(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 106599062 A
(43)申请公布日 2017.04.26

(21)申请号 201611028735.0

(22)申请日 2016.11.18

(71)申请人 北京奇虎科技有限公司
地址 100088 北京市西城区新街口外大街28号D座112室(德胜园区)
申请人 奇智软件(北京)有限公司

(72)发明人 李远策 李振炜 白泉 王锋 武志刚

(74)专利代理机构 北京市隆安律师事务所 11323
代理人 权鲜枝 何立春

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

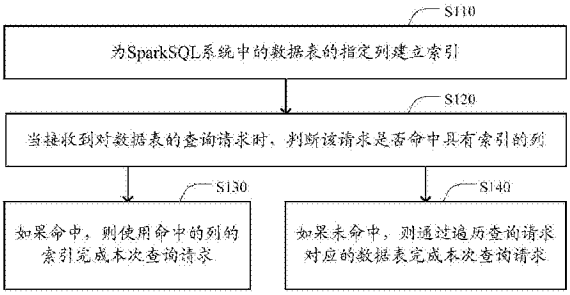
权利要求书2页 说明书8页 附图1页

(54)发明名称

一种SparkSQL系统中的数据处理方法和装置

(57)摘要

本发明公开了一种SparkSQL系统中的数据处理方法和装置，其中方法包括：为SparkSQL系统中的数据表的指定列建立索引；当接收到对数据表的查询请求时，判断该请求是否命中具有索引的列；如果命中，则使用命中的列的索引完成本次查询请求；如果未命中，则通过遍历所述查询请求对应的数据表完成本次查询请求。该技术方案通过为SparkSQL系统中数据表的指定列建立索引的方式显著提高了对SparkSQL系统中数据表的查询速度，在大数据场景下，如果查询请求命中具有索引的列，相比于利用SparkSQL提供的查询方式会有指数级的性能提升，对大数据查询具有极大意义。



CN 106599062 A

1.一种SparkSQL系统中的数据处理方法,其中,该方法包括:

为SparkSQL系统中的数据表的指定列建立索引;

当接收到对数据表的查询请求时,判断该请求是否命中具有索引的列;

如果命中,则使用命中的列的索引完成本次查询请求;

如果未命中,则通过遍历所述查询请求对应的数据表完成本次查询请求。

2.如权利要求1所述的方法,其中,所述为SparkSQL系统中的数据表的指定列建立索引包括:

为数据表的指定列建立JSON格式的索引,所述索引包括:索引号和数据位置标识。

3.如权利要求2所述的方法,其中,

所述索引号包括:指定列的列名、该条索引对应的列值;

所述数据位置标识包括:该数据表的存储路径、该条索引对应的数据在该数据表中的偏移量。

4.如权利要求2或3所述的方法,其中,所述为SparkSQL系统中的数据表的指定列建立索引还包括:

将建立的索引保存在指定搜索服务器的文件系统中。

5.如权利要求4所述的方法,其中,所述使用命中的列的索引完成本次查询请求包括:

将所述查询请求发送至所述指定搜索服务器,接收所述指定搜索服务器返回的查询结果;其中,所述查询结果为空,或者为一个或多个数据位置标识;

当所述查询结果不为空时,根据所述指定搜索服务器返回的一个或多个数据位置标识,从数据表中读取与相应数据位置标识匹配的数据。

6.如权利要求1-5中任一项所述的方法,其中,所述判断该请求是否命中具有索引的列包括:

从所述查询请求中解析出待查询的表名和列名;

根据所述待查询的表名,判断相应的数据表是否存在与待查询的列名对应的索引,若存在则判断为命中,若不存在则判断为未命中。

7.如权利要求6所述的方法,其中,所述为SparkSQL系统中的数据表的指定列建立索引还包括:

在该数据表的表结构中标识已建立索引的列;

所述根据所述待查询的表名,判断相应的数据表是否存在与待查询的列名对应的索引包括:

根据所述待查询的表名,从相应数据表的表结构中读取该数据表已建立的索引的列,根据所述待查询的列名判断所述表结构中是否存在与该列名对应的索引。

8.一种SparkSQL系统中的数据处理装置,其中,该装置包括:

索引建立单元,适于为SparkSQL系统中的数据表的指定列建立索引;

