(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 105589920 B
(45)授权公告日 2019.10.01

(21)申请号 201510630493.1

(22)申请日 2015.09.29

(65)同一申请的已公布的文献号
申请公布号 CN 105589920 A

(43)申请公布日 2016.05.18

(73)专利权人 中国银联股份有限公司
地址 200135 上海市浦东新区含笑路36号银联大厦

(72)发明人 何东杰

(74)专利代理机构 中国专利代理(香港)有限公司 72001
代理人 方世栋 付曼

(51)Int.Cl.
G06F 16/2453(2019.01)
G06F 16/245(2019.01)

(56)对比文件
CN 102819589 A,2012.12.12,
CN 102737033 A,2012.10.17,

审查员 郑岩

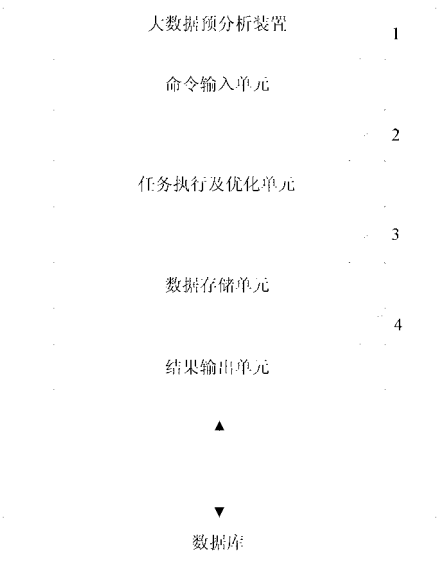
权利要求书1页 说明书4页 附图2页

(54)发明名称

用于大数据预分析的方法和装置

(57)摘要

本发明提出了一种用于大数据预分析的方法和装置,所述方法包括:接收来自用户的数据查询及分析命令,其中,所述用户能够选择式地或者直接地输入所述数据查询及分析命令;解析所述数据查询及分析命令以确定其定义的数据处理任务,并随之基于预定规则优化并执行所述数据处理任务;基于存储优化算法存储所述数据处理任务执行过程中使用的数据;向用户输出所述数据处理任务的执行结果。本发明所公开的用于大数据预分析的方法和装置能够显著地提高大数据预分析结果的有效性和准确性。



CN 105589920 B

1.一种大数据预分析装置,所述大数据预分析装置包括:

命令输入单元,所述命令输入单元接收来自用户的数据查询及分析命令,其中,所述用户能够选择式地或者直接地输入所述数据查询及分析命令;

任务执行及优化单元,所述任务执行及优化单元解析所述数据查询及分析命令以确定其定义的数据处理任务,并随之基于预定规则优化并执行所述数据处理任务;

数据存储单元,所述数据存储单元基于存储优化算法存储所述数据处理任务执行过程中使用的数据;

结果输出单元,所述结果输出单元向用户输出所述数据处理任务的执行结果,

其中,基于预定规则优化并执行所述数据处理任务的步骤包括:根据待执行任务的具体操作并基于预统计的结果进行优化,其中,所述预统计被周期性地或不定期的执行以识别数据取值分布以及数据表的数据量大小,并且所述优化限定针对表关联操作优先加载数据量小的数据表并且限定针对数据过滤操作优先针对取值分布较多的字段进行过滤。

2.根据权利要求1所述的大数据预分析装置,其特征在于,所述数据查询及分析命令是基于SQL语言的命令。

3.根据权利要求1所述的大数据预分析装置,其特征在于,基于预定规则优化并执行所述数据处理任务的步骤进一步包括:根据待执行任务操作的数据字段的数量选择不同存储方式的数据进行操作,即当待执行任务操作的数据字段的数量小于预定阈值时选择列式存储的数据,而当待执行任务操作的数据字段的数量不小于所述预定阈值时选择行列混合式存储的数据。

