



(12) 发明专利申请

(10) 申请公布号 CN 114461663 A

(43) 申请公布日 2022. 05. 10

(21) 申请号 202011246333.4

(22) 申请日 2020.11.10

(71) 申请人 华为云计算技术有限公司

地址 550025 贵州省贵阳市贵安新区黔中
大道交兴功路华为云数据中心

(72) 发明人 乔波

(74) 专利代理机构 北京龙双利达知识产权代理
有限公司 11329

专利代理师 周乔 王君

(51) Int. Cl.

G06F 16/242 (2019.01)

G06F 16/22 (2019.01)

G06F 16/28 (2019.01)

G06F 40/205 (2020.01)

权利要求书2页 说明书11页 附图3页

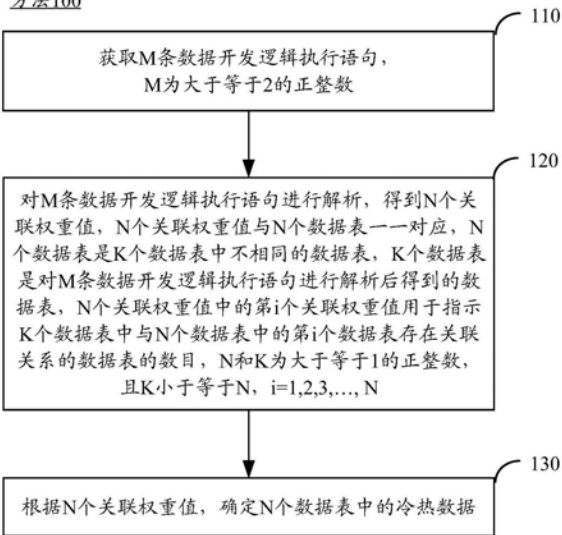
(54) 发明名称

确定冷热数据的方法、装置及系统

(57) 摘要

本申请提供了一种确定冷热数据的方法、装置及系统。该方法包括：获取M条数据开发逻辑执行语句，M为大于等于2的正整数；对M条数据开发逻辑执行语句进行解析，得到N个关联权重值，N个关联权重值与N个数据表一一对应，N个数据表是K个数据表中不相同的数据表，K个数据表是对M条数据开发逻辑执行语句进行解析后得到的数据表，第i个关联权重值用于指示K个数据表中与N个数据表中的第i个数据表存在关联关系的数据表的数目，N和K为大于等于1的正整数，且K小于等于N， $i=1, 2, 3, \dots, N$ ；根据N个关联权重值，确定N个数据表中的冷热数据。该方法在保证冷热数据的确定结果具有较高的准确度的前提下，能够有效降低系统开销。

方法100



1. 一种确定冷热数据的方法,其特征在于,所述方法包括:

获取M条数据开发逻辑执行语句,M为大于等于2的正整数;

对所述M条数据开发逻辑执行语句进行解析,得到N个关联权重值,所述N个关联权重值与N个数据表一一对应,所述N个数据表是K个数据表中不相同的数据表,所述K个数据表是对所述M条数据开发逻辑执行语句进行解析后得到的数据表,所述N个关联权重值中的第i个关联权重值用于指示所述K个数据表中与所述N个数据表中的第i个数据表存在关联关系的数据表的数目,N和K为大于等于1的正整数,且K小于等于N, $i=1,2,3,\dots,N$;

根据所述N个关联权重值,确定所述N个数据表中的冷热数据。

2. 根据权利要求1所述的方法,其特征在于,所述对所述M条数据开发逻辑执行语句进行解析,得到N个关联权重值,包括:

根据所述M条数据开发逻辑执行语句的参数信息对所述M条数据开发逻辑执行语句进行解析,得到每条数据开发逻辑执行语句对应的数据表;

统计所述每条数据开发逻辑执行语句对应的数据表中的与所述第i个数据表存在关联关系的数据表的数目;

根据与所述第i个数据表存在关联关系的数据表的数目,确定所述第i个关联权重值。

3. 根据权利要求2所述的方法,其特征在于,所述M条数据开发逻辑执行语句的参数信息至少包括以下信息:所述M条数据开发逻辑执行语句的操作类型信息以及所述数据开发逻辑执行语句的操作参数信息。

4. 根据权利要求2或3所述的方法,其特征在于,所述根据与所述第i个数据表存在关联关系的数据表的数目,确定所述N个关联权重值,包括:

如果查询到所述K个数据表中与所述第i个数据表存在关联关系的数据表的数目为P,则确定所述第i个关联权重值等于P,P为大于等于零的整数。

5. 根据权利要求1-4任一项所述的方法,其特征在于,所述根据所述N个关联权重值,确定所述N个数据表中的冷热数据,包括:

如果所述第i个关联权重值大于等于预设关联权重值,则确定所述第i个关联权重值对应的数据表包括的数据为热数据, $i=1,2,3,\dots,N$;

如果所述第i个关联权重值小于预设关联权重值,则确定所述第i个关联权重值对应的数据表包括的数据为冷数据。

6. 根据权利要求1-5任一项所述的方法,其特征在于,所述M条数据开发逻辑执行语句不相同。

7. 根据权利要求1-6任一项所述的方法,其特征在于,所述M条数据开发逻辑执行语句包括结构化查询语句SQL。

8. 一种确定冷热数据的装置,其特征在于,包括处理器和存储器,所述存储器用于存储计算机执行指令,所述处理器用于读取所述存储器中存储的所述计算机执行指令,以实现如权利要求1至7中任一项所述的方法。