请求处理单元,适于在接收到对数据表的查询请求时,判断该请求是否命中具有索引的列;如果命中,则使用命中的列的索引完成本次查询请求;如果未命中,则通过遍历所述查询请求对应的数据表完成本次查询请求。

9.如权利要求8所述的装置,其中,

所述索引建立单元,适于为数据表的指定列建立JSON格式的索引,所述索引包括:索引

号和数据位置标识。

10.如权利要求9所述的装置,其中,

所述索引号包括:指定列的列名、该条索引对应的列值;

所述数据位置标识包括:该数据表的存储路径、该条索引对应的数据在该数据表中的偏移量。

一种SparkSQL系统中的数据处理方法和装置

技术领域

[0001] 本发明涉及计算机技术领域，具体涉及一种SparkSQL系统中的数据处理方法和装置。

背景技术

[0002] SparkSQL是一个使用SQL进行大数据分析的系统，可以进行TB至PB级的数据统计。但是SparkSQL批处理的计算模型，限制了其进行SQL查询的速度。这是由于SparkSQL系统中数据表的原始数据是以文件的形式存储在HDFS(Hadoop Distributed File System，Hadoop分布式文件系统)上的，在对SQL查询请求进行处理时，为从数据表中提取符合条件的记录，需要把相应数据表中的数据逐条读取进行判断，因此不能很好地实现交互式即时查询的效果。

发明内容

[0003] 鉴于上述问题，提出了本发明以便提供一种克服上述问题或者至少部分地解决上述问题的SparkSQL系统中的数据处理方法和装置。

[0004] 依据本发明的一个方面，提供了一种SparkSQL系统中的数据处理方法，包括：

[0005] 为SparkSQL系统中的数据表的指定列建立索引；

[0006] 当接收到对数据表的查询请求时，判断该请求是否命中具有索引的列；

[0007] 如果命中，则使用命中的列的索引完成本次查询请求；

[0008] 如果未命中，则通过遍历所述查询请求对应的数据表完成本次查询请求。

[0009] 可选地，所述为SparkSQL系统中的数据表的指定列建立索引包括：

[0010] 为数据表的指定列建立JSON格式的索引，所述索引包括：索引号和数据位置标识。

[0011] 可选地，所述索引号包括：指定列的列名、该条索引对应的列值；

[0012] 所述数据位置标识包括：该数据表的存储路径、该条索引对应的数据在该数据表中的偏移量。

[0013] 可选地，所述为SparkSQL系统中的数据表的指定列建立索引还包括：

[0014] 将建立的索引保存在指定搜索服务器的文件系统中。

[0015] 可选地，所述使用命中的列的索引完成本次查询请求包括：

[0016] 将所述查询请求发送至所述指定搜索服务器，接收所述指定搜索服务器返回的查询结果；其中，所述查询结果为空，或者为一个或多个数据位置标识；

[0017] 当所述查询结果不为空时，根据所述指定搜索服务器返回的一个或多个数据位置标识，从数据表中读取与相应数据位置标识匹配的数据。

[0018] 可选地，所述判断该请求是否命中具有索引的列包括：

[0019] 从所述查询请求中解析出待查询的表名和列名；

[0020] 根据所述待查询的表名，判断相应的数据表是否存在与待查询的列名对应的索引，若存在则判断为命中，若不存在则判断为未命中。

[0021] 可选地，所述为SparkSQL系统中的数据表的指定列建立索引还包括：

[0022] 在该数据表的表结构中标识已建立索引的列；

[0023] 所述根据所述待查询的表名，判断相应的数据表是否存在与待查询的列名对应的索引包括：

[0024] 根据所述待查询的表名，从相应数据表的表结构中读取该数据表已建立的索引的列，根据所述待查询的列名判断所述表结构中是否存在与该列名对应的索引。

[0025] 依据本发明的另一方面，提供了一种SparkSQL系统中的数据处理装置，包括：

[0026] 索引建立单元，适于为SparkSQL系统中的数据表的指定列建立索引；

[0027] 请求处理单元，适于在接收到对数据表的查询请求时，判断该请求是否命中具有索引的列；如果命中，则使用命中的列的索引完成本次查询请求；如果未命中，则通过遍历所述查询请求对应的数据表完成本次查询请求。

