(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 106104525 A
(43)申请公布日 2016.11.09

(21)申请号 201480077256.0
(22)申请日 2014.03.31
(85)PCT国际申请进入国家阶段日
2016.09.18
(86)PCT国际申请的申请数据
PCT/EP2014/056425 2014.03.31
(87)PCT国际申请的公布数据
WO2015/149830 EN 2015.10.08
(71)申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼
(72)发明人 埃利泽·利维 阿哈龙·埃威佐 卢卡斯·布劳恩 托马斯·埃特 乔治斯·加斯帕里斯 马丁·考夫曼 唐纳德·科斯曼 丹尼尔·威德默
(51)Int.Cl.
G06F 17/30(2006.01)

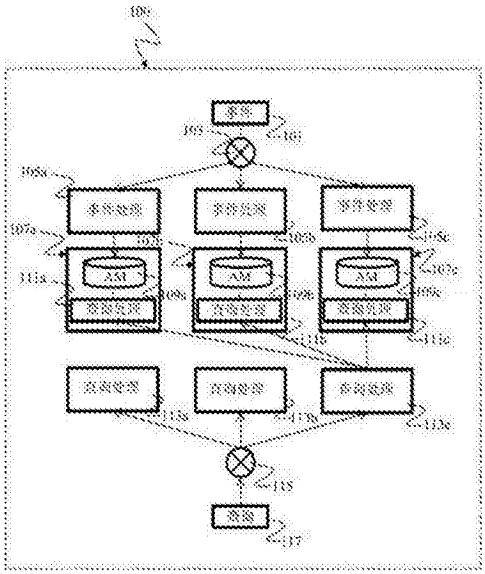
权利要求书2页 说明书10页 附图2页

(54)发明名称

事件处理系统

(57)摘要

一种事件处理系统(100)用于处理在数据库系统上操作的流事件(101),所述事件处理系统(100)包括事件负载均衡单元(103)、多个事件计算节点(105a、105b、105c)以及与所述事件计算节点(105a、105b、105c)分离的多个事件状态存储(109a、109b、109c),其中所述事件负载均衡单元(103)用于根据事件负载均衡标准将流事件(101)路由至所述多个事件计算节点(105a、105b、105c);所述多个事件状态存储(109a、109b、109c)用于存储所述多个事件计算节点(105a、105b、105c)的状态以保存所述事件处理的状态;以及所述多个事件计算节点(105a、105b、105c)用于处理从所述事件负载均衡单元(103)接收的所述流事件(101),根据所述事件处理改变它们的状态,以及基于它们改变后的状态更新所述多个事件状态存储(109a、109b、109c)。



CN 106104525 A

1.一种事件处理系统(100),用于处理在数据库系统上操作的流事件(101),其特征在于,所述事件处理系统(100)包括事件负载均衡单元(103)、多个事件计算节点(105a、105b、105c)以及与所述事件计算节点(105a、105b、105c)分离的多个事件状态存储(109a、109b、109c),其中:

所述事件负载均衡单元(103)用于根据事件负载均衡标准将所述流事件(101)路由至所述多个事件计算节点(105a、105b、105c);

所述多个事件状态存储(109a、109b、109c)用于存储所述多个事件计算节点(105a、105b、105c)的状态以保存所述事件处理的状态;以及

所述多个事件计算节点(105a、105b、105c)用于处理从所述事件负载均衡单元(103)接收的所述流事件(101),根据所述事件处理改变它们的状态,以及基于它们改变后的状态更新所述多个事件状态存储(109a、109b、109c)。

2.根据任一前述权利要求所述的事件处理系统(100),其特征在于,还包括查询负载均衡单元(115)和多个查询处理节点(113a、113b、113c),其中:

所述查询负载均衡单元(115)用于根据查询负载均衡标准将多个查询(117)路由至所述多个查询处理节点(113a、113b、113c);

所述多个查询处理节点(113a、113b、113c)用于处理从所述查询负载均衡单元(115)接收的所述查询(117)。

3.根据权利要求2所述的事件处理系统(100),其特征在于,所述查询负载均衡单元(115)用于将每个查询(117)准确转发给一个查询处理节点(113a、113b、113c)。

4.根据权利要求2或3所述的事件处理系统(100),其特征在于,所述多个查询处理节点(113a、113b、113c)中的一个查询处理节点用于访问所述多个事件状态存储(109a、109b、109c)中的至少一个事件状态存储以处理从所述查询负载均衡单元(115)接收的所述查询(117)。

5.根据权利要求4所述的事件处理系统(100),其特征在于,所述查询处理节点(113a、113b、113c)用于访问更多数据,尤其是所述数据库系统中的用于处理所述查询(117)的客户主数据。

6.根据任一前述权利要求所述的事件处理系统(100),其特征在于,所述多个查询处理节点(113a、113b、113c)用于处理随即查询以进行实时分析。

7.根据任一前述权利要求所述的事件处理系统(100),其特征在于,所述多个事件状态存储(109a、109b、109c)包括基于分布式主存储器的键值存储。

8.根据任一前述权利要求所述的事件处理系统(100),其特征在于,所述多个事件计算节点(105a、105b、105c)用于处理规则和连续查询以在同一时间实时地一起进行复杂事件处理和所述事件处理。

9.根据任一前述权利要求所述的事件处理系统(100),其特征在于,所述事件负载均衡单元(103)用于基于应用定义分区,尤其是基于客户键,以及规则复制来路由所述流事件(101)。