9. 一种计算机可读存储介质,其特征在于,包括计算机程序,当其在计算机上运行时,使得所述计算机执行如权利要求1至7中任一项所述的方法。

10. 一种芯片系统,其特征在于,包括至少一个处理器和接口;所述至少一个所述处理器,用于调用并运行计算机程序,以使所述芯片系统执行如权利要求1至7中任一项所述的

方法。

确定冷热数据的方法、装置及系统

技术领域

[0001] 本申请涉及存储技术领域,并且更具体地,涉及一种确定冷热数据的方法、装置及系统。

背景技术

[0002] 随着云计算等技术的快速发展和广泛应用,数据呈现出爆炸性增长,但是数据的价值并非完全相同。因此,需要对海量数据(例如,数据湖中的数据)中的冷热数据进行有效区分,基于该区分结果制定更好的存储策略,从而实现节省系统内存的目的。

[0003] 现有技术中,通过访问数据底层引擎,以及根据数据访问的频率,确定数据的冷热程度。但上述方法在确定冷热数据时,存在系统开销大、判断不够准确的问题。

发明内容

[0004] 本申请提供一种确定冷热数据的方法、装置及系统,该方法在保证冷热数据的确定结果具有较高的准确度的前提下,能够有效降低系统开销。

[0005] 第一方面,提供了一种确定冷热数据的方法,其特征在于,该方法包括:

[0006] 获取M条数据开发逻辑执行语句,M为大于等于2的正整数;

[0007] 对该M条数据开发逻辑执行语句进行解析,得到N个关联权重值,该N个关联权重值与N个数据表一一对应,该N个数据表是K个数据表中不相同的数据表,该K个数据表是对该M条数据开发逻辑执行语句进行解析后得到的数据表,该N个关联权重值中的第i个关联权重值用于指示该K个数据表中与该N个数据表中的第i个数据表存在关联关系的数据表的数目,N和K为大于等于1的正整数,且K小于等于N, $i=1,2,3,\dots,N$;

[0008] 根据该N个关联权重值,确定该N个数据表中的冷热数据。

[0009] 其中,对获取M条数据开发逻辑执行语句的数据平台不作具体限定。例如,数据平台可以是数据湖、数据仓库或数据中台等。

[0010] 在上述技术方案中,对获取的数据开发逻辑执行语句直接进行解析,根据解析结果确定数据平台中各数据表之间的关联关系,根据关联关系确定对应的数据表的关联权重值,并根据关联权重值确定数据表的冷热属性。该方法避免了现有技术中需要访问数据底层引擎的操作,在保证冷热数据的确定结果具有较高的准确度的前提下,能够有效降低系统开销。

[0011] 结合第一方面,在第一方面的某些实现方式中,对该M条数据开发逻辑执行语句进行解析,得到N个关联权重值,包括:

[0012] 根据该M条数据开发逻辑执行语句的参数信息对该M条数据开发逻辑执行语句进行解析,得到每条数据开发逻辑执行语句对应的数据表;

[0013] 统计该每条数据开发逻辑执行语句对应的数据表中的与该第i个数据表存在关联关系的数据表的数目;

[0014] 根据与该第i个数据表存在关联关系的数据表的数目,确定该第i个关联权重值。

[0015] 结合第一方面,在第一方面的某些实现方式中,该M条数据开发逻辑执行语句的参数信息至少包括以下信息:该M条数据开发逻辑执行语句的操作类型信息以及该数据开发逻辑执行语句的操作参数信息。

[0016] 结合第一方面,在第一方面的某些实现方式中,根据与该第i个数据表存在关联关系的数据表的数目,确定该N个关联权重值,包括:

[0017] 如果查询到该K个数据表中与该第i个数据表存在关联关系的数据表的数目为P,则确定该第i个关联权重值等于P,P为大于等于零的整数。

[0018] 在上述技术方案中,通过根据确定的数据表之间的关联关系,可以确定每个数据表的关联权重值。

[0019] 结合第一方面,在第一方面的某些实现方式中,根据该N个关联权重值,确定该N个数据表中的冷热数据,包括:

[0020] 如果该第i个关联权重值大于等于预设关联权重值,则确定该第i个关联权重值对应的数据表包括的数据为热数据, $i=1,2,3,\dots,N$;

[0021] 如果该第i个关联权重值小于预设关联权重值,则确定该第i个关联权重值对应的数据表包括的数据为冷数据。

[0022] 其中,预设关联权重值可以根据具体的应用场景进行设置,本申请实施例中对此不作具体限定。

[0023] 在上述技术方案中,通过比较预设关联权重值与每个数据表的关联权重值,能够快速且简单的确定每个数据表的冷热属性。

[0024] 结合第一方面,在第一方面的某些实现方式中,该M条数据开发逻辑执行语句不相同。

[0025] 结合第一方面,在第一方面的某些实现方式中,该M条数据开发逻辑执行语句包括结构化查询语句SQL。

[0026] 第二方面,提供了一种确定冷热数据的装置,所述装置包括用于执行上述第一方面以及第一方面中的任意一种实现方式中的方法的模块或单元。

[0027] 第三方面,提供了一种确定冷热数据的设备,该设备包括:存储器,用于存储程序;处理器,用于执行所述存储器存储的程序,当所述存储器存储的程序被执行时,所述处理器用于执行上述第一方面以及第一方面中的任意一种实现方式中的方法。

[0028] 第四方面,提供一种计算机可读介质,该计算机可读介质存储用于设备执行的程序代码,该程序代码包括用于执行上述第一方面以及第一方面中的任意一种实现方式中的方法。

[0029] 第五方面,提供一种包含指令的计算机程序产品,当该计算机程序产品在计算机上运行时,使得计算机执行上述第一方面以及第一方面中的任意一种实现方式中的方法。

[0030] 第六方面,提供一种芯片,所述芯片包括处理器与数据接口,所述处理器通过所述数据接口读取存储器上存储的指令,执行上述第一方面以及第一方面中的任意一种实现方式中的方法。