[0028] 可选地，所述索引建立单元，适于为数据表的指定列建立JSON格式的索引，所述索引包括：索引号和数据位置标识。

[0029] 可选地，所述索引号包括：指定列的列名、该条索引对应的列值；

[0030] 所述数据位置标识包括：该数据表的存储路径、该条索引对应的数据在该数据表中的偏移量。

[0031] 可选地，所述索引建立单元，适于将建立的索引保存在指定搜索服务器的文件系统中。

[0032] 可选地，所述请求处理单元，适于将所述查询请求发送至所述指定搜索服务器，接收所述指定搜索服务器返回的查询结果；其中，所述查询结果为空，或者为一个或多个数据位置标识；当所述查询结果不为空时，根据所述指定搜索服务器返回的一个或多个数据位置标识，从数据表中读取与相应数据位置标识匹配的数据。

[0033] 可选地，所述请求处理单元，适于从所述查询请求中解析出待查询的表名和列名；根据所述待查询的表名，判断相应的数据表是否存在与待查询的列名对应的索引，若存在则判断为命中，若不存在则判断为未命中。

[0034] 可选地，所述索引建立单元，还适于在该数据表的表结构中标识已建立索引的列；

[0035] 所述请求处理单元，适于根据所述待查询的表名，从相应数据表的表结构中读取该数据表已建立的索引的列，根据所述待查询的列名判断所述表结构中是否存在与该列名对应的索引。

[0036] 由上述可知，本发明的技术方案，通过为SparkSQL系统中的数据表的指定列建立索引，在接收到对数据表的查询请求时，判断该请求是否命中具有索引的列，如果命中，则使用命中的列的索引完成本次查询请求；如果未命中，则通过遍历查询请求对应的数据表完成本次查询请求。该技术方案通过为SparkSQL系统中数据表的指定列建立索引的方式显著提高了对SparkSQL系统中数据表的查询速度，在大数据场景下，如果查询请求命中具有索引的列，相比于利用SparkSQL提供的查询方式会有指数级的性能提升，对大数据查询具有极大意义。

[0037] 上述说明仅是本发明技术方案的概述，为了能够更清楚了解本发明的技术手段，而可依照说明书的内容予以实施，并且为了让本发明的上述和其它目的、特征和优点能够更明显易懂，以下特举本发明的具体实施方式。

附图说明

[0038] 通过阅读下文优选实施方式的详细描述，各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的，而并不认为是对本发明的限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

[0039] 图1示出了根据本发明一个实施例的一种SparkSQL系统中的数据处理方法的流程示意图；

[0040] 图2示出了根据本发明一个实施例的一种SparkSQL系统中的数据处理装置的结构示意图。

具体实施方式

[0041] 下面将参照附图更详细地描述本公开的示例性实施例。虽然附图中显示了本公开的示例性实施例，然而应当理解，可以以各种形式实现本公开而不应被这里阐述的实施例所限制。相反，提供这些实施例是为了能够更透彻地理解本公开，并且能够将本公开的范围完整的传达给本领域的技术人员。

[0042] 图1示出了根据本发明一个实施例的一种SparkSQL系统中的数据处理方法的流程示意图，如图1所示，该方法包括：

[0043] 步骤S110，为SparkSQL系统中的数据表的指定列建立索引。

[0044] 步骤S120，当接收到对数据表的查询请求时，判断该请求是否命中具有索引的列。

[0045] 步骤S130，如果命中，则使用命中的列的索引完成本次查询请求。

[0046] 步骤S140，如果未命中，则通过遍历查询请求对应的数据表完成本次查询请求。

[0047] 可见，图1所示的方法，通过为SparkSQL系统中的数据表的指定列建立索引，在接收到对数据表的查询请求时，判断该请求是否命中具有索引的列，如果命中，则使用命中的列的索引完成本次查询请求；如果未命中，则通过遍历查询请求对应的数据表完成本次查询请求。该技术方案通过为SparkSQL系统中数据表的指定列建立索引的方式显著提高了对SparkSQL系统中数据表的查询速度，在大数据场景下，如果查询请求命中具有索引的列，相比于利用SparkSQL提供的查询方式会有指数级的性能提升，对大数据查询具有极大意义。

