(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 105706047 A
(43)申请公布日 2016.06.22

(21)申请号 201480061587.5
(22)申请日 2014.11.11
(30)优先权数据
14/077,167 2013.11.11 US
(85)PCT国际申请进入国家阶段日
2016.05.11
(86)PCT国际申请的申请数据
PCT/US2014/065057 2014.11.11
(87)PCT国际申请的公布数据
WO2015/070236 EN 2015.05.14
(71)申请人 亚马逊科技公司
地址 美国内华达
(72)发明人 M·M·泰默 G·D·高雷
J·D·杜纳根 G·伯吉斯 熊颖
(74)专利代理机构 中国国际贸易促进委员会专利商标事务所 11038
代理人 郑宗玉
(51)Int.Cl.
G06F 7/00(2006.01)

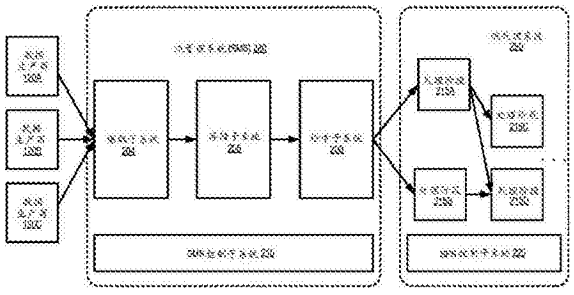
权利要求书3页 说明书47页 附图31页

(54)发明名称

基于分区的数据流处理框架

(57)摘要

一种多租户流处理服务的控制节点接收请求，所述请求指示有待在特定的数据流的数据记录上执行的操作。基于流分区策略，所述控制节点确定有待使用的工作节点的初始数量。所述控制节点配置工作节点以在接收的数据上执行所述操作。响应于所述工作节点正处于不健康的状态中的确定，所述控制节点配置替换工作节点。



CN 105706047 A

1.一种方法,其包括:

由一个或多个计算装置执行以下各项:

在多租户流处理服务处从特定的客户端接收有待在指定的处理阶段处在特定的数据流的数据记录上执行特定操作的指示,和用于所述特定操作的结果的特定的输出分配描述符;

至少部分基于所述特定的操作来确定有待配置用于所述指定的处理阶段的工作节点的初始数量;

配置所述初始数量的工作节点的特定的工作节点来:(a)在所述特定的数据流的一个或多个分区的接收的数据记录上执行所述特定的操作,(b)存储进度记录,所述进度记录指示已在所述工作节点处处理过的所述一个或多个分区的部分,以及(c)根据所述特定的输出分配描述符将所述特定的操作的结果传递至一个或多个目的地;以及

响应于所述特定的工作节点正处于不健康的状态中的确定,选择替换工作节点以替换所述特定的工作节点,其中所述替换工作节点访问由所述特定的工作节点存储的进度记录,以识别所述一个或多个分区的至少一个数据记录,在所述一个或多个分区上有待由所述替换工作节点执行所述特定的操作。

2.如权利要求1所述的方法,其还包括由所述一个多个计算装置执行:

调用由多租户流管理服务所实现的一个或多个编程数据记录检索接口,以接收所述一个或多个分区的数据记录,包括特定的编程数据记录检索接口,其包括请求的数据记录的分区内的序列号的指示作为参数。

3.如权利要求1所述的方法,其还包括由所述一个多个计算装置执行:

实现一个或多个编程接口以使得所述流处理服务的客户端能够指定用于一个或多个数据流的数据记录的处理阶段的有向无环图。

4.如权利要求1所述的方法,其还包括由所述一个多个计算装置执行:

从负责所述特定的数据流的所述数据记录的存储的多租户流管理服务获取分区策略正在用于所述特定的数据流的指示;以及

至少部分基于所述分区策略来确定所述工作节点的初始数量。

5.如权利要求1所述的方法,其中所述特定的输出分配描述符指示有待根据不同的分区策略将所述特定操作的结果作为不同的数据流的数据记录分配至或更多个被配置用于所述不同的数据流的摄取节点。