[0031] 可选地,作为一种实现方式,所述芯片还可以包括存储器,所述存储器中存储有指令,所述处理器用于执行所述存储器上存储的指令,当所述指令被执行时,所述处理器用于执行上述第一方面以及第一方面中的任意一种实现方式中的方法。

[0032] 上述芯片具体可以是现场可编程门阵列(field-programmable gate array, FPGA)或者专用集成电路(application-specific integrated circuit,ASIC)。

[0033] 第七方面,提供一种系统,包括上述第二方面的确定冷热数据的装置和/或上述第三方面的确定冷热数据的设备。

附图说明

[0034] 图1是本申请实施例提供的一种确定冷热数据的方法100的示意性流程图。

[0035] 图2是本申请实施例提供的另一种确定冷热数据的方法200的示意性流程图。

[0036] 图3是本申请实施例提供的一种确定冷热数据的装置3000的示意性结构图。

[0037] 图4是本申请实施例提供的一种确定冷热数据的设备4000的示意性结构图。

[0038] 图5是本申请实施例提供的一种系统5000的结构示意图。

具体实施方式

[0039] 下面将结合附图,对本申请中的技术方案进行描述。

[0040] 本申请的实施方式部分使用的术语仅用于对本申请的具体实施例进行解释,而非旨在限定本申请。

[0041] 本申请中术语“第一”“第二”“第三”等字样用于对作用和功能基本相同的相同项或相似项进行区分,应理解,“第一”、“第二”和“第三”之间不具有逻辑或时序上的依赖关系,也不对数量和执行顺序进行限定。

[0042] 本申请将围绕可包括多个设备、组件、模块等的系统来呈现各个方面、实施例或特征。应当理解和明白的是,各个系统可以包括另外的设备、组件、模块等,并且/或者可以并不包括结合附图讨论的所有设备、组件、模块等。此外,还可以使用这些方案的组合。

[0043] 另外,在本申请实施例中,“示例的”、“例如”等词用于表示作例子、例证或说明。本申请中被描述为“示例”的任何实施例或设计方案不应被解释为比其它实施例或设计方案更优选或更具优势。确切而言,使用示例一词旨在以具体方式呈现概念。

[0044] 本申请实施例中,“相应的(corresponding, relevant)”和“对应的(corresponding)”有时可以混用,应当指出的是,在不强调其区别时,其所要表达的含义是一致的。

[0045] 本申请实施例中,有时候下标如 W_1 可能会笔误为非下标的形式如 $W1$,在不强调其区别时,其所要表达的含义是一致的。

[0046] 在本说明书中描述的参考“一个实施例”或“一些实施例”等意味着在本申请的一个或多个实施例中包括结合该实施例描述的特定特征、结构或特点。由此,在本说明书中的不同之处出现的语句“在一个实施例中”、“在一些实施例中”、“在其他一些实施例中”、“在另外一些实施例中”等不是必然都参考相同的实施例,而是意味着“一个或多个但不是所有的实施例”,除非是以其他方式另外特别强调。术语“包括”、“包含”、“具有”及它们的变形都意味着“包括但不限于”,除非是以其他方式另外特别强调。

[0047] 本申请中,“至少一个”是指一个或者多个,“多个”是指两个或两个以上。“和/或”,描述关联对象的关联关系,表示可以存在三种关系,例如,A和/或B,可以表示:单独存在A,同时存在A和B,单独存在B的情况,其中A,B可以是单数或者复数。字符“/”一般表示前后关

联对象是一种“或”的关系。“以下至少一项(个)”或其类似表达,是指的这些项中的任意组合,包括单项(个)或复数项(个)的任意组合。例如,a,b,或c中的至少一项(个),可以表示:a,b,c,a-b,a-c,b-c,或a-b-c,其中a,b,c可以是单个,也可以是多个。

[0048] 下面,介绍本申请的相关技术:

[0049] 为便于理解,在描述本申请实施例提供的确定冷热数据的方法之前,首先对本申请实施例中涉及的相关术语进行简单介绍。

[0050] 1、冷热数据

[0051] 冷热数据,是指数据被访问的频率不同。一般而言,被访问频率高的数据称为热数据,被访问频率低的数据称为冷数据。

[0052] 2、结构化查询语言(structured query language,SQL)

[0053] SQL,是一种数据库查询和程序设计语言,用于存取数据以及查询、更新和管理关系数据库系统。

[0054] SQL是高级的非过程化编程语言,允许用户在高层数据结构上工作。它不要求用户指定对数据的存放方法,也不需要用户了解具体的数据存放方式,所以具有完全不同底层结构的不同数据库系统,可以使用相同的结构化查询语言作为数据输入与管理的接口。SQL可以嵌套,这使它具有极大的灵活性和强大的功能。

[0055] 3、数据湖

[0056] 数据湖,是一个存储企业的各种各样原始数据的大型仓库。数据湖中的数据可供存取、处理、分析及传输。

[0057] 数据湖中的数据是从企业的多个数据源中获取的原始数据,并且针对不同的目的,同一份原始数据还可能有多多种满足特定内部模型格式的数据副本。因此,数据湖中被处理的数据可能是任意类型的信息,从结构化数据到完全非结构化数据。

[0058] 4、数据仓库(data warehouse,DW或DWH)

[0059] 数据仓库,是企业所有级别的决策制定过程,提供所有类型数据支持的战略集合。它是单个数据存储,出于分析性报告和决策支持目的而创建。为需要业务智能的企业,提供指导业务流程改进、监视时间、成本、质量以及控制。

[0060] 5、数据中台

[0061] 数据中台,是指通过数据技术,对海量数据进行采集、计算、存储、加工,同时统一标准和口径。数据中台把数据统一之后,会形成标准数据,再进行存储,形成大数据资产层,进而为客户提供高效服务。

[0062] 6、关系型数据库(relational database,RD)