[0048] 在本发明的一个实施例中，图1所示的方法中，为SparkSQL系统中的数据表的指定列建立索引包括：为数据表的指定列建立JSON格式的索引，索引包括：索引号和数据位置标识。

[0049] 每条索引对应于数据表中的一条数据，JSON格式的索引可以方便地建立和修改。具体地，在本发明的一个实施例中，上述方法中，索引号包括：指定列的列名、该条索引对应的列值；数据位置标识包括：该数据表的存储路径、该条索引对应的数据在该数据表中的偏移量。

[0050] 例如，下面给出了一条索引的示例：

[0051] {index:{name:"tom"},value:"/home/user/file.txt|0"}

[0052] 其中，index为索引号，包括name这个指定列的列名，以及tom这个与该条索引对应的列值。value为数据位置标识，其中“/home/user/file.txt”部分为该数据表的存储路径，而0则代表该条索引对应的数据在该数据表中的偏移量为0，也就是第1行。该条索引代表着

在user表中,name为Tom的数据的存储位置为/home/user/file.txt中的第1行。

[0053] 当然,也可以为多列建立索引,例如,下面还给出了另一条索引的示例:

[0054] {index:{name:"tom",age:"18"},value:"/home/user/file.txt|0"}

[0055] 该条索引代表着在user表中,name为Tom且age为18的数据的存储位置为/home/user/file.txt中的第1行。

[0056] 在本发明的一个实施例中,上述方法中,为SparkSQL系统中的数据表的指定列建立索引还包括:将建立的索引保存在指定搜索服务器的文件系统中。

[0057] 其中,搜索服务器可以为Elastic Search搜索服务器,该服务器提供了基于RESTful web接口的分布式多用户能力的全文搜索引擎,并且还可以保存建立的索引。那么在本发明的一个实施例中,上述方法中,使用命中的列的索引完成本次查询请求包括:将查询请求发送至指定搜索服务器,接收指定搜索服务器返回的查询结果;其中,查询结果为空,或者为一个或多个数据位置标识;当查询结果不为空时,根据指定搜索服务器返回的一个或多个数据位置标识,从数据表中读取与相应数据位置标识匹配的数据。

[0058] 可以看到,当查询请求命中索引时,将该请求发送至搜索服务器,由该服务器进行索引的搜索并直接返回查询结果,如果查询结果不为空,那么具体地查询结果为数据位置标识,可以直接定位数据的位置。例如,在检索user表中名字为Tom且年龄为18的数据时,由于该查询请求命中了name和age列,而已经为这两个列建立了索引,那么举例而言仅查询到如下索引:{index:{name:"tom",age:"18"},value:"/home/user/file.txt|0"},那么可以直接提取/home/user/file.txt|0作为查询结果,根据该查询结果,再调用SparkSQL的接口直接从HDFS中读取到数据表中相应的数据,就完成了本次查询。

[0059] 在本发明的一个实施例中,上述方法中,判断该请求是否命中具有索引的列包括:从查询请求中解析出待查询的表名和列名;根据待查询的表名,判断相应的数据表是否存在与待查询的列名对应的索引,若存在则判断为命中,若不存在则判断为未命中。

[0060] 由于SparkSQL使用类似SQL的查询语句,因此也会对查询语句进行解析。例如查询请求为:select*from user where name="TOM"and age=18,首先对查询请求进行AST(abstract syntax tree,抽象语法树)解析,得到查询请求的树形结构的表现形式,其中每一个节点为一个查询请求中的一个单词,而树的结构体现了查询请求的语法。进一步地,根据树形结构生成逻辑查询计划,可以对查询请求进行一些优化,再进一步生成物理查询计划。在这个过程中,就可以根据解析出的待查询的表名,判断相应的数据表是否存在与待查询的列名对应的索引。根据出的表名user和解析出的列名name和age,可以判断请求命中为这两列建立的索引。

[0061] 可以看出,在请求的解析过程中完成了对查询请求是否命中索引的判断。而如何判断相应的数据表是否存在与待查询的列名对应的索引,下面给出了一种示例:

[0062] 在本发明的一个实施例中,上述方法中,为SparkSQL系统中的数据表的指定列建立索引还包括:在该数据表的表结构中标识已建立索引的列;根据待查询的表名,判断相应的数据表是否存在与待查询的列名对应的索引包括:根据待查询的表名,从相应数据表的表结构中读取该数据表已建立的索引的列,根据待查询的列名判断表结构中是否存在与该列名对应的索引。