4.根据权利要求3所述的大数据预分析装置,其特征在于,所述数据存储单元自动地确定所述数据处理任务执行过程中不同操作使用列式存储的数据和行列混合式存储的数据时的处理效率,并根据所确定的处理效率针对特定的数据字段执行数据优化操作,其中所述数据优化操作包括:(1)对经常处理分析的数据字段进行压缩;(2)对关联的数据字段进行联合存储。

5.根据权利要求4所述的大数据预分析装置,其特征在于,所述结果输出单元能够通过显示器向用户呈现所述数据处理任务的执行结果,并且能够基于用户的指令提供所述数据处理任务的执行结果的下载服务。

6.一种用于大数据预分析的方法,所述用于大数据预分析的方法包括下列步骤:

(A1)接收来自用户的数据查询及分析命令,其中,所述用户能够选择式地或者直接地输入所述数据查询及分析命令;

(A2)解析所述数据查询及分析命令以确定其定义的数据处理任务,并随之基于预定规则优化并执行所述数据处理任务;

(A3)基于存储优化算法存储所述数据处理任务执行过程中使用的数据;

(A4)向用户输出所述数据处理任务的执行结果,

其中,基于预定规则优化并执行所述数据处理任务的步骤包括:根据待执行任务的具体操作并基于预统计的结果进行优化,其中,所述预统计被周期性地或不定期的执行以识别数据取值分布以及数据表的数据量大小,并且所述优化限定针对表关联操作优先加载数据量小的数据表并且限定针对数据过滤操作优先针对取值分布较多的字段进行过滤。

用于大数据预分析的方法和装置

技术领域

[0001] 本发明涉及数据分析方法和装置，更具体地，涉及用于大数据预分析的方法和装置。

背景技术

[0002] 目前，随着计算机和网络应用的日益广泛以及不同领域的业务种类的日益丰富，在实际使用海量数据(即大数据)之前对其进行预分析变得越来越重要。

[0003] 在现有的技术方案中，通常采用数据抽样方式对大数据进行预分析(例如分析目标数据的内容、分布、关联关系等等)，即从目标大数据中随机地或基于预定规则抽取样本数据，并随之针对该样本数据执行分析操作。

[0004] 然而，现有的技术方案存在如下问题：由于基于样本数据执行数据预分析，故预分析结果的准确性直接取决于所抽取的样本数据的质量和代表性，由此预分析结果的有效性和准确性难于控制并且是不稳定的。

[0005] 因此，存在如下需求：提供能够显著地提高大数据预分析结果的有效性和准确性的用于大数据预分析的方法和装置。

发明内容

[0006] 为了解决上述现有技术方案所存在的问题，本发明提出了能够显著地提高大数据预分析结果的有效性和准确性的用于大数据预分析的方法和装置。

[0007] 本发明的目的是通过以下技术方案实现的：

[0008] 一种大数据预分析装置，所述大数据预分析装置包括：

[0009] 命令输入单元，所述命令输入单元接收来自用户的数据查询及分析命令，其中，所述用户能够选择式地或者直接地输入所述数据查询及分析命令；

[0010] 任务执行及优化单元，所述任务执行及优化单元解析所述数据查询及分析命令以确定其定义的数据处理任务，并随之基于预定规则优化并执行所述数据处理任务；

[0011] 数据存储单元，所述数据存储单元基于存储优化算法存储所述数据处理任务执行过程中使用的数据；

[0012] 结果输出单元，所述结果输出单元向用户输出所述数据处理任务的执行结果。

[0013] 在上面所公开的方案中，优选地，所述数据查询及分析命令是基于SQL语言的命令。

[0014] 在上面所公开的方案中，优选地，所述优化所述数据处理任务包括：根据待执行任务的具体操作并基于预统计的结果优化各个操作执行的先后顺序以及数据加载的顺序，其中，所述预统计被周期性地或不定期的执行以识别数据取值分布以及数据表的数据量大小，并且所述优化限定针对表关联操作优先加载数据量小的数据表并且限定针对数据过滤操作优先针对取值分布较多的字段进行过滤。