[0063] 关系型数据库,是指采用了关系模型来组织数据的数据库,其以行和列的形式存储数据,以便于用户理解,关系型数据库这一系列的行和列被称为表,一组表组成了数据库。用户通过查询来检索数据库中的数据,而查询是一个用于限定数据库中某些区域的执行代码。关系模型可以简单理解为二维表格模型,而一个关系型数据库就是由二维表及其之间的关系组成的一个数据组织。例如,常见的关系型数据库有Oracle数据库或SQLServer数据库等。

[0064] 现有技术中,通过访问数据底层引擎,以及基于访问数据的频率,确定数据的冷热程度。例如,为了提升对数据湖(例如,智能数据湖运营平台DAYU)的管理效率,在对数据湖

治理的过程中,通常需要确定数据湖中的数据哪些属于冷数据,哪些数据属于热数据。该方法通过记录数据访问频率来判断数据湖中的冷热数据,但这种判断方法依赖数据引擎底层提供接口,存在系统开销大、判断不够准确的问题。

[0065] 本申请实施例提供了一种确定冷热数据的方法。该方法在保证冷热数据的确定结果具有较高的准确度的前提下,能够有效降低系统开销。

[0066] 下面,结合图1对本申请实施例提供的确定冷热数据的方法进行详细介绍。

[0067] 图1是本申请实施例提供的一种确定冷热数据的方法100的示意性流程图。

[0068] 如图1所示,方法100可以包括步骤110至步骤130。下面对步骤110至步骤130进行详细介绍。

[0069] 步骤110,获取M条数据开发逻辑执行语句,M为大于等于2的正整数。

[0070] 在本申请实施例中,对获取M条数据开发逻辑执行语句的数据平台不作具体限定。

[0071] 例如,数据平台可以是数据湖,即从数据湖中获取M条数据开发逻辑执行语句。例如,数据平台可以是数据仓库,即也可以从数据仓库中获取M条数据开发逻辑执行语句。例如,上数据平台可以是数据中台,即还可以从数据中台中获取M条数据开发逻辑执行语句。

[0072] 可选的,在一些实现方式中,上述M条数据开发逻辑执行语句包括SQL。也就是说,上述M条数据开发逻辑执行语句可以是SQL语句。

[0073] 可选的,在另一些实现方式中,上述M条数据开发逻辑执行语句也可以是其它类型的数据库语言。例如,在另一个示例中,上述M条数据开发逻辑执行语句还可以是Oracle语句。

[0074] 例如,通常情况下数据湖中存储的语句为SQL语句,当从该数据湖中获取获取M条数据开发逻辑执行语句时,即是从该数据湖中获取获取M条SQL语句。

[0075] 在本申请实施例中,对上述M条数据开发逻辑执行语句的类型不作具体限定。

[0076] 例如,当上述M条数据开发逻辑执行语句是SQL语句时,上述M条SQL语句的类型可以包括但不限于以下类型:查询类型(select)、插入类型(insert)、更新类型(update)或删除类型(delete)等。

[0077] 例如,在一个示例中,当上述M条SQL语句中的一条SQL语句为select类型时,可以表示为:select cno,cname from courses where credit=3,该语句可以理解为找出3个学分的课程号和课程名。例如,在一个示例中,当上述M条SQL语句中的一条SQL语句为insert类型时,可以表示为:insert into infos values('s100102','林冲','男',22,2,to_date('2020-5-906:30:10','yyyy-mm-dd hh24:mi:ss'))。

[0078] 可选的。在一些实现方式中,上述获取的M条数据开发逻辑执行语句不相同。其中,M条数据开发逻辑执行语句不相同,可以理解为,M条数据开发逻辑执行语句中每条数据开发逻辑执行语句对应的数据表的字段不完全相同。

[0079] 例如,当上述数据开发逻辑执行语句为SQL语句时,从数据湖中获取了3条SQL语句,为了便于描述,分别记为SQL#1语句、SQL#2语句和SQL#3语句。如果对SQL#1语句解析得到SQL#1语句对应的数据表的字段为:标识id字段和姓名name字段,对SQL#2语句解析得到SQL#2语句对应的数据表的字段为:id字段、年龄age字段和name字段,在此情况下,可以认为SQL#1语句和SQL#2语句不相同。如果对SQL#1语句解析得到SQL#1语句对应的数据表的字段为id字段,对SQL#2语句解析得到SQL#2语句对应的数据表的字段为name字段,在此情况

下,则可以认为SQL#1语句和SQL#3语句不相同。可以理解的是,如果对SQL#1语句解析得到SQL#1语句对应的数据表的字段为id字段,对SQL#2语句解析得到SQL#2语句对应的数据表的字段为id字段,在此情况下,可以认为SQL#1语句和SQL#2语句是相同的。

[0080] 可选的,在步骤110之前还可以包括如下步骤:

[0081] 获取K条数据开发逻辑执行语句,其中,该K条数据开发逻辑执行语句不完全相同,K为大于等于1的正整数。

[0082] 从该K条数据开发逻辑执行语句中,选取不相同的M条数据开发逻辑执行语句。

[0083] 步骤120,对M条数据开发逻辑执行语句进行解析,得到N个关联权重值,N个关联权重值与N个数据表一一对应,N个数据表是K个数据表中不相同的数据表,K个数据表是对M条数据开发逻辑执行语句进行解析后得到的数据表,N个关联权重值中的第i个关联权重值用于指示K个数据表中与N个数据表中的第i个数据表存在关联关系的数据表的数目,N和K为大于等于1的正整数,且K小于等于N, $i=1,2,3,\dots,N$;

[0084] 上述N个关联权重值与N个数据表一一对应,可以理解的是,N个关联权重值中的第i个关联权重值与N个数据表中的第i个数据表对应。也就是说,第i个关联权重值是对应的第i个数据表的关联权重值。