[0063] 在SparkSQL系统中,数据表的实数据是存储在HDFS中的,也就是待查询的对象;而

元数据中记录表结构，对该表结构进行修改，可以标识已建立索引的列。这样在解析请求得到待查询的表名和列名后，查询与表名对应的表结构，就可以判断是否已经为待查询的列名建立了索引。

[0064] 在上述实施例中，索引的建立可以是在建表时就确定的，也可以是在建表之后再选择的。当数据被修改时，相关的索引也需要被修改，例如，索引{index:{name:"tom",age:"18"},value:"/home/user/file.txt|0"}对应的数据将age的列值修改为17，那么该条索引会被修改为{index:{name:"tom",age:"17"},value:"/home/user/file.txt|0"}。

[0065] 图2示出了根据本发明一个实施例的一种SparkSQL系统中的数据处理装置的结构示意图，如图2所示，SparkSQL系统中的数据处理装置200包括：

[0066] 索引建立单元210，适于为SparkSQL系统中的数据表的指定列建立索引。

[0067] 请求处理单元220，适于在接收到对数据表的查询请求时，判断该请求是否命中具有索引的列；如果命中，则使用命中的列的索引完成本次查询请求；如果未命中，则通过遍历查询请求对应的数据表完成本次查询请求。

[0068] 可见，图2所示的装置，通过各单元的相互配合，为SparkSQL系统中的数据表的指定列建立索引，在接收到对数据表的查询请求时，判断该请求是否命中具有索引的列，如果命中，则使用命中的列的索引完成本次查询请求；如果未命中，则通过遍历查询请求对应的数据表完成本次查询请求。该技术方案通过为SparkSQL系统中数据表的指定列建立索引的方式显著提高了对SparkSQL系统中数据表的查询速度，在大数据场景下，如果查询请求命中具有索引的列，相比于利用SparkSQL提供的查询方式会有指数级的性能提升，对大数据查询具有极大意义。

[0069] 在本发明的一个实施例中，图2所示的装置中，索引建立单元210，适于为数据表的指定列建立JSON格式的索引，索引包括：索引号和数据位置标识。

[0070] 在本发明的一个实施例中，上述装置中，索引号包括：指定列的列名、该条索引对应的列值；数据位置标识包括：该数据表的存储路径、该条索引对应的数据在该数据表中的偏移量。

[0071] 在本发明的一个实施例中，上述装置中，索引建立单元210，适于将建立的索引保存在指定搜索服务器的文件系统中。

[0072] 在本发明的一个实施例中，上述装置中，请求处理单元220，适于将查询请求发送至指定搜索服务器，接收指定搜索服务器返回的查询结果；其中，查询结果为空，或者为一个或多个数据位置标识；当查询结果不为空时，根据指定搜索服务器返回的一个或多个数据位置标识，从数据表中读取与相应数据位置标识匹配的数据。

[0073] 在本发明的一个实施例中，上述装置中，请求处理单元220，适于从查询请求中解析出待查询的表名和列名；根据待查询的表名，判断相应的数据表是否存在与待查询的列名对应的索引，若存在则判断为命中，若不存在则判断为未命中。

[0074] 在本发明的一个实施例中，上述装置中，索引建立单元210，还适于在该数据表的表结构中标识已建立索引的列；请求处理单元220，适于根据待查询的表名，从相应数据表的表结构中读取该数据表已建立的索引的列，根据待查询的列名判断表结构中是否存在与该列名对应的索引。

[0075] 需要说明的是，上述各装置实施例的具体实施方式与前述对应方法实施例的具体

实施方式相同，在此不再赘述。

[0076] 综上所述，本发明的技术方案，通过为SparkSQL系统中的数据表的指定列建立索引，在接收到对数据表的查询请求时，判断该请求是否命中具有索引的列，如果命中，则使用命中的列的索引完成本次查询请求；如果未命中，则通过遍历查询请求对应的数据表完成本次查询请求。该技术方案通过为SparkSQL系统中数据表的指定列建立索引的方式显著提高了对SparkSQL系统中数据表的查询速度，在大数据场景下，如果查询请求命中具有索引的列，相比于利用SparkSQL提供的查询方式会有指数级的性能提升，对大数据查询具有极大意义。

