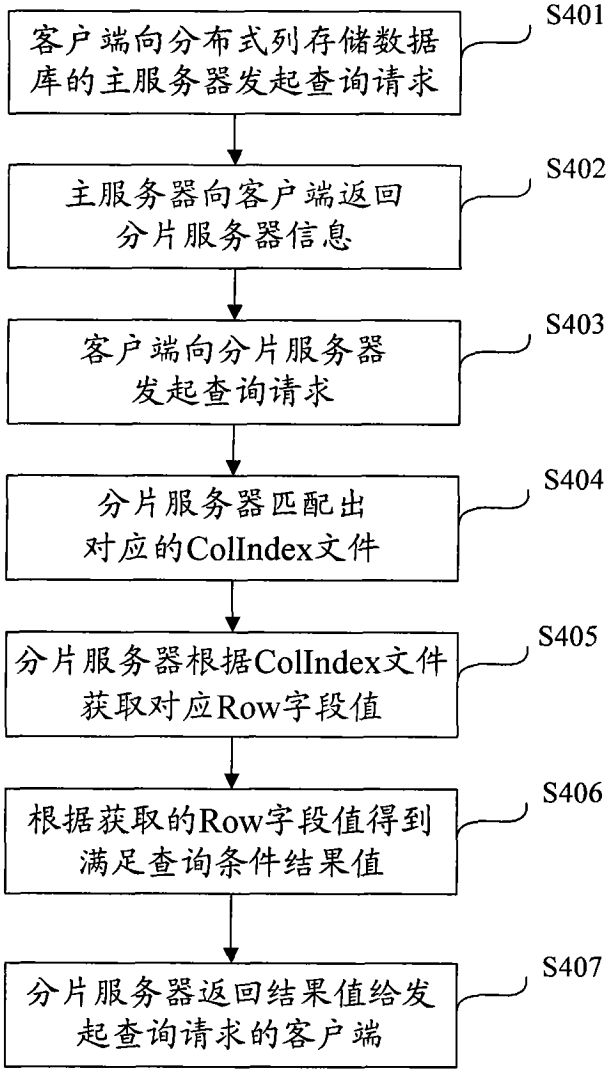

客户端向分布式列存储数据库的主服务器发起查询请求
S401
主服务器向客户端返回分片服务器信息
S402
客户端向分片服务器发起查询请求
S403
分片服务器匹配出对应的ColIndex文件
S404
分片服务器根据ColIndex文件获取对应Row字段值
S405
根据获取的Row字段值得到满足查询条件结果值
S406
分片服务器返回结果值给发起查询请求的客户端
S407

图 4

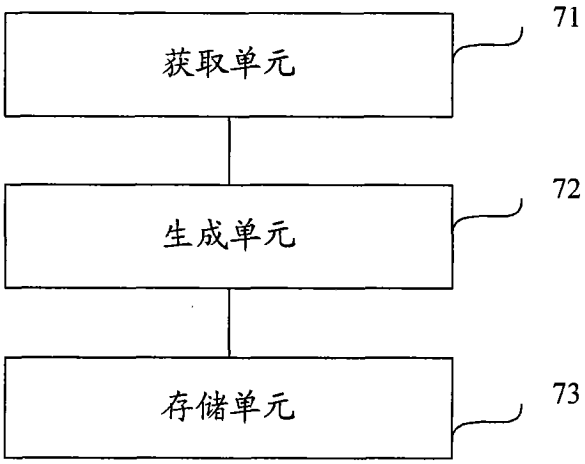

获取单元
71
生成单元
72
存储单元
73

图 5

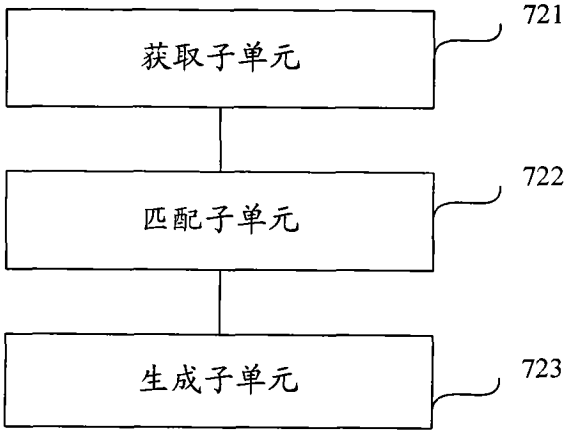

获取子单元
721
匹配子单元
722
生成子单元
723

图 6

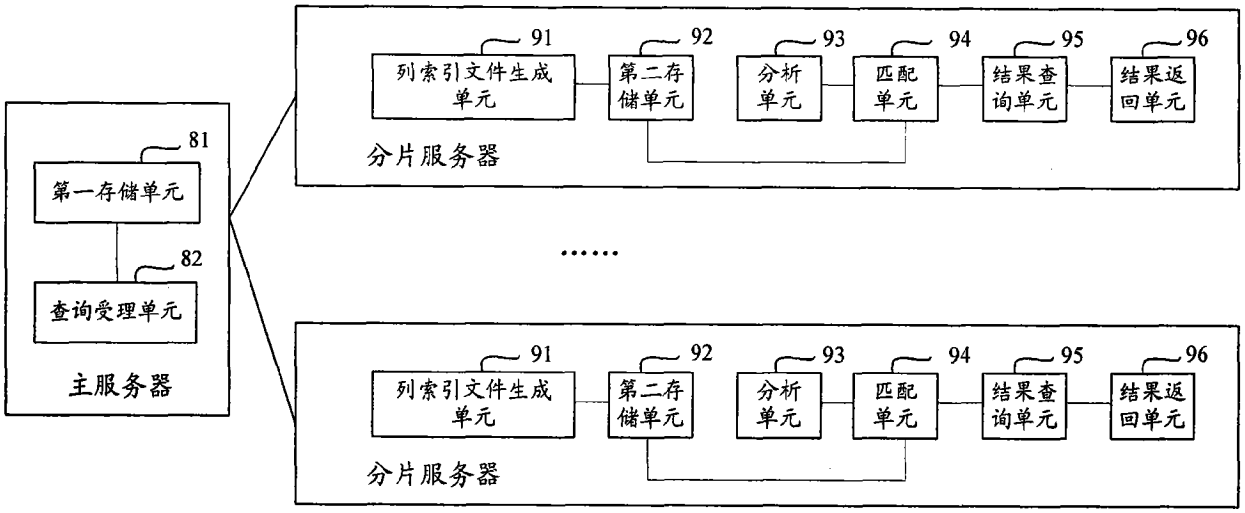

91
92
93
94
95
96
列索引文件生成单元
第二存储单元
分析单元
匹配单元
结果查询单元
结果返回单元
分片服务器
81
第一存储单元
82
查询受理单元
主服务器
.....
91
92
93
94
95
96
列索引文件生成单元
第二存储单元
分析单元
匹配单元
结果查询单元
结果返回单元
分片服务器

图 7