[0085] 上述K个数据表中与第i个数据表存在关联关系的数据表,可以理解为,K个数据表中存在与第i个数据表包括一个或多个相同字段的数据表。

[0086] 例如,在一个示例中,当上述M条数据开发逻辑执行语句是SQL语句时,对M条SQL语句进行解析后得到了3个数据表,分别记为数据表1、数据表2和数据表3,且数据表1仅包括id字段,数据表2包括id字段和age字段,数据表3仅包括name字段。在此情况下,可以认为与数据表1存在关联关系的数据表是数据表2,与数据表2存在关联关系的数据表是数据表1,没有与数据表3存在关联关系的数据表。

[0087] 在本申请实施例中,对M条数据开发逻辑执行语句进行解析,N个关联权重值,可以包括如下步骤:

[0088] 根据M条数据开发逻辑执行语句的参数信息对M条数据开发逻辑执行语句进行解析,得到每条数据开发逻辑执行语句对应的数据表;

[0089] 统计每条数据开发逻辑执行语句对应的数据表中的与第i个数据表存在关联关系的数据表的数目;

[0090] 根据与第i个数据表存在关联关系的数据表的数目,确定第i个关联权重值。

[0091] 其中,上述M条数据开发逻辑执行语句的参数信息至少可以包括以下信息:M条数据开发逻辑执行语句的操作类型信息以及数据开发逻辑执行语句的操作参数信息。应理解的是,本申请实施例中对数据开发逻辑执行语句的参数信息不作具体限定。例如,数据开发逻辑执行语句的参数信息还可以包括除去操作类型信息以及操作参数信息之外的其它数据开发逻辑执行语句的参数信息。

[0092] 可选的,在另一些实现方式中,对M条数据开发逻辑执行语句进行解析,N个关联权重值,可以包括如下步骤:

[0093] 对M条数据开发逻辑执行语句中的第1条数据开发逻辑执行语句进行解析,得到第1条数据开发逻辑执行语句对应的数据表;

[0094] 对M条数据开发逻辑执行语句中的第p条数据开发逻辑执行语句进行解析,得到第

p条数据开发逻辑执行语句对应的数据表, $p=2,3,4,\dots,M$;

[0095] 根据第1条数据开发逻辑执行语句至第p条数据开发逻辑执行语句对应的数据表, 确定第p条数据开发逻辑执行语句对应的数据表中的每个数据表的关联权重值, 以及对第1条数据开发逻辑执行语句至第p-1条数据开发逻辑执行语句对应的数据表中的每个数据表的关联权重值进行更新。

[0096] 在本申请实施例中, 对数据开发逻辑执行语句的操作类型信息和数据开发逻辑执行语句的操作参数信息不作具体限定。

[0097] 在一个示例中, 当上述M条数据开发逻辑执行语句是SQL语句时, SQL的操作类型信息可以包括但不限于以下类型: 查询类型、插入类型、更新类型或删除类型。例如, SQL的操作参数信息可以包括但不限于以下类型: SQL引用数据表的字段、SQL引用的数据表以及SQL引用的数据库, SQL语句类型等。

[0098] 在本申请实施例中, 根据与第i个数据表存在关联关系的数据表的数目, 确定N个关联权重值, 可以包括如下步骤:

[0099] 如果查询到K个数据表中与第i个数据表存在关联关系的数据表的数目为P, 则确定第i个关联权重值等于P, P为大于等于零的整数。

[0100] 例如, 查询到K个数据表中与第i个数据表存在关联关系的数据表的数目为3, 则确定第i个关联权重值等于3。例如, 查询到K个数据表中与第i个数据表存在关联关系的数据表的数目为1, 则确定第i个关联权重值等于1。

[0101] 步骤130, 根据N个关联权重值, 确定N个数据表中的冷热数据。

[0102] 在本申请实施例中, 根据N个关联权重值, 确定N个数据表中的冷热数据, 可以包括如下步骤:

[0103] 如果第i个关联权重值大于等于预设关联权重值, 则确定第i个关联权重值对应的数据表包括的数据为热数据, $i=1,2,3,\dots,N$;

[0104] 如果第i个关联权重值小于预设关联权重值, 则确定第i个关联权重值对应的数据表包括的数据为冷数据。

[0105] 其中, 预设关联权重值可以根据具体的应用场景进行设置, 本申请实施例中对此不作具体限定。例如, 当上述M条数据开发逻辑执行语句是SQL语句时, 对M条SQL语句进行解析后得到了20张数据表, 且这20张数据表中有10张不相同的数据表, 在此情况下, 可以设置预设关联权重值等于3, 4或5等等。

[0106] 应理解, 上述图1仅为示意并不对本申请实施例构成任何限定。例如, 上述SQL语句还可以替换为数据库中存储数据表的其它语句(例如, Oracle语句)。

[0107] 本申请实施例提供的确定冷热数据的方法, 避免了现有技术中需要访问数据底层引擎的操作。该方法通过对从数据平台中获取的数据开发逻辑执行语句直接进行解析, 根据解析结果确定数据平台中各数据表之间的关联关系, 根据关联关系确定对应的数据表的关联权重值, 并根据关联权重值确定数据表的冷热属性。该方法在保证冷热数据的确定结果具有较高的准确度的前提下, 能够有效降低系统开销。

[0108] 下面, 结合图2以数据开发逻辑执行语句是SQL语句为例, 介绍本申请提供的确定冷热数据的一个具体的实施例。

[0109] 图2是本申请实施例提供的另一种确定冷热数据的方法200的示意性流程图。

[0110] 如图2所示,方法200可以包括步骤210至步骤240。下面对210至步骤240进行详细介绍。应理解,图2仅为示意并不对本申请构成任何限定。例如,方法200中的SQL语句还可以替换为存储数据表的其它类型的语句。