6.如权利要求1所述的方法,其还包括由所述一个多个计算装置执行:

响应于在所述处理阶段的不同工作节点处的工作量水平符合触发标准的确定,实现以下各项中的一个或多个:(a)所述特定的数据流的动态重新分区,(b)替代工作节点到在所述不同的工作节点处先前处理的至少一个分区的分配,(c)被配置用于所述处理阶段的多个工作节点的改变,或者(d)工作节点从一个服务器到另一个服务器的转移。

7.如权利要求1所述的方法,其中将所述特定的工作节点配置来将条目存储在持久性数据资源库中,所述条目代表积累的应用程序状态信息,所述状态信息对应于已在所述特定的工作节点处处理过的多个数据记录,并且将所述特定的工作节点配置成包括进度记录中的所述条目的指示。

8.如权利要求1所述的方法,其还包括由所述一个多个计算装置执行:

响应于通过客户端库部件的调用的流处理配置请求，在所述多租户流处理服务处注册指定的资源作为用于不同的处理阶段的工作节点。

9.如权利要求1所述的方法，其还包括由所述一个多个计算装置执行：

在所述多租户流处理服务处从所述特定的客户端接收有待在不同的处理阶段处在不同的数据流的数据记录上执行特定的非幂等操作的指示；以及

配置所述不同处理阶段的第一工作节点，以在接收的数据记录上执行所述非幂等操作。

10.如权利要求9所述的方法，其还包括由所述一个多个计算装置执行：

配置所述不同处理阶段的所述第一工作节点以：(a)执行清除操作以将所述非幂等操作的结果存储至一个或多个目的地，和(b)将清除操作定时的指示保存在持久性存储地点处；以及

使用所述清除操作定时的所述指示来配置替换工作节点，以在所述第一工作节点故障之后的恢复期间重演所述清除操作。

11.一种包括一个或多个处理器和一个或多个存储器的系统，所述一个或多个存储器包括程序指令，当在所述一个或多个处理器上执行所述程序指令时实现多租户流处理服务的控制节点，其中所述控制节点可操作来：

通过编程接口从特定的客户端接收有待在特定的数据流的数据记录上执行的特定操作的指示；

至少部分基于与所述特定的数据流相关联的分区策略来确定在处理阶段处用于所述指定的数据流的工作节点的初始数量；

配置所述初始数量的工作节点的特定的工作节点，以在所述特定的数据流的一个或多个分区的接收的数据记录上执行所述特定操作；以及

响应于所述特定的工作节点正处于不健康的状态中的确定，配置替换工作节点以替换所述特定的工作节点。

12.如权利要求11所述的系统，其中所述控制节点可操作来：

配置冗余组，所述冗余组包括处理不同数据流的不同分区的数据记录的多个工作节点，其中将所述多个工作节点中的至少一个工作节点指定为接收所述不同分区的所述数据记录的主要节点，并且其中将所述多个工作节点中的至少另一个工作节点配置为响应于触发事件来承担主要节点的责任的备用节点。

13.如权利要求11所述的系统，其中所述控制节点可操作来：

响应于在所述处理阶段的不同工作节点处的工作量水平符合触发标准的确定，实现以下各项中的一个或多个：(a)所述特定的数据流的动态重新分区，(b)替代工作节点到在所述不同的工作节点处先前处理的至少一个分区的分配，(c)被配置用于所述处理阶段的多个工作节点的改变，或者(d)工作节点从一个服务器到另一个服务器的转移。

14.如权利要求11所述的系统，其中所述特定的输出分配描述符指示有待根据不同的分区策略将所述特定操作的结果作为短暂数据流的数据记录分配至或更多个被配置用于所述短暂数据流的摄取节点，对于所述短暂数据流来说到持久性存储装置的存储是不需要的。