10.根据权利要求9所述的事件处理系统(100),其特征在于,所述事件负载均衡单元(103)用于路由所述流事件(101),使得所述多个事件计算节点(105a、105b、105c)中的一个事件计算节点处理所述流事件(101)的一个特定子集以及处理所述流事件(101)的所述特

定子集上的所有规则。

11.根据权利要求1至8中任一权利要求所述的事件处理系统(100),其特征在于,所述事件负载均衡单元(103)用于基于事件复制和规则分区来路由所述流事件(101),使得每个事件计算节点(105a、105b、105c)处理所有流事件(101)以及规则的子集。

12.根据权利要求1至8之一所述的事件处理系统(100),其特征在于,所述事件负载均衡单元103用于基于事件101的循环分区和规则复制来路由流事件(101),使得多个事件计算节点(105a、105b、105c)中的一个事件计算节点在具有空闲容量时处理流事件(101)。

13.根据权利要求1至8中任一权利要求所述的事件处理系统(100),其特征在于,所述事件负载均衡单元(103)用于通过多个事件计算节点(105a、105b、105c)基于所述流事件(101)和规则分区和复制来路由所述流事件(101)。

14.一种事件处理方法(200),其特征在于,所述方法包括:

根据事件负载均衡标准将多个流事件(101)路由(201)至多个事件计算节点(105a、105b、105c);

将所述多个事件计算节点(105a、105b、105c)的状态存储(202)在与所述事件计算节点(105a、105b、105c)分离的多个事件状态存储(109a、109b、109c)中以保存所述事件处理的状态;以及

处理(203)所述多个事件计算节点(105a、105b、105c)中的已接收流事件(101),根据所述流事件(101)的处理改变所述多个事件计算节点(105a、105b、105c)的状态,以及更新所述多个事件状态存储(109a、109b、109c)中的所述多个事件计算节点(105a、105b、105c)的状态。

15.一种计算机程序产品,包括存储在其上的程序代码的可读存储介质,其特征在于,所述程序代码包括执行以下操作的指令:

根据事件负载均衡标准将多个流事件(101)路由至多个事件计算节点(105a、105b、105c);

将所述多个事件计算节点(105a、105b、105c)的状态存储在与所述事件计算节点(105a、105b、105c)分离的多个事件状态存储(109a、109b、109c)中;以及

处理所述多个事件计算节点(105a、105b、105c)中的已接收流事件(101),根据所述事件的处理改变所述多个事件计算节点(105a、105b、105c)的状态,以及更新所述多个事件状态存储(109a、109b、109c)中的所述多个事件计算节点(105a、105b、105c)的状态。

事件处理系统

技术领域

[0001] 本发明涉及一种事件处理系统和一种事件处理方法。本发明还涉及结合数据库管理系统(database management system,DBMS)(尤其是使用SQL(Structured Query Language,结构化查询语言)的DBMS)实时分析的可扩展流处理。

背景技术

[0002] 在大规模数据处理中,三个重要特征:高容量有状态事件处理、对复杂声明式查询(例如,SQL查询)的实时分析以及工作负载的可扩展性,对许多应用至关重要。工作负载即查询和事件的数目和数据库大小。例如,电信运营商可能需要每秒捕获和处理多达数百万呼叫记录以及提供支持来分析其客户的行为与复杂随即查询。最新的数据库技术最多能够应对这三个特征中的两个。具体而言,存在两类系统:(a)数据流管理系统(Data Stream Management System,DSMS)和(b)数据库管理系统(Database Management System,DBMS)。DSMS适合于复杂事件处理,但是不支持实时分析。最高级的DBMS(在最好的情况下)支持实时分析,但并不很适合于复杂事件处理。两种系统通常都具有有限的可扩展性。

[0003] 数据流管理系统(Data Stream Management System,DSMS)的目标是实时跟踪离散事件的高带宽流。最新的系统使用StreamSQL等SQL查询语言的扩展来提供声明式界面。通过使用StreamSQL,应用开发人员定义规则(还称为连续查询)来规定输入事件流上的可能复杂模式。每当输入事件流公开这种模式时,这些规则触发并产生一个结果,应用将此结果作为一个事件流。

[0004] 为了执行DSMS的功能,DSMS必须保存数目可能非常大的需要用来评估这些规则的状态变量。原则上,当每个事件发生时,DSMS必须更新相关状态变量并检查新的全局状态是否要求触发预先指定的规则。相关计算意味着计算新的值和更新这些状态变量;然而,相关存储意味着保存状态变量,这样可以实时进行新值的频繁计算。

[0005] 现代DSMS能够在标准的单机上每秒处理多达几十万个事件以及数百条规则。它们通过应用多个优化进行。具体而言,它们在内部优化专门针对给定规则集合的状态变量的管理。为了随机器的数目扩展,这些系统复制事件流并在不同机器上处理不同规则。也就是说,它们复制数据并对工作负载进行分区。这些系统的设计以在下面项目环境中进行的研究为基础,包括:依据“Brian Babcock、Shivnath Babu、Mayur Datar、Rajeev Motwani、Jennifer Widom:数据流系统中的模型和问题。PODS,2002年:1至16页”的流处理项目,依据“Sailesh Krishnamurthy、Sirish Chandrasekaran、Owen Cooper、Amol Deshpande、Michael J.Franklin、Joseph M.Hellerstein、Wei Hong、Samuel Madden、Frederick Reiss、Mehul A.Shah:TelegraphCQ:架构状态报告。IEEE Data Eng.Bull.26(1):11至18页(2003年)”的TelegraphCQ项目以及依据“Hari Balakrishnan、Magdalena Balazinska、Donald Carney、Ugur Çetintemel、Mitch Cherniack、Christian Convey、Eduardo F.Galvez、Jon Salz、Michael Stonebraker、Nesime Tatbul、Richard Tibbetts、Stanley B.Zdonik:Retrospective on Aurora。VLDB J.13(4):370至383页(2004年)”和“Daniel