[0077] 需要说明的是：

[0078] 在此提供的算法和显示不与任何特定计算机、虚拟装置或者其它设备固有相关。各种通用装置也可以与基于在此的示教一起使用。根据上面的描述，构造这类装置所要求的结构是显而易见的。此外，本发明也不针对任何特定编程语言。应当明白，可以利用各种编程语言实现在此描述的本发明的内容，并且上面对特定语言所做的描述是为了披露本发明的最佳实施方式。

[0079] 在此处所提供的说明书中，说明了大量具体细节。然而，能够理解，本发明的实施例可以在没有这些具体细节的情况下实践。在一些实例中，并未详细示出公知的方法、结构和技术，以便不模糊对本说明书的理解。

[0080] 类似地，应当理解，为了精简本公开并帮助理解各个发明方面中的一个或多个，在上面对本发明的示例性实施例的描述中，本发明的各个特征有时被一起分组到单个实施例、图、或者对其的描述中。然而，并不应将该公开的方法解释成反映如下意图：即所要求保护的本发明要求比在每个权利要求中所明确记载的特征更多的特征。更确切地说，如下面的权利要求书所反映的那样，发明方面在于少于前面公开的单个实施例的所有特征。因此，遵循具体实施方式的权利要求书由此明确地并入该具体实施方式，其中每个权利要求本身都作为本发明的单独实施例。

[0081] 本领域那些技术人员可以理解，可以对实施例中的设备中的模块进行自适应性地改变并且把它们设置在与该实施例不同的一个或多个设备中。可以把实施例中的模块或单元或组件组合成一个模块或单元或组件，以及此外可以把它们分成多个子模块或子单元或子组件。除了这样的特征和/或过程或者单元中的至少一些是相互排斥之外，可以采用任何组合对本说明书(包括伴随的权利要求、摘要和附图)中公开的所有特征以及如此公开的任何方法或者设备的所有过程或单元进行组合。除非另外明确陈述，本说明书(包括伴随的权利要求、摘要和附图)中公开的每个特征可以由提供相同、等同或相似目的的替代特征来代替。

[0082] 此外，本领域的技术人员能够理解，尽管在此所述的一些实施例包括其它实施例中所包括的某些特征而不是其它特征，但是不同实施例的特征的组合意味着处于本发明的范围之内并且形成不同的实施例。例如，在下面的权利要求书中，所要求保护的实施例的任意之一都可以以任意的组合方式来使用。

[0083] 本发明的各个部件实施例可以以硬件实现，或者以在一个或者多个处理器上运行的软件模块实现，或者以它们的组合实现。本领域的技术人员应当理解，可以在实践中使用微处理器或者数字信号处理器(DSP)来实现根据本发明实施例的SparkSQL系统中的数据处

理装置中的一些或者全部部件的一些或者全部功能。本发明还可以实现为用于执行这里所描述的方法的一部分或者全部的设备或者装置程序(例如,计算机程序和计算机程序产品)。这样的实现本发明的程序可以存储在计算机可读介质上,或者可以具有一个或者多个信号的形式。这样的信号可以从因特网网站上下载得到,或者在载体信号上提供,或者以任何其他形式提供。

[0084] 应该注意的是上述实施例对本发明进行说明而不是对本发明进行限制,并且本领域技术人员在不脱离所附权利要求的范围的情况下可设计出替换实施例。在权利要求中,不应将位于括号之间的任何参考符号构造成对权利要求的限制。单词“包含”不排除存在未列在权利要求中的元件或步骤。位于元件之前的单词“一”或“一个”不排除存在多个这样的元件。本发明可以借助于包括有若干不同元件的硬件以及借助于适当编程的计算机来实现。在列举了若干装置的单元权利要求中,这些装置中的若干个可以是通过同一个硬件项来具体体现。单词第一、第二、以及第三等的使用不表示任何顺序。可将这些单词解释为名称。

[0085] 本发明的实施例公开了A1、一种SparkSQL系统中的数据处理方法,其中,该方法包括:

[0086] 为SparkSQL系统中的数据表的指定列建立索引;