[0111] 步骤210,SQL采集器从数据湖中获取M条不相同的SQL语句,M为大于等于2的正整数。

[0112] 其中,M条不相同的SQL语句,可以理解为,M条SQL语句中每条SQL语句对应的数据表的字段不完全相同。

[0113] 步骤220,使用关系型数据库对M条SQL语句进行存储。

[0114] 在一个示例中,上述关系型数据库可以是SQL Server数据库。

[0115] 步骤230,使用SQL解析器对关系型数据库中存储的M条SQL语句进行解析,得到K个数据表,K个数据表中包括N个不相同的数据表,N和K为大于等于1的正整数,且K小于等于N。

[0116] 其中,具体的解析方法与上述方法100中的步骤120中的方法相同,此处不再详细赘述。

[0117] 步骤240,使用图形数据库对得到的K个数据表进行处理得到第一记录表,第一记录表是用于记录上述N个不相同的数据表中的各个数据表之间的关联关系以及对应的N个关联权重值的网状表。

[0118] 其中,确定N个不相同的数据表中的各个数据表之间的关联关系以及对应的关联权重值与上述方法100中的步骤120中的方法相同,此处不再详细赘述。

[0119] 在上述技术方案中,最终在图形数据库中记录了一个网状结构的表关系(即,上述第一记录表),基于建立好的该网状结构表,可以判断数据湖中各个数据表被关联的程度,根据该关联程度与预设阈值可以进一步确定数据湖中的每个数据表的冷热属性(即,冷数据或热数据)。

[0120] 在一个示例中,如果第一记录表中记录的第i个关联权重值大于等于预设关联权重值,则确定第i个关联权重值对应的数据表包括的数据为热数据, $i=1,2,3,\dots,N$;

[0121] 如果第i个关联权重值小于预设关联权重值,则确定第i个关联权重值对应的数据表包括的数据为冷数据。

[0122] 其中,预设关联权重值可以根据具体的应用场景进行设置,本申请实施例中对此不作具体限定。例如,对M条SQL语句进行解析后得到了20张数据表,且这20张数据表中有10张不相同的数据表,在此情况下,可以设置预设关联权重值等于3,4或5等等。

[0123] 可选的,在上述步骤210至上述步骤240之后还可以包括如下操作:

[0124] 使用热度服务器对外公布上述确定的N个不相同的数据表中每个数据表的冷热属性;

[0125] 使用解析服务器根据上述冷热属性,得到N个不相同的数据表的冷热情况。

[0126] 上文结合图1和图2,详细描述了本申请提供的确定冷热数据的方法。下面,结合图3至图5详细介绍本申请提供的确定冷热数据的装置、设备和系统。应理解,方法实施例的描述与装置实施例的描述相互对应,因此,未详细描述的部分可以参见前面方法实施例。

[0127] 图3是本申请提供的一种确定冷热数据的装置3000的示意性结构图。

[0128] 如图3所示,该装置3000包括:获取单元3001和处理单元3002。

[0129] 获取单元3001,用于获取M条数据开发逻辑执行语句,M为大于等于2的正整数;

[0130] 处理单元3002,用于对该M条数据开发逻辑执行语句进行解析,得到N个关联权重值,该N个关联权重值与N个数据表一一对应,该N个数据表是K个数据表中不相同的数据表,该K个数据表是对该M条数据开发逻辑执行语句进行解析后得到的数据表,该N个关联权重值中的第i个关联权重值用于指示该K个数据表中与该N个数据表中的第i个数据表存在关联关系的数据表的数目,N和K为大于等于1的正整数,且K小于等于N, $i=1,2,3,\cdots,N$;

[0131] 该处理单元3002,还用于根据该N个关联权重值,确定该N个数据表中的冷热数据。

[0132] 可选的,在一些实现方式中,该处理单元3002,还用于执行如下操作:

[0133] 根据该M条数据开发逻辑执行语句的参数信息对该M条数据开发逻辑执行语句进行解析,得到每条数据开发逻辑执行语句对应的数据表;

[0134] 统计该每条数据开发逻辑执行语句对应的数据表中的与该第i个数据表存在关联关系的数据表的数目;

[0135] 根据与该第i个数据表存在关联关系的数据表的数目,确定该第i个关联权重值。

[0136] 可选的,在一些实现方式中,该M条数据开发逻辑执行语句的参数信息至少包括以下信息:该M条数据开发逻辑执行语句的操作类型信息以及该数据开发逻辑执行语句的操作参数信息。

[0137] 可选的,在一些实现方式中,该处理单元3002,还用于执行如下操作:

[0138] 如果查询到该K个数据表中与该第i个数据表存在关联关系的数据表的数目为P,则确定该第i个关联权重值等于P,P为大于等于零的整数。

[0139] 可选的,在一些实现方式中,该处理单元3002,还用于执行如下操作:

[0140] 如果该第i个关联权重值大于等于预设关联权重值,则确定该第i个关联权重值对应的数据表包括的数据为热数据, $i=1,2,3,\cdots,N$;

[0141] 如果该第i个关联权重值小于预设关联权重值,则确定该第i个关联权重值对应的数据表包括的数据为冷数据。

[0142] 可选的,在一些实现方式中,该M条数据开发逻辑执行语句不相同。

[0143] 可选的,在一些实现方式中,该M条数据开发逻辑执行语句包括结构化查询语句SQL。

[0144] 下面,结合图4,以确定冷热数据的设备中包括收发器、处理器和存储器为例进行介绍。

[0145] 图4是本申请提供的一种确定冷热数据的设备4000的示意性结构图。如图4所示,该设备4000包括:收发器4010、处理器4020和存储器4030。其中,收发器4010、处理器4020和存储器4030之间通过内部连接通路互相通信,传递控制和/或数据信号,该存储器4030用于存储计算机程序,该处理器4010用于从该存储器4030中调用并运行该计算机程序,以控制该收发器4020收发信号。