15.如权利要求11所述的系统，其中所述控制节点可操作来：

实现一个或多个编程接口以使得所述流处理服务的客户端能够指定用于一个或多个数据流的数据记录的处理阶段的有向无环图。

基于分区的数据流处理框架

背景技术

[0001] 由于多年来数据存储的成本已降低并且由于将计算机基础结构的各种元件互连的能力已改进，能够潜在地采集和分析与很多种应用程序有关的越来越多的数据。例如，移动电话可产生指示它们的地点、正在由电话用户使用的应用程序等的数据，可采集和分析所述数据中的至少一些来为用户呈现定制优惠券、广告等。由监控摄像机采集的数据的分析对于防止和/或解决犯罪可能是有用的，并且从在飞机引擎、汽车或复杂机械内嵌入在各种地点处的传感器采集的数据可用于各种目的，诸如预防性维护、提高效率和降低成本。

[0002] 流数据的体积增加已伴随有商用硬件的增加的使用(并且在一些情况下可通过商用硬件的增加的使用而成为可能)。用于商用硬件的虚拟化技术的出现已为管理用于许多类型的应用程序的大规模计算资源提供了益处，从而允许各种计算资源高效且安全地由多个客户端共享。例如，虚拟化技术可以通过为每个用户提供由单一物理计算机托管的一个或多个虚拟机而允许所述单一物理计算机在多个用户之间共享，其中每个这样的虚拟机充当不同逻辑计算系统的软件模拟，所述软件模拟为用户提供了以为自己是给定硬件计算资源的唯一操作者和管理员的错觉，同时还提供了各种虚拟机之间的应用程序隔离和安全性。此外，一些虚拟化技术能够提供跨越两个或更多个物理资源的虚拟资源，如具有跨越多个不同物理计算系统的多个虚拟处理器的单一虚拟机。除了计算平台，一些大型组织还提供使用虚拟化技术建立的各种类型的存储服务。通过使用这种存储服务，海量数据可存储具有所需的耐久性水平。

[0003] 虽然从各种供应商可以相对较低的成本获得虚拟计算和/或存储资源，然而较大的动态波动的数据流的采集、存储和处理的管理和编排出于各种原因仍然是挑战性命题。当更多资源被添加到系统设置用于处理较大的数据流时，例如可能出现系统的不同部分之间的工作量的不均衡。如果不加以解决，除了其他资源的未充分利用(和因此的损耗)以外，那么这种不均衡可能导致在一些资源处的严重性能问题。客户还可能对他们的流数据的安全性担心，或者如果这种数据或结果被存储在客户无法控制的设施处，那么还对分析流数据的结果担心。在当分布式系统尺寸上增大时增加频率的情况下自然倾向于发生的故障(诸如连接性的偶尔失去和/或硬件故障)也可能必须有效地解决，以防止流数据采集、存储或分析的昂贵的中断。