[0087] 当接收到对数据表的查询请求时,判断该请求是否命中具有索引的列;

[0088] 如果命中,则使用命中的列的索引完成本次查询请求;

[0089] 如果未命中,则通过遍历所述查询请求对应的数据表完成本次查询请求。

[0090] A2、如A1所述的方法,其中,所述为SparkSQL系统中的数据表的指定列建立索引包括:

[0091] 为数据表的指定列建立JSON格式的索引,所述索引包括:索引号和数据位置标识。

[0092] A3、如A2所述的方法,其中,

[0093] 所述索引号包括:指定列的列名、该条索引对应的列值;

[0094] 所述数据位置标识包括:该数据表的存储路径、该条索引对应的数据在该数据表中的偏移量。

[0095] A4、如A2或A3所述的方法,其中,所述为SparkSQL系统中的数据表的指定列建立索引还包括:

[0096] 将建立的索引保存在指定搜索服务器的文件系统中。

[0097] A5、如A4所述的方法,其中,所述使用命中的列的索引完成本次查询请求包括:

[0098] 将所述查询请求发送至所述指定搜索服务器,接收所述指定搜索服务器返回的查询结果;其中,所述查询结果为空,或者为一个或多个数据位置标识;

[0099] 当所述查询结果不为空时,根据所述指定搜索服务器返回的一个或多个数据位置标识,从数据表中读取与相应数据位置标识匹配的数据。

[0100] A6、如A1-A5中任一项所述的方法,其中,所述判断该请求是否命中具有索引的列包括:

[0101] 从所述查询请求中解析出待查询的表名和列名;

[0102] 根据所述待查询的表名,判断相应的数据表是否存在与待查询的列名对应的索引,若存在则判断为命中,若不存在则判断为未命中。

[0103] A7、如A6所述的方法，其中，所述为SparkSQL系统中的数据表的指定列建立索引还包括：

[0104] 在该数据表的表结构中标识已建立索引的列；

[0105] 所述根据所述待查询的表名，判断相应的数据表是否存在与待查询的列名对应的索引包括：

[0106] 根据所述待查询的表名，从相应数据表的表结构中读取该数据表已建立的索引的列，根据所述待查询的列名判断所述表结构中是否存在与该列名对应的索引。

[0107] 本发明的实施例还公开了B8、一种SparkSQL系统中的数据处理装置，其中，该装置包括：

[0108] 索引建立单元，适于为SparkSQL系统中的数据表的指定列建立索引；

[0109] 请求处理单元，适于在接收到对数据表的查询请求时，判断该请求是否命中具有索引的列；如果命中，则使用命中的列的索引完成本次查询请求；如果未命中，则通过遍历所述查询请求对应的数据表完成本次查询请求。

[0110] B9、如B8所述的装置，其中，

[0111] 所述索引建立单元，适于为数据表的指定列建立JSON格式的索引，所述索引包括：索引号和数据位置标识。

[0112] B10、如B9所述的装置，其中，

[0113] 所述索引号包括：指定列的列名、该条索引对应的列值；

[0114] 所述数据位置标识包括：该数据表的存储路径、该条索引对应的数据在该数据表中的偏移量。

[0115] B11、如B9或B10所述的装置，其中，

[0116] 所述索引建立单元，适于将建立的索引保存在指定搜索服务器的文件系统中。

[0117] B12、如B11所述的装置，其中，

[0118] 所述请求处理单元，适于将所述查询请求发送至所述指定搜索服务器，接收所述指定搜索服务器返回的查询结果；其中，所述查询结果为空，或者为一个或多个数据位置标识；当所述查询结果不为空时，根据所述指定搜索服务器返回的一个或多个数据位置标识，从数据表中读取与相应数据位置标识匹配的数据。

[0119] B13、如B8-B12中任一项所述的装置，其中，

[0120] 所述请求处理单元，适于从所述查询请求中解析出待查询的表名和列名；根据所述待查询的表名，判断相应的数据表是否存在与待查询的列名对应的索引，若存在则判断为命中，若不存在则判断为未命中。

[0121] B14、如B13所述的装置，其中，

[0122] 所述索引建立单元，还适于在该数据表的表结构中标识已建立索引的列；

[0123] 所述请求处理单元，适于根据所述待查询的表名，从相应数据表的表结构中读取该数据表已建立的索引的列，根据所述待查询的列名判断所述表结构中是否存在与该列名对应的索引。