J.Abadi、Yanif Ahmad、Magdalena Balazinska、UgurÇetintemel、Mitch Cherniack、Jeong-Hyon Hwang、Wolfgang Lindner、Anurag Maskey、Alex Rasin、Esther Ryvkina、Nesime Tatbul、Ying Xing、Stanley B.Zdonik:北极流处理引擎的设计。CIDR,2005年:277至289页"的Aurora/Borealis项目。

[0006] 现代DSMS的设计能够使其适合于特定应用,这些应用被设计用于现代DSMS,但是使这些应用不适合支持实时分析。通过封装所有状态变量和调谐专门用于规则的状态变量的存储,现代DSMS几乎变得不可能在数据上处理任何随即查询。因此,现代DSMS没有足够的实时分析能力。"数据复制—工作负载分区"原则限制了这些系统的可扩展性。如果数据速率增长超过单机的能力,则这些系统无法再维持维持工作负载。

[0007] 为了克服现代DSMS的可扩展性问题,依据"Daniel Peng、Frank Dabek:使用分布式事务和通知的大规模增量处理。OSDI 2010年:251至264页"的Percolator系统被开发为"谷歌"公司的专有系统。Percolator系统提供了一种用于分布式和可扩展流和复杂事件处理的灵活编程框架。通过Percolator,应用开发人员可选择遵循"数据分区—工作负载分区"模型来实现更好的数据可扩展性。然而,Percolator并不提供任何声明式机制来支持规则和连续查询。此外,Percolator并不对实时分析提供任何支持。遵循大多数Percolator设计原则,依据"www.storm-project.net"的风暴(Storm)系统被"Twitter"公司开发为开源系统。

[0008] 作为DSMS的替代系统,理论上有可能使用现代DBMS处理事件流并进行随即实时分析查询。在该方法中,每个事件被发布为DBMS的一个事务。随即查询随后可由相同的DBMS执行。现代主内存数据库系统,诸如依据"Per-ÅkeLarson、Mike Zwilling、Kevin Farlee:Hekaton内存优化的OLTP引擎。IEEE Data Eng.Bull.36(2):34至40页(2013年)"的"Hekaton"在单机上每秒能够处理成千上万个简单事务(诸如插入一个事件)。此外,它们可通过对若干节点上的数据进行分区来扩展。然而,它们对随即查询的支持是有限的,因为,在另一方面,这类高端事务处理系统的内部数据结构特别适于处理数据导入的许多小事务。此外,这些系统通常不支持与存储在若干机器上的数据有关的查询。更糟的是,这些系统完全不支持处理规则和连续查询。此外,这一缺陷是其架构固有的,也是起初开发DSMS的关键动力,如"Michael Stonebraker:技术观点—一体适用:时代到来和过去的观念。通信ACM 51(12):76页(2008年)"所描述。这些系统最多支持所谓的在某个事件发生时触发的触发条件(参见"Jennifer Widom、Stefano Ceri:主动数据库系统的介绍。主动数据库系统:高级数据库处理的触发条件和规则,1996年:1至41页"),但是这些触发条件无法检测通常在复杂事件处理应用的规则中发现的复杂模式。

发明内容

[0009] 本发明的目的是提供一种改进的事件处理技术,从而提供高容量有状态事件处理、复杂声明式查询的实时分析以及工作负载和数据大小的可扩展性。

[0010] 该目的由独立权利要求的特征来实现。其它实施方式从从属权利要求、描述内容和附图中显而易见。

[0011] 如下文所描述的本发明是基于以下基本观察:提供改进的事件处理系统有两个关键点,第一点是事件处理的逻辑和状态的分离,第二点是构构建分析矩阵(analytics

matrix,AM)及出于实时分析目的而在单独状态上增加查询处理能力的可能性。

[0012] 由于流处理系统的状态和逻辑可能不耦合,所以有可能以系统化横向扩展方式扩展系统。横向扩展相对简单并效率更高,因为AM的更新可能按照键进行且AM可按照键进行分区。另外,无状态逻辑和规则索引可重复扩展。最新的DSMS无法展示最终的可扩展性。同样地,存储层可分别按照数据大小和/或工作负载要求进行扩展。

[0013] 如果流处理子系统的状态被保存在主存储器数据库系统(相反于分布式键值存储)中,则该状态可用于在新数据上进行查询处理,从而满足实时分析要求。在任何情况下,实时分析层都可通过与扩展流处理层相同的方式随工作负载而扩展;可在任何时间点添加和移除节点以应对不断增加或不断减少的工作负载。

[0014] 在下文中,描述一种称为分析矩阵(Analytics Matrix,AM)的机制,用于将事件处理的状态保存在特殊专用存储中。具体而言,状态变量可存储在基于分布式主内存的键值存储中。此外,处理节点的单独层被描述为处理复杂事件处理的规则和连续查询。处理节点的另一单独层被描述为处理实时分析的随即查询。

[0015] 为了详细地描述本发明,将使用以下术语、缩略语和表示:

[0016] AM: 分析矩阵。

[0017] DBMS: 数据库管理系统。

[0018] DSMS: 数据流管理系统。

[0019] SQL: 结构化查询语言。

[0020] SEP: 流和事件处理系统。

[0021] RTA: 实时分析系统。

[0022] AIM: 运动分析

[0023] ETL: 提取、转换、加载。

[0024] CRM: 客户关系管理。

[0025] 数据库管理系统(database management System,DBMS)是经特殊设计的应用,这些应用与用户、其它应用和数据库本身交互以获取和分析数据。通用数据库管理系统(database management system,DBMS)是一种设计为支持数据库定义、创建、查询、更新和管理的软件系统。不同的DBMS可通过使用诸如SQL和ODBC或JDBC等标准互操作以允许单个应用与一个以上数据库一起工作。

[0026] 数据流管理系统(Data stream management system,DSMS)是一种管理连续数据流的计算机程序。它类似于数据库管理系统(database management system,DBMS),然而,DSMS设计用于传统数据库中的静态数据。DSMS还提供灵活的查询处理,使得信息需求可以通过查询来表达。然而,相比于DBMS,DSMS执行连续查询,连续查询不仅执行一次,而且永久保存。因此,查询被连续执行,直到其显式地被移除。由于多数DSMS是数据驱动的,所以只要新数据到达系统,连续查询就会产生新结果。该基本概念类似于复杂事件处理,使得两种技术部分合并。

[0027] 事件处理、事件计算。

[0028] 事件处理是一种跟踪和分析(处理)信息(数据)流以及从中推导出结论的方法,信息是关于发生的事情(事件)。复杂事件处理,或CEP,是组合来自多个源的数据以推断表明更复杂情形的事件或模式的事件处理。复杂事件处理的目标是识别有意义的事件(诸如机

遇或威胁)并尽可能快地响应这些事件。

[0029] SQL(Structured Query Language,结构化查询语言)是一种专用编程语言,设计用于管理保存在关系数据库管理系统(relational database management system,RDBMS)中的数据。

[0030] 最初以关系代数和元组关系演算为基础,SQL由数据定义语言和数据操作语言组成。SQL的范围包括数据插入、查询、更新和删除、模型创建和修改以及数据访问控制。

[0031] InfiniBand是一种在高性能计算和企业数据中心中使用的计算机交换网络通信链路。其特征包括高吞吐量、低延迟、服务质量和故障接管,并且InfiniBand被设计为可扩展的。InfiniBand架构规范定义了处理器节点和诸如存储设备等高性能I/O节点之间的连接。

[0032] 根据第一方面,本发明涉及一种事件处理系统,用于处理在数据库系统上操作的多个事件,所述事件处理系统包括事件负载均衡单元、多个事件计算节点以及与所述事件计算节点分离的多个事件状态存储,其中:所述事件负载均衡单元用于根据事件负载均衡标准将所述多个事件路由至所述多个事件计算节点;所述多个事件状态存储用于存储所述多个事件计算节点的状态以保存事件处理的状态;以及所述多个事件计算节点用于:处理从所述事件负载均衡单元接收的事件,根据所述事件的所述处理改变它们的状态以及基于它们改变后的状态更新所述多个事件状态存储。

[0033] 事件状态存储与事件计算节点的分离提供了可扩展的流处理。有状态流处理,即,意识到其内部状态的流处理,可通过可扩展方式来实现。

[0034] 扩展有状态流处理使得事件处理系统能够承受极高的事件发生率。在单个系统中将通过SQL等进行的实时分析的能力与有状态流处理相结合能够降低总成本和系统复杂性。

[0035] 事件计算节点不与各个状态节点关联。状态和计算之间的分离/解耦允许任何计算节点处理任何传入事件。在传入事件上,计算节点能够透明地访问任何状态存储以获取当前状态。由此,事件处理系统是可扩展的且均衡的,因为任何计算节点都可处理任何传入事件并可访问其中一个状态存储上的任何状态。

[0036] 事件处理系统因此提供足够快的单独状态存储以在相对于吞吐量和响应事件的最佳性能下处理大量事件流。

[0037] 通过引入这种分离,有助于实时查询域中的各种客户使用任一种复杂分析查询来查询这种合并状态存储。

[0038] 根据所述第一方面,在所述事件处理系统的第一可能实施形式中,所述事件处理系统还包括查询负载均衡单元和多个查询处理节点,其中:所述查询负载均衡单元用于根据查询负载均衡标准将多个查询路由至所述多个查询处理节点;所述多个查询处理节点用于处理从所述查询负载均衡单元接收的所述查询。

[0039] 事件状态存储与事件计算节点的分离提供了结合实时分析的可扩展流处理。

[0040] 根据所述第一方面的所述第一实施形式,在所述事件处理系统的第二可能实施形式中,所述查询负载均衡单元用于将每个查询准确转发给一个查询处理节点。

[0041] 当将每个查询准确转发给一个查询处理节点时,只有该处理节点是忙碌的,使得系统效率得以提高。

[0042] 根据所述第一方面的所述第一或所述第二实施形式，在所述事件处理系统的第三可能实施形式中，所述多个查询处理节点中的一个查询处理节点用于访问所述多个事件状态存储中的至少一个事件状态存储以处理从所述查询负载均衡单元接收的所述查询。

[0043] 所述查询处理节点在访问至少一个事件状态存储时可以增加其处理查询的响应时间。

[0044] 根据所述第一方面的所述第三实施形式，在所述事件处理系统的第四可能实施形式中，所述查询处理节点用于访问更多数据，尤其是数据库系统处理查询的客户主数据。

[0045] 当访问诸如数据库系统的客户主数据等更多数据时，处理查询的准确性得以提高。

[0046] 根据如上所述第一方面或根据所述第一方面的任一前述实施形式，在所述事件处理系统的第五可能实施形式中，所述多个查询处理节点用于处理随即查询以进行实时分析。