[0015] 在上面所公开的方案中，优选地，所述优化所述数据处理任务进一步包括：根据待

执行任务操作的数据字段的数量选择不同存储方式的数据进行操作，即当待执行任务操作的数据字段的数量小于预定阈值时选择列式存储的数据，而当待执行任务操作的数据字段的数量不小于所述预定阈值时选择行列混合式存储的数据。

[0016] 在上面所公开的方案中，优选地，所述数据存储单元自动地确定所述数据处理任务执行过程中不同操作使用列式存储的数据和行列混合式存储的数据时的处理效率，并根据所确定的处理效率针对特定的数据字段执行数据优化操作，其中所述数据优化操作包括：(1)对经常处理分析的数据字段进行压缩；(2)对关联的数据字段进行联合存储。

[0017] 在上面所公开的方案中，优选地，所述结果输出单元能够通过显示器向用户呈现所述数据处理任务的执行结果，并且能够基于用户的指令提供所述数据处理任务的执行结果的下载服务。

[0018] 本发明的目的也能够通过以下技术方案实现：

[0019] 一种用于大数据预分析的方法，所述用于大数据预分析的方法包括下列步骤：

[0020] (A1)接收来自用户的数据查询及分析命令，其中，所述用户能够选择式地或者直接地输入所述数据查询及分析命令；

[0021] (A2)解析所述数据查询及分析命令以确定其定义的数据处理任务，并随之基于预定规则优化并执行所述数据处理任务；

[0022] (A3)基于存储优化算法存储所述数据处理任务执行过程中使用的数据；

[0023] (A4)向用户输出所述数据处理任务的执行结果

[0024] 本发明所公开的用于大数据预分析的方法和装置具有以下优点：能够显著地提高大数据预分析结果的有效性和准确性，并且提高了数据处理操作的效率。

附图说明

[0025] 结合附图，本发明的技术特征以及优点将会被本领域技术人员更好地理解，其中：

[0026] 图1是根据本发明的实施例的大数据预分析装置的示意性结构图；

[0027] 图2是根据本发明的实施例的用于大数据预分析的方法的流程图。

具体实施方式

[0028] 图1是根据本发明的实施例的大数据预分析装置的示意性结构图。如图1所示，本发明所公开的大数据预分析装置包括命令输入单元1、任务执行及优化单元2、数据存储单元3以及结果输出单元4。其中，所述命令输入单元1接收来自用户的数据查询及分析命令，其中，所述用户能够选择式地(例如通过下拉框选择)或者直接地输入所述数据查询及分析命令。所述任务执行及优化单元2解析(例如语句的拼装和拆分)所述数据查询及分析命令以确定其定义的数据处理任务，并随之基于预定规则优化并执行所述数据处理任务。所述数据存储单元3基于存储优化算法存储所述数据处理任务执行过程中使用的数据。所述结果输出单元4向用户输出所述数据处理任务的执行结果。

[0029] 优选地，在本发明所公开的大数据预分析装置中，所述数据查询及分析命令是基于SQL(结构化查询语言)的命令。

[0030] 优选地，在本发明所公开的大数据预分析装置中，所述优化所述数据处理任务包括：根据待执行任务的具体操作并基于预统计的结果优化各个操作执行的先后顺序以及数

据加载的顺序，其中，所述预统计被周期性地或不定期的执行以识别数据取值分布以及数据表的数据量大小，并且所述优化限定针对表关联操作优先加载数据量小的数据表并且限定针对数据过滤操作优先针对取值分布较多的字段进行过滤。

[0031] 优选地，在本发明所公开的大数据预分析装置中，所述优化所述数据处理任务进一步包括：根据待执行任务操作的数据字段的数量选择不同存储方式的数据进行操作，即当待执行任务操作的数据字段的数量小于预定阈值(例如15个数据字段)时选择列式存储的数据，而当待执行任务操作的数据字段的数量不小于预定阈值(例如15个数据字段)时选择行列混合式存储的数据。