[0146] 具体的,收发器4010的功能与图3所示的获取单元3001的具体功能相对应,此处不再赘述。

[0147] 具体的,处理器4020的功能与图3所示的处理单元3002的具体功能相对应,此处不再赘述。

[0148] 在本申请实施例中,确定冷热数据的设备4000中应包括处理器4020。其中,可选的,在一些实现方式中,该确定冷热数据的设备4000中还可以包括收发器4010。

[0149] 可选的,在一些实现方式中,该确定冷热数据的设备4000中还可以包括存储器4030。

[0150] 图5是本申请实施例提供的一种系统5000的结构示意图。如图5所示,该系统5000包括:上文中确定冷热数据的装置3000和/或确定冷热数据的设备4000。

[0151] 本申请实施例提供了一种计算机程序产品,当该计算机程序产品在确定冷热数据的装置3000上运行时,使得确定冷热数据的装置3000可以执行上述方法实施例中的方法100和/或方法200。

[0152] 本领域普通技术人员可以意识到,结合本文中所公开的实施例中描述的各方法步骤和单元,能够以电子硬件、计算机软件或者二者的结合来实现,为了清楚地说明硬件和软件的可互换性,在上述说明中已经按照功能一般性地描述了各实施例的步骤及组成。这些功能究竟以硬件还是软件方式来执行,取决于技术方案的特定应用和设计约束条件。本领域普通技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能,但是这种实现不应认为超出本申请的范围。

[0153] 所属领域的技术人员可以清楚地了解到,为了描述的方便和简洁,上述描述的系统、装置和单元的具体工作过程,可以参见前述方法实施例中的对应过程,在此不再赘述。

[0154] 在本申请所提供的几个实施例中,所揭露的系统、装置和方法,可以通过其它的方式实现。例如,以上所描述的装置实施例仅仅是示意性的,例如,该单元的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式,例如多个单元或组件可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另外,所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口、装置或单元的间接耦合或通信连接,也可以是电的,机械的或其它的形式连接。

[0155] 该作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本申请实施例方案的目的。

[0156] 另外,在本申请各个实施例中的各功能单元可以集成在一个处理单元中,也可以是各个单元单独物理存在,也可以是两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现,也可以采用软件功能单元的形式实现。

[0157] 该集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用,可以存储在一个计算机可读存储介质中。基于这样的理解,本申请的技术方案本质上或者说对现有技术做出贡献的部分,或者该技术方案的全部或部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行本申请各个实施例中方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(read-only memory,ROM)、随机存取存储器(random access memory,RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0158] 以上描述,仅为本申请的具体实施方式,但本申请的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本申请揭露的技术范围内,可轻易想到各种等效的修改或替换,这些修改或替换都应涵盖在本申请的保护范围之内。因此,本申请的保护范围应以权利

要求的保护范围为准。

[0159] 在上述实施例中,可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件实现时,可以全部或部分地以计算机程序产品的形式实现。该计算机程序产品包括一个或多个计算机程序指令。在计算机上加载和执行该计算机程序指令时,全部或部分地产生按照本申请实施例中的流程或功能。该计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。该计算机指令可以存储在计算机可读存储介质中,或者从一个计算机可读存储介质向另一个计算机可读存储介质传输,例如,该计算机程序指令可以从一个网站站点、计算机、服务器或数据中心通过有线或无线方式向另一个网站站点、计算机、服务器或数据中心进行传输。该计算机可读存储介质可以是计算机能够存取的任何可用介质或者是包含一个或多个可用介质集成的服务器、数据中心等数据存储设备。该可用介质可以是磁性介质(例如软盘、硬盘、磁带)、光介质(例如,数字视频光盘(digital video disc,DVD)、或者半导体介质(例如固态硬盘)等。

[0160] 本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成,也可以通过程序来指令相关的硬件完成,该程序可以存储于一种计算机可读存储介质中,上述提到的存储介质可以是只读存储器,磁盘或光盘等。

[0161] 所属领域的技术人员可以清楚地了解到,为描述的方便和简洁,上述描述的系统、装置和单元的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0162] 另外,本申请中术语“和/或”,仅仅是一种描述关联对象的关联关系,表示可以存在三种关系,例如,A和/或B,可以表示:单独存在A,同时存在A和B,单独存在B这三种情况。另外,本文中字符“/”,一般表示前后关联对象是一种“或”的关系;本申请中术语“至少一个”,可以表示“一个”和“两个或两个以上”,例如,A、B和C中至少一个,可以表示:单独存在A,单独存在B,单独存在C、同时存在A和B,同时存在A和C,同时存在C和B,同时存在A和B和C,这七种情况。

[0163] 以上所述,仅为本申请的具体实施方式,但本申请的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本申请揭露的技术范围内,可轻易想到变化或替换,都应涵盖在本申请的保护范围之内。因此,本申请的保护范围应以所述权利要求的保护范围为准。