[0047] 当处理随即查询时，所述事件处理系统能够执行实时分析。

[0048] 根据如上所述第一方面或根据所述第一方面的任一前述实施形式，在所述事件处理系统的第六可能实施形式中，所述多个事件状态存储包括基于分布式主内存的键值存储。

[0049] 分布式主内存允许更快的访问时间，从而加速事件处理系统。

[0050] 根据如上所述第一方面或根据所述第一方面的任一前述实施形式，在所述事件处理系统的第七可能实施形式中，所述多个事件计算节点用于处理规则和连续查询以进行复杂事件处理以及在同一时间实时进行事件处理。

[0051] 因此，所述事件处理系统能够在同一系统中同时进行实时查询分析和可扩展事件处理。

[0052] 根据如上所述第一方面或根据所述第一方面的任一前述实施形式，在所述事件处理系统的第八可能实施形式中，所述事件复杂均衡单元用于基于应用定义分区，尤其基于客户键和规则复制，来路由所述事件。

[0053] 基于应用定义分区来路由所述事件允许快速路由事件，因此系统的效率更高。

[0054] 根据如上所述第一方面或根据所述第一方面的所述第八实施形式，在所述事件处理系统的第九可能实施形式中，所述事件负载均衡单元用于路由所述事件，使得所述多个事件计算节点中的一个事件处理节点处理所述事件的特定子集并且处理所述事件的所述特定子集上的所有规则。

[0055] 当事件计算节点处理事件的特定子集时，各个计算节点的复杂性得以降低，因此降低了事件处理系统的整体复杂性。

[0056] 根据如上所述第一方面或根据所述第一方面的所述第一至所述第七实施形式的任一实施形式，在所述事件处理系统的第十可能实施形式中，所述事件负载均衡单元用于基于事件复制和规则分区路由所述事件，使得每个事件计算节点处理所有事件以及规则的子集。

[0057] 当每个事件计算节点处理所有事件以及规则的子集时，每个事件计算节点的软件可相同，从而降低开发复杂性。

[0058] 根据如上所述第一方面或根据所述第一方面的所述第一至所述第七实施形式任

一实施形式，在所述事件处理系统的第十一可能实施形式中，所述事件负载均衡单元用于基于所述事件的循环分区以及规则复制路由所述事件，使得所述多个事件计算节点中的一个事件计算节点在其具有可用容量时处理事件。

[0059] 当基于循环分区来路由所述事件时，所有事件可在大约相同的时延内处理。

[0060] 根据如上所述第一方面和根据所述第一方面的所述第一至所述第七实施形式的任一实施形式，在所述事件处理系统的第十二可能实施形式中，所述事件负载均衡单元用于基于事件和规则分区和复制通过所述多个事件计算节点路由所述事件。

[0061] 基于所述事件和规则分区和复制来路由所述事件允许通过简单的规则描述所述系统。

[0062] 根据第二方面，本发明涉及一种事件处理方法，所述方法包括：根据事件负载均衡标准将多个事件路由至多个事件计算节点；将所述多个事件计算节点的状态存储在与所述事件计算节点分离的多个事件状态存储中以保存所述事件处理的状态；以及处理所述多个事件计算节点中的所述接收的事件，根据所述事件的所述处理改变所述多个事件计算节点的状态以及更新所述多个事件状态存储中的所述多个事件计算节点的所述状态。

[0063] 通过分离事件的状态存储与事件的计算节点，所述方法每秒能够处理更高速率的流事件，即，吞吐量增加。

[0064] 根据第三方面，本发明涉及一种计算机程序产品，包括在其上存储程序代码的可读存储介质以供计算机使用，所述程序代码包括执行以下操作的指令：根据事件负载均衡标准将多个事件路由至多个事件计算节点；将所述多个事件计算节点的状态存储在与所述事件计算节点分离的多个事件状态存储中；以及处理所述多个事件计算节点中的所述接收事件，根据所述事件的所述处理改变所述多个事件计算节点的状态以及更新所述多个事件状态存储中的所述多个事件计算节点的所述状态。

[0065] 事件状态存储与事件计算节点的分离提供了可扩展的流处理。扩展有状态流处理使得事件处理系统能够承受极高的事件发生率。在单个系统中将实时分析的能力与有状态流处理相结合能够降低总成本和系统复杂性。计算机程序可灵活地设计，从而很容易地实现需求更新。所述计算机程序产品可在如下文所描述的事件处理系统上运行。

[0066] 因此，本发明的各方面提供一种改进的事件处理技术，从而提供高容量有状态事件处理、复杂声明式查询的实时分析以及工作负载和数据大小的可扩展性。

附图说明

[0067] 本发明的具体实施方式将结合以下附图进行描述，其中：

[0068] 图1所示为根据一实施形式的图示一种事件处理系统100的方框图；以及

[0069] 图2所示为根据一实施形式的图示一种事件处理方法200的方框图。

具体实施方式

[0070] 以下结合附图进行详细描述，所述附图是描述的一部分，并通过图解说明的方式示出可以实施本发明的具体方面。可以理解的是，在不脱离本发明范围的情况下，可以利用其他方面，并可以做出结构上或逻辑上的改变。因此，以下详细的描述并不当作限定，本发明的范围由所附权利要求书界定。

[0071] 本文中描述的设备和方法可基于事件处理节点和事件状态存储。据了解，结合所描述的方法进行的评论同样对于一种执行所述方法的对应设备或系统成立，反之亦然。例如，如果描述了特定方法步骤，对应的设备可包括执行所描述的方法步骤的单元，即使此单元没有在图中详细描述或图示。此外，应理解，本文中所描述的各种示例性方面的特征可相互组合，除非另有特殊说明。