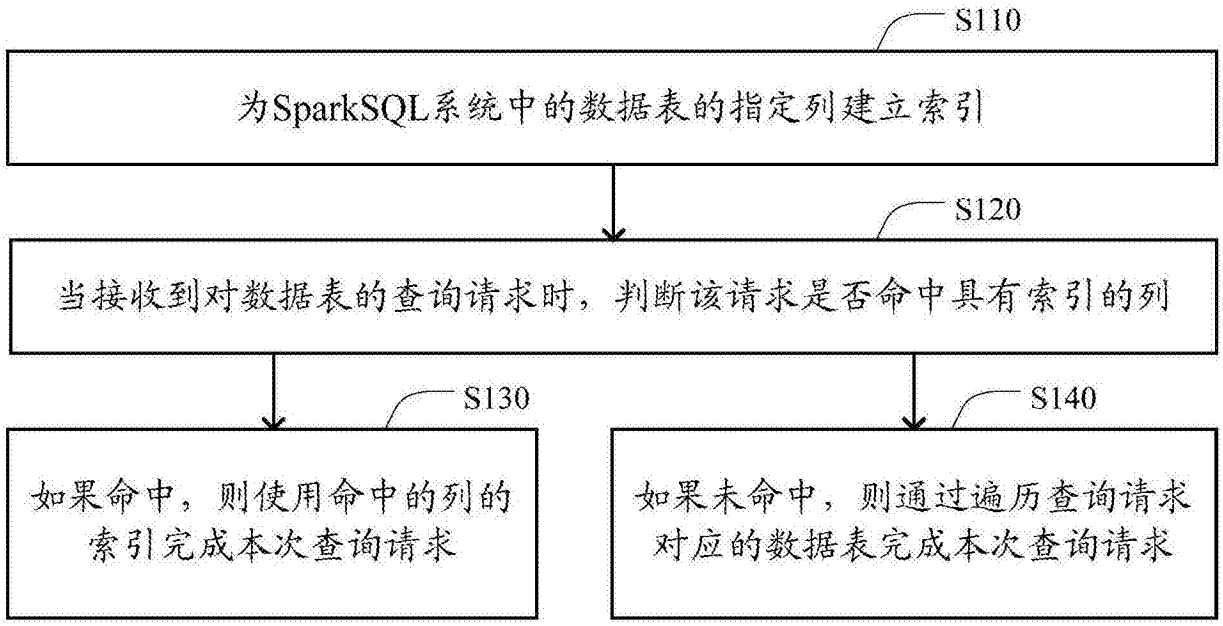

S110
为SparkSQL系统中的数据表的指定列建立索引
S120
当接收到对数据表的查询请求时，判断该请求是否命中具有索引的列
S130
如果命中，则使用命中的列的索引完成本次查询请求
S140
如果未命中，则通过遍历查询请求对应的数据表完成本次查询请求

图1

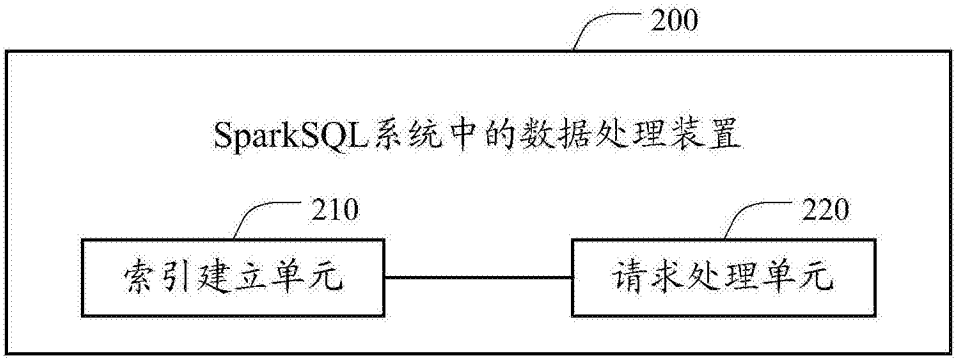

200
SparkSQL系统中的数据处理装置
210
索引建立单元
220
请求处理单元

图2