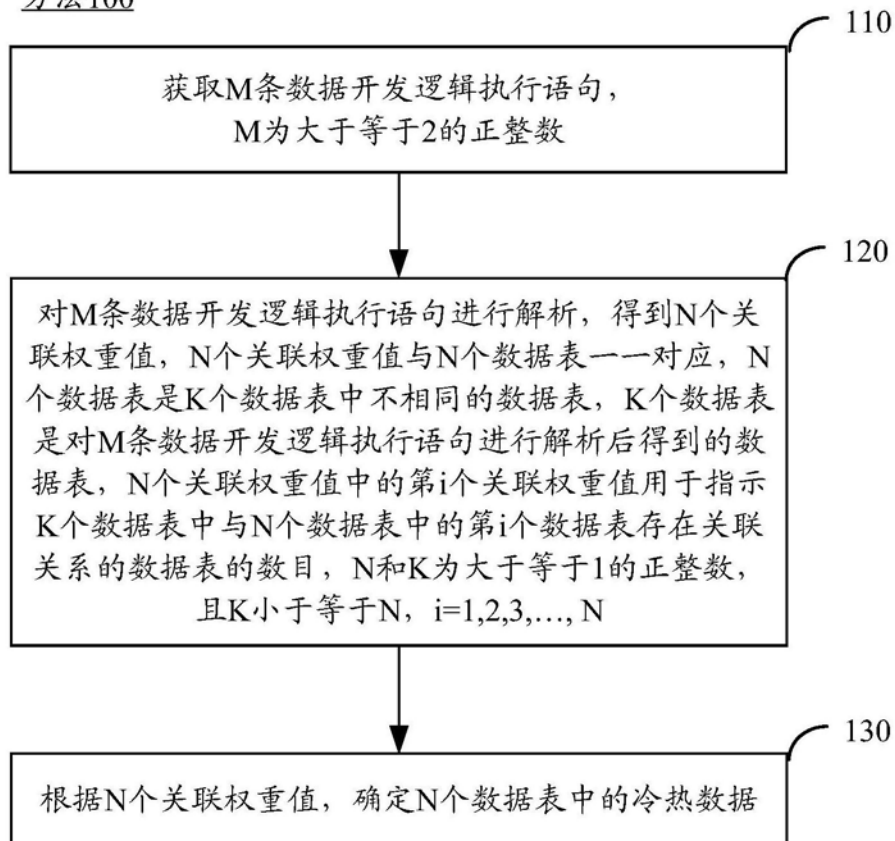
方法100

图1

方法200

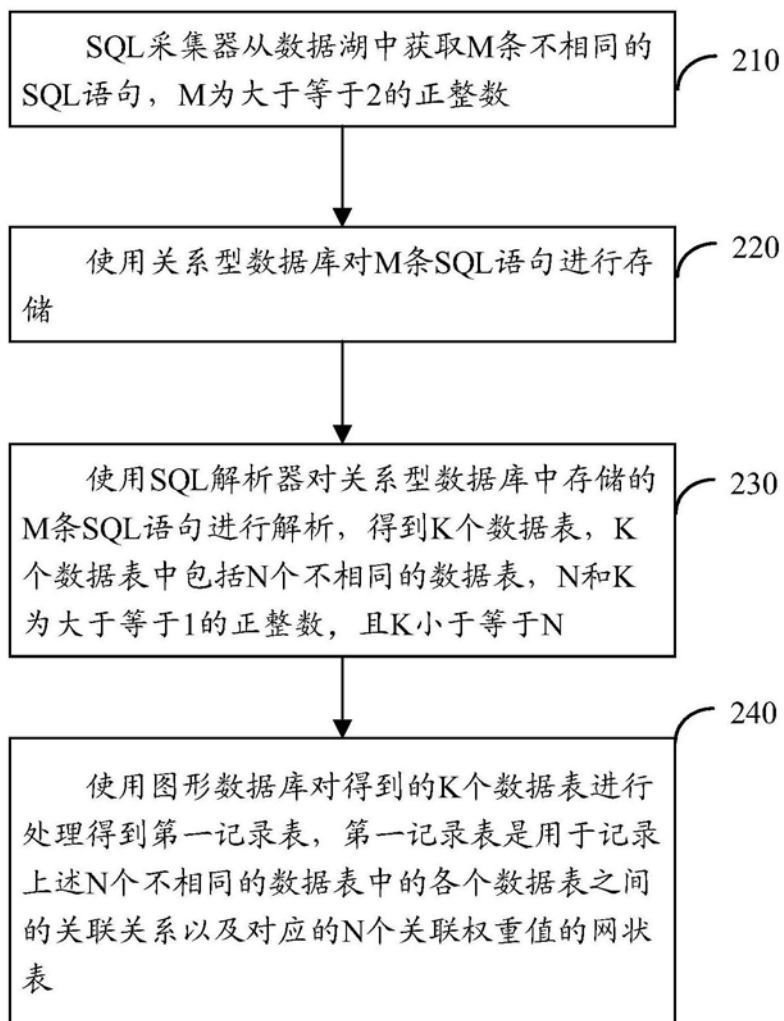


图2

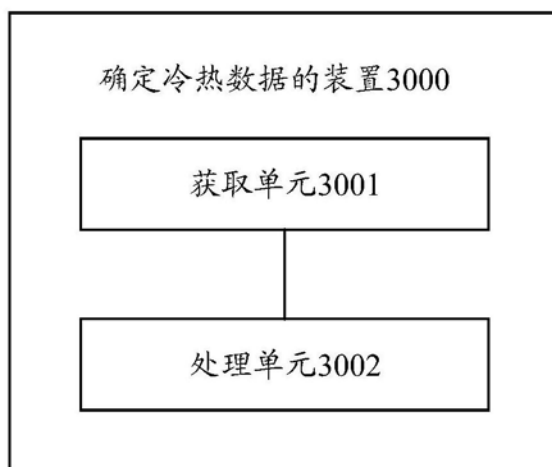


图3

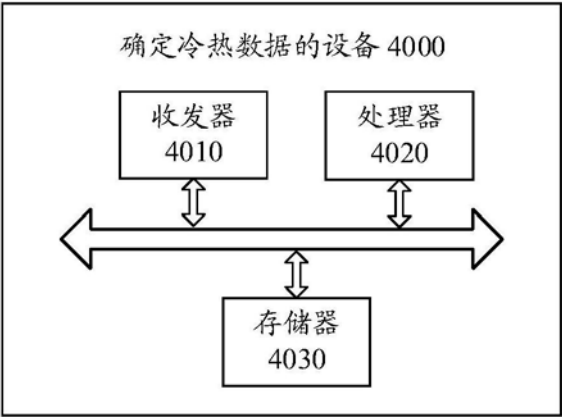


图4



图5