[0072] 本文中所描述的方法和设备可在数据库管理系统中，尤其是使用SQL的DBMS中实施。所描述的设备和系统可以包括集成电路和/或无源器件且可以根据各种技术制造。例如，电路可设计为逻辑集成电路、模拟集成电路、混合信号集成电路、光电路、存储器电路和/或集成无源器件。

[0073] 图1所示为根据一实施形式的图示事件一种处理系统100的方框图。事件处理系统100可处理在数据库系统上操作的多个事件101。事件处理系统100可包括事件负载均衡单元103、多个事件计算节点105a、105b、105c以及与事件计算节点105a、105b、105c分离的多个事件状态存储109a、109b、109c。事件负载均衡单元103可根据事件负载均衡标准将多个事件101路由至多个事件计算节点105a、105b、105c。多个事件状态存储109a、109b、109c可存储多个事件计算节点105a、105b、105c的状态以保存事件处理的状态。多个事件计算节点105a、105b、105c可处理从事件负载均衡单元103接收的事件101，根据事件的处理改变它们的状态以及基于它们改变后的状态更新多个事件状态存储109a、109b、109c。

[0074] 事件处理系统100还可包括查询负载均衡单元115和多个查询处理节点113a、113b、113c。查询负载均衡单元115可根据查询负载均衡标准将多个查询117路由至多个查询处理节点113a、113b、113c。多个查询处理节点113a、113b、113c可处理从查询负载均衡单元115接收的查询117。

[0075] 在一个示例中，查询负载均衡单元115可将每个查询117准确转发给一个查询处理节点113a、113b、113c。在一个示例中，查询处理节点113a、113b、113c可访问多个事件状态存储109a、109b、109c中的至少一个事件状态存储以处理从查询负载均衡单元115接收的查询117。在一个示例中，查询处理节点113a、113b、113c可访问更多数据，尤其是数据库系统中处理查询117的客户主数据。在一个示例中，多个查询处理节点113a、113b、113c可处理随即查询以进行实时分析。在一个示例中，多个事件状态存储109a、109b、109c可包括基于分布式主存储器的键值存储。在一个示例中，多个事件计算节点105a、105b、105c可在同一时间实时地处理规则和连续查询以一起进行复杂事件处理和事件处理。在一个示例中，事件负载均衡单元103可基于应用定义分区，尤其基于客户键和规则复制，来路由事件101。在一个示例中，事件负载均衡单元103可路由事件101，使得多个事件计算节点105a、105b、105c中的一个事件计算节点处理事件101的特定子集并且处理事件101的特定子集上的所有规则。在一个示例中，事件负载均衡单元103可基于事件复制和规则分区来路由事件101，使得每个事件计算节点105a、105b、105c处理所有事件101以及规则的子集。在一个示例中，事件负载均衡单元103可基于事件101的循环分区以及规则复制来路由事件101，使得多个事件计算节点105a、105b、105c中的一个事件计算节点在具有剩余容量时处理事件101。在一个示例中，事件负载均衡单元103可基于事件101和规则的分区和复制通过多个事件计算节点105a、105b、105c路由事件101。

[0076] 图1描绘了整个架构。包含事件计算节点105a、105b、105c的顶层可专用于复杂事

件处理。负载均衡器，下文中表示为事件负载均衡单元103，可将多个事件101作为输入并且将事件101路由至处理节点，下文中表示为事件计算节点105a、105b、105c。该层中的处理节点105a、105b、105c可处理用于复杂事件处理的规则并且可计算状态变量的新状态。它们可更新存储在中间层单元107a、107b、107c中的AM(analytics matrix，分析矩阵)。底层单元，下文中表示为查询处理节点113a、113b、113c，可专用于处理随即查询以进行实时分析。此外，存在一负载均衡器，下文中表示为查询负载均衡单元115，其可将多个随即查询117作为输入并可将随即查询117准确转发给该层中的一个处理节点113a、113b、113c。该处理节点113a、113b、113c随后可处理该查询117，从而访问AM 109a、109b、109c，并可能访问其它维度数据，例如客户主数据。在获得高性能的实施形式中，特殊且分布式的查询处理技术可用于在AM 109a、109b、109c上执行随即查询。

[0077] 可实施流处理中的负载均衡的四种不同变体：在第一变体中，可采用(按照，例如客户键)事件的应用定义分区以及规则复制。也就是说，流处理层中的处理节点可处理所有事件的特定子集并可对处理该事件集上的所有规则。在第二变体中，可采用复制事件以及规则分区。也就是说，流处理层中的每个处理节点均可处理所有事件但只处理规则的子集。在第三变体中，可采用事件的循环分区以及复制规则。也就是说，流处理层中的处理节点每当具有剩余容量时才可处理事件。在第四变体中，可采用混合方案，即，在处理层中对事件和规则进行分区和复制。

[0078] 哪种变体最好取决于应用工作负载的特征。在第一和第二变体以及第四变体的某些例示中，处理节点可缓存状态变量以获得更高性能。在第三变体中，处理节点可从每个事件的存储层读取所需状态变量。

[0079] 实验已经表明，在现代硬件上，例如InfiniBand网络，即使第三变体也会表现得非常好。AM的确切存储布局可取决于选定的用于流处理中的负载均衡的选择。在第一变体中，状态变量可按照客户键进行分区并可存储在对应的节点上，而在第二和第三变体中，状态变量可存储在相同节点上作为可使用它们的规则。