[0032] 优选地，在本发明所公开的大数据预分析装置中，所述数据存储单元3自动地确定所述数据处理任务执行过程中不同操作使用列式存储的数据和行列混合式存储的数据时的处理效率，并根据所确定的处理效率针对特定的数据字段执行数据优化操作，其中所述数据优化操作包括：(1)对经常处理分析的数据字段进行压缩(例如，在金融领域中，对卡号字段的值进行数值转换并将其压缩成哈夫曼编码)；(2)对关联的数据字段进行联合存储(例如，在金融领域中，卡品牌和卡属性经常同时出现，则将这两个数据字段进行组合后存储)。

[0033] 优选地，在本发明所公开的大数据预分析装置中，所述结果输出单元4能够通过显示器向用户呈现所述数据处理任务的执行结果，并且能够基于用户的指令提供所述数据处理任务的执行结果的下载服务。

[0034] 由上可见，本发明所公开的大数据预分析装置具有下列优点：能够显著地提高大数据预分析结果的有效性和准确性，并且提高了数据处理操作的效率。

[0035] 图2是根据本发明的实施例的用于大数据预分析的方法的流程图。如图2所示，本发明所公开的用于大数据预分析的方法包括下列步骤：(A1)接收来自用户的数据查询及分析命令，其中，所述用户能够选择式地(例如通过下拉框选择)或者直接地输入所述数据查询及分析命令；(A2)解析(例如语句的拼装和拆分)所述数据查询及分析命令以确定其定义的数据处理任务，并随之基于预定规则优化并执行所述数据处理任务；(A3)基于存储优化算法存储所述数据处理任务执行过程中使用的数据；(A4)向用户输出所述数据处理任务的执行结果。

[0036] 优选地，在本发明所公开的用于大数据预分析的方法中，所述数据查询及分析命令是基于SQL(结构化查询语言)的命令。

[0037] 优选地，在本发明所公开的用于大数据预分析的方法中，所述优化所述数据处理任务包括：根据待执行任务的具体操作并基于预统计的结果优化各个操作执行的先后顺序以及数据加载的顺序，其中，所述预统计被周期性地或不定期的执行以识别数据取值分布以及数据表的数据量大小，并且所述优化限定针对表关联操作优先加载数据量小的数据表并且限定针对数据过滤操作优先针对取值分布较多的字段进行过滤。

[0038] 优选地，在本发明所公开的用于大数据预分析的方法中，所述优化所述数据处理任务进一步包括：根据待执行任务操作的数据字段的数量选择不同存储方式的数据进行操作，即当待执行任务操作的数据字段的数量小于预定阈值(例如15个数据字段)时选择列式存储的数据，而当待执行任务操作的数据字段的数量不小于预定阈值(例如15个数据字段)时选择行列混合式存储的数据。

[0039] 优选地，在本发明所公开的用于大数据预分析的方法中，所述步骤(A3)包括：自动地确定所述数据处理任务执行过程中不同操作使用列式存储的数据和行列混合式存储的数据时的处理效率，并根据所确定的处理效率针对特定的数据字段执行数据优化操作，其中所述数据优化操作包括：(1)对经常处理分析的数据字段进行压缩(例如，在金融领域中，对卡号字段的值进行数值转换并将其压缩成哈夫曼编码)；(2)对关联的数据字段进行联合存储(例如，在金融领域中，卡品牌和卡属性经常同时出现，则将这两个数据字段进行组合后存储)。

[0040] 优选地，在本发明所公开的用于大数据预分析的方法中，所述步骤(A4)进一步包括：通过显示器向用户呈现所述数据处理任务的执行结果，并且基于用户的指令提供所述数据处理任务的执行结果的下载服务。