[0080] 为了实现分布式状态变量读取和更新的高性能，可使用主存储器哈希技术；例如，在“Ion Stoica、Robert Morris、David Liben-Nowell、David R.Karger、M.Frans Kaashoek、Frank Dabek、Hari Balakrishnan:Chord:互联网应用的可扩展端到端查找协议。IEEE/ACM会刊网络，11(1):17至32页(2003年)”以及“Antony I.T.Rowstron、Peter Druschel:Pastry:可扩展、分散化对象位置和大规模端到端系统的路由。中间件，2001年:329至350页”中描述的Chortle和Pastry系统提出的那些协议等。或者，可使用任何其它主存储器数据库系统(例如，微软的Hekaton、Oracle的Times Ten，或SAP的Hana产品)。此外，可使用现代低延迟网络技术，诸如InfiniBand。

[0081] 新查询处理技术可应用于在这种分布式主存储器存储系统上处理随即、实时分析查询。为了在流处理层的处理节点中有效地处理事件，可应用特殊的规则编译技术。这些技术可考虑每个状态变量的特殊语义。例如，一些状态变量可完全增量处理；其它的可能需要追踪一些变更历史。编译器可利用每种状态变量，从而最小化空间和时间。

[0082] 事件处理系统100可组成实时决策子系统决策系统，其可为电信运营商的CRM系统的一部分。子系统可支撑同时来自计费系统和CRM系统用户提交的若干不同查询的混合工作负载。子系统可划分为两个部分。第一，流和事件处理系统(Stream and Event

Processing System,SEP),以适于快速评估业务规则的方式处理和存储事件。第二,实时分析系统(real-time analytics System,RTA),估计更复杂的分析查询。

[0083] 本发明中提出的新颖方法,还称为“运动分析(Analytics in Motion,AIM)”,不遵循传统数据仓库技术,但是会使RTA直接访问SEP的存储,从而使其能够实时地答复分析查询。在传统数据仓库技术中,RTA是通过连续ETL(Extract,Transform,Load,提取、变换、加载)操作由SEP导入的,

[0084] SEP处理的规则可支持决定根据特定条件运算向客户推广哪类产品信息。这种查询的条件可能具有高选择性,聚合可能很频繁。接着,描述了RTA查询。例如,子系统可支持查询以将客户划分为组,从而帮助营销人员设计和测量活动。全表扫描、聚合和多表连接可能很频繁。

[0085] 图2所示为根据一实施形式的图示一种事件处理方法200的方框图。方法200可包括根据事件负载均衡标准将多个事件101路由201至多个事件计算节点105a、105b、105c(参见图1)。方法200可包括将多个事件计算节点105a、105b、105c的状态存储202在与事件计算节点105a、105b、105c分离的多个事件状态存储109a、109b、109c中以保存事件处理的状态。方法200可包括处理203多个事件计算节点105a、105b、105c中的已接收事件101,根据事件101的处理改变多个事件计算节点105a、105b、105c的状态,以及更新多个事件状态存储109a、109b、109c中的多个事件计算节点105a、105b、105c的状态。

[0086] 本文中所描述的方法、系统和设备可以作为数字信号处理器(Digital Signal Processor,DSP)、微控制器或任何其它边处理器中的软件或作为专用集成电路(application specific integrated circuit,ASIC)内的硬件电路来实现。

[0087] 本发明可以在数字电子电路,或计算机硬件、固件、软件,或其组合中实现,例如,在传统移动设备的可用硬件或专用于处理本文中所描述方法的新硬件中实现。

[0088] 本发明还支持一种包括计算机可执行代码或计算机可执行指令的计算机程序产品,当执行这些指令时,使得至少一个计算机执行本文中所描述的执行和计算步骤,尤其是上文结合图2所描述的方法200和上文结合图1所描述的技术。这种计算机程序产品可包括存储在其上的以供计算机使用的程序代码的可读存储介质,该程序代码可包括执行以下操作的指令:根据事件负载均衡标准将多个事件路由至多个事件计算节点;将多个事件计算节点的状态存储在于事件计算节点分离的多个事件状态存储中;以及处理多个事件计算节点中的已接收事件,根据事件的处理改变多个事件计算节点的状态,以及更新多个事件状态存储中的多个事件计算节点的状态。

[0089] 尽管本发明的特定特征或方面可能已经仅结合几种实现方式中的一种进行公开,但此类特征或方面可以和其它实现方式中的一个或多个特征或方面相结合,只要对于任何给定或特定的应用是有需要或有利。而且,在一定程度上,术语“包括”、“有”、“具有”或这些词的其它变形在详细的说明书或权利要求书中使用,这类术语和所述术语“包含”是类似的,都是表示包括的含义。同样,术语“示例性地”,“例如”仅表示为示例,而不是最好或最佳的。

[0090] 尽管本文中已经图示和描述了具体方面,但是本领域普通技术人员将会理解各种替代和/或等效实施形式可以代替所示出和描述的具体方面,而不脱离本发明的范围。该申请旨在覆盖本文论述的具体实施方式的任何修改或变更。

[0091] 尽管以下权利要求书中的各元素是借助对应的标签按照特定顺序列举的，除非对权利要求的阐述另有暗示用于实现部分或所有这些元素的特定顺序，否则这些元素并不一定限于以所述特定顺序来实现。

[0092] 通过以上启示，对于本领域技术人员来说，许多替代产品、修改及变体是显而易见的。当然，所属领域的技术人员容易意识到除本文所述的应用之外，还存在本发明的众多其它应用。虽然已参考一个或多个特定实施例描述了本发明，但所属领域的技术人员将认识到在不偏离本发明的范围的前提下，仍可对本发明做出许多改变。因此，应理解，只要是在所附权利要求书及其等效文句的范围内，可以用不同于本文具体描述的方式来实践本发明。