[0041] 由上可见，本发明所公开的用于大数据预分析的方法具有下列优点：能够显著地提高大数据预分析结果的有效性和准确性，并且提高了数据处理操作的效率。

[0042] 尽管本发明是通过上述的优选实施方式进行描述的，但是其实现形式并不局限于上述的实施方式。应该认识到：在不脱离本发明主旨和范围的情况下，本领域技术人员可以对本发明做出不同的变化和修改。

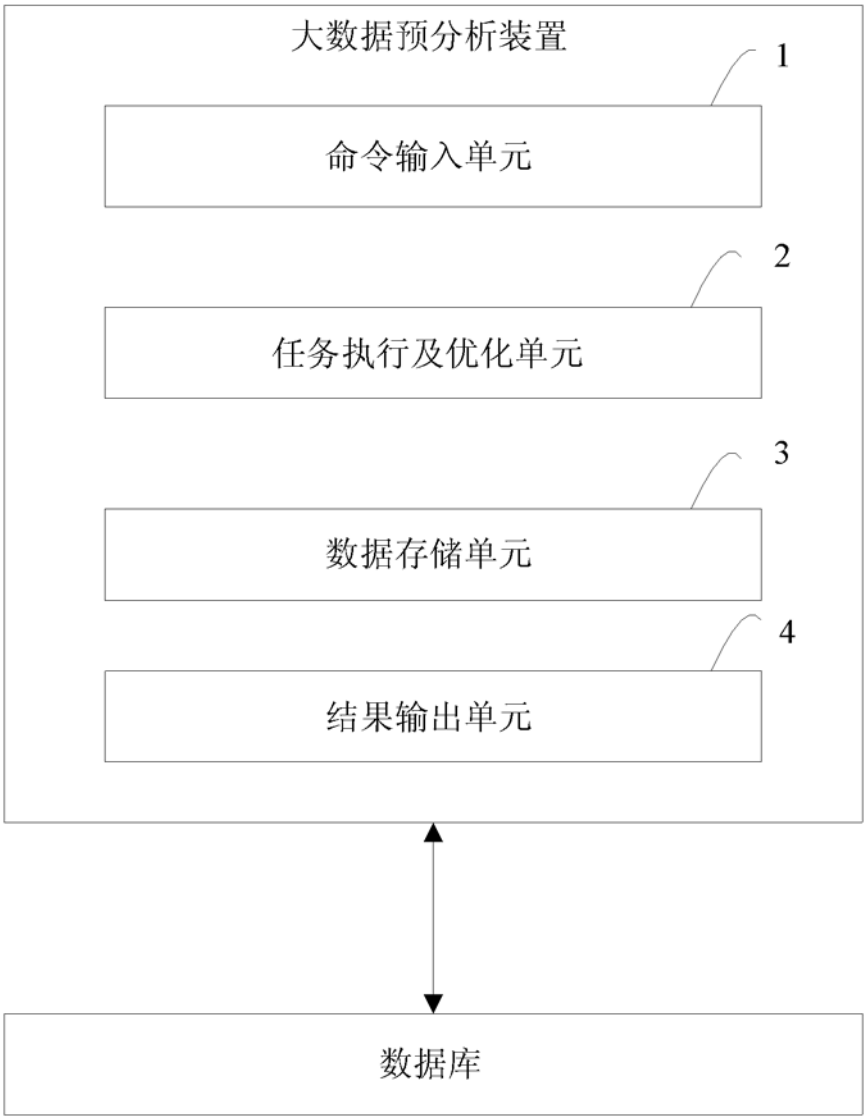


图 1

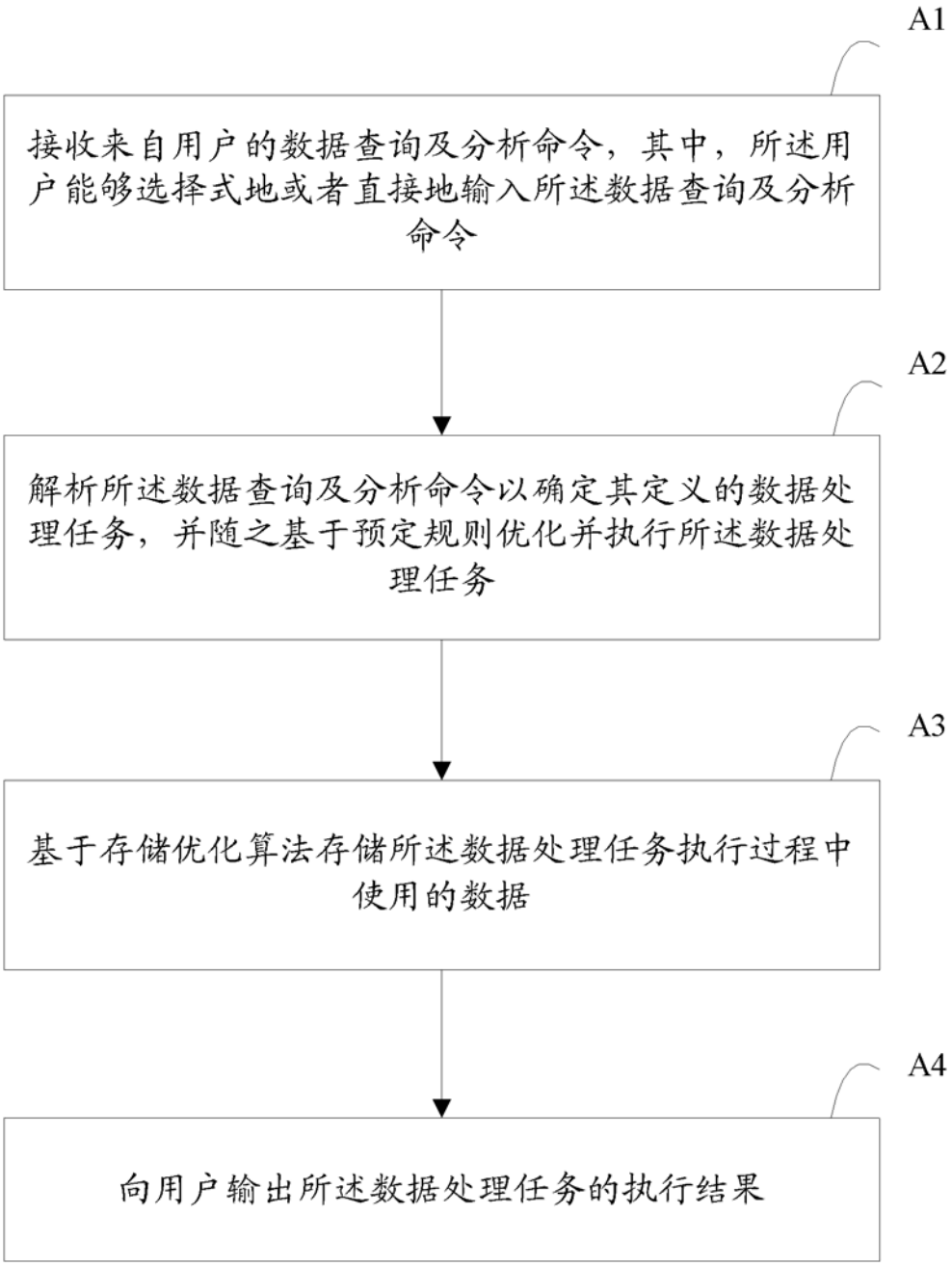

A1
接收来自用户的数据查询及分析命令，其中，所述用户能够选择式地或者直接地输入所述数据查询及分析命令
A2
解析所述数据查询及分析命令以确定其定义的数据处理任务，并随之基于预定规则优化并执行所述数据处理任务
A3
基于存储优化算法存储所述数据处理任务执行过程中使用的数据
A4
向用户输出所述数据处理任务的执行结果

图 2