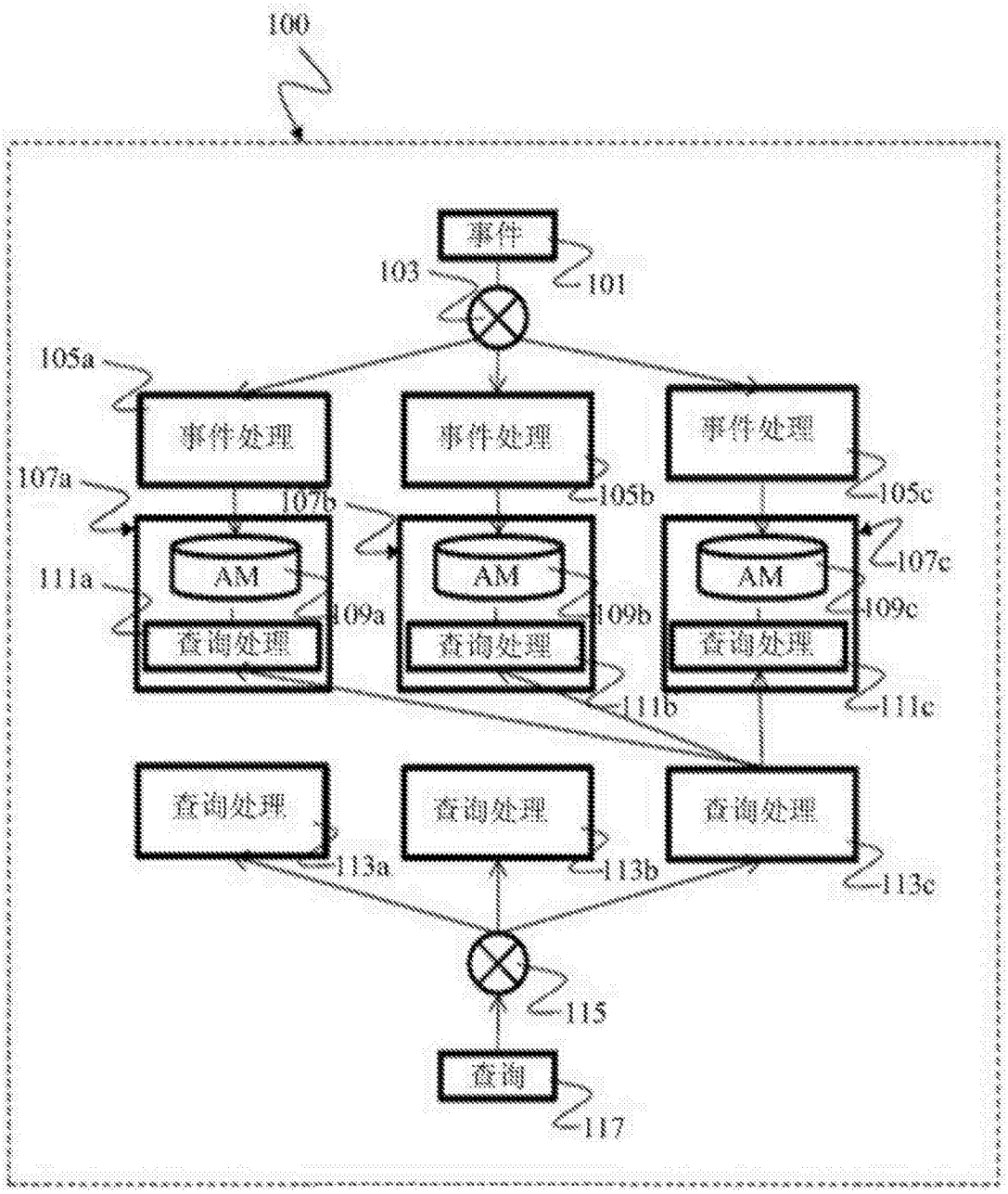

100
事件
101
103
105a
事件处理
事件处理
事件处理
105b
105c
107a
107b
107c
AM
AM
AM
111a
109a
109b
109c
查询处理
查询处理
查询处理
111b
111c
查询处理
查询处理
查询处理
113a
113b
113c
115
查询
117

图1

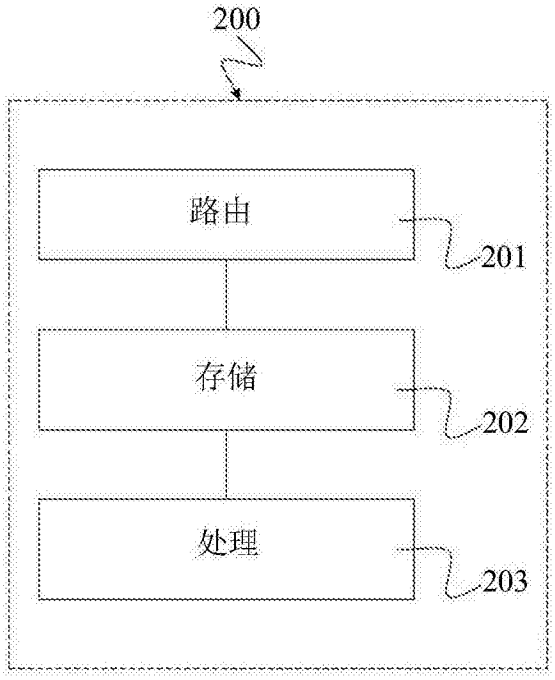


图2