附图说明

[0004] 图1提供了根据至少一些实施方案的数据流概念的简化概述。

[0005] 图2提供了根据至少一些实施方案的在流管理系统(SMS)和包括流处理阶段的采集的流处理系统(SPS)的各种子部件之中的数据流的概述。

[0006] 图3示出根据至少一些实施方案的在SMS SPS处可实现的相应组编程接口的实例。

[0007] 图4示出根据至少一些实施方案的示例性基于网络的接口，所述接口可实现为使得SPS客户端能够产生流处理阶段的图形。

[0008] 图5示出根据至少一些实施方案的在SMS处可实现的编程记录提交接口和记录检

索接口的实例。

[0009] 图6示出根据至少一些实施方案的SMS的摄取子系统的示例性元件。

[0010] 图7示出根据至少一些实施方案的SMS的存储子系统的示例性元件。

[0011] 图8示出根据至少一些实施方案的SMS的检索子系统的示例性元件和检索子系统与SPS的交互的实例。

[0012] 图9示出根据至少一些实施方案的可建立用于SMS或SPS的节点的冗余组的实例。

[0013] 图10示出根据至少一些实施方案的供应商网络环境，其中给定冗余组的节点可分布在多个数据中心中。

[0014] 图11示出根据至少一些实施方案的可以被选择用于SMS或SPS的节点的多个放置目的地。

[0015] 图12a和图12b分别示出根据至少一些实施方案的可以由SPS客户端和SMS客户端提交的安全选项请求的实例。

[0016] 图13a示出根据至少一些实施方案的在流数据生产商与SMS的摄取节点之间的示例性交互。

[0017] 图13b示出根据至少一些实施方案的可以在SMS处被产生用于摄取的数据记录的序列号的示例性要素。

[0018] 图14示出根据至少一些实施方案的在SMS处的流数据记录的有序存储和检索的实例。

[0019] 图15示出根据至少一些实施方案的流分区映射和可以针对SMS和SPS节点做出的对应配置决策的实例。

[0020] 图16示出根据至少一些实施方案的动态流重新分区的实例。

[0021] 图17是根据至少一些实施方案的示出可执行来支持用于数据记录摄取和数据记录检索的相应组编程接口的操作方面的流程图。

[0022] 图18a是根据至少一些实施方案的示出可执行来配置流处理阶段的操作方面的流程图。

[0023] 图18b是根据至少一些实施方案的示出响应于用于流处理工作节点的配置的客户端库的部件调用可执行的操作方面的流程图。

[0024] 图19是根据至少一些实施方案的示出可执行来实现用于流处理的一种或多种恢复策略的操作方面的流程图。

[0025] 图20是根据至少一些实施方案的示出可执行来实现用于数据流的多种安全选项的操作方面的流程图。

[0026] 图21是根据至少一些实施方案的示出可执行来实现用于数据流的分区策略的操作方面的流程图。

[0027] 图22是根据至少一些实施方案的示出可执行来实现数据流的动态重新分区的操作方面的流程图。

[0028] 图23是根据至少一些实施方案的示出可执行来实现用于数据流记录的至少一次记录摄取策略的操作方面的流程图。

[0029] 图24是根据至少一些实施方案的示出可执行来实现用于数据流的多种持久性策略的操作方面的流程图。

[0030] 图25示出根据至少一些实施方案的流处理系统的实例，其中处理阶段的工作节点使用数据库表来协调它们的工作量。

[0031] 图26示出根据至少一些实施方案的可存储在用于工作量协调的分区分配表中的示例性条目。

[0032] 图27示出根据至少一些实施方案的可由流处理阶段的工作节点执行来选择在其上执行处理操作的分区的操作方面。

[0033] 图28示出根据至少一些实施方案的可由流处理阶段的工作节点执行来基于从流管理服务控制子系统获取的信息更新分区分配表的操作方面。

[0034] 图29示出根据至少一些实施方案的可由流处理阶段的工作节点执行的负载均衡操作的方面。

[0035] 图30是示出可以在至少一些实施方案中使用的示例性计算装置的框图。

[0036] 虽然在本文中通过列举若干实施方案和示意性附图的实例的方式描述了实施方案，本领域的技术人员应认识到，实施方案并不限于所描述的实施方案或附图。应理解，附图和对其的详细描述并非意图将实施方案限于所公开的特定形式，而是相反，其意图在于涵盖落入由所附权利要求书所界定的精神和范围内的所有修改、等同物以及替代方案。本文中使用的任何标题都仅用于组织目的，并且并不意图用于限制描述或权利要求书的范围。如贯穿本申请所用，词语“可”是以允许意义(即，意味着有可能)而不是强制意义(即，意味着必须)使用。类似地，词语“包括(include/including/includes)”意味着包括但不限于。

具体实施方式

[0037] 描述了用于管理被设计来操纵数百个或者甚至数千个并发数据生产商和数据消费者的大规模数据流的创建、存储、检索和处理的方法和设备的各种实施方案。如本文使用的术语“数据流”是指可由一个或多个数据生产商产生并由一个或多个数据消费者访问的数据记录的序列，其中每个数据记录假定为不变序列的字节。流管理服务(SMS)可提供编程接口(例如应用程序编程接口(API)、网页或网址、图形用户接口或命令行工具)以使得能够进行流的创建、配置和删除，以及在一些实施方案中的流数据记录的提交、存储和检索。涉及与SMS控制部件的交互的一些类型的流操作(诸如流创建或删除，或者下文描述的各种动态重新分区操作)可以在本文被称为“控制平面”操作，而通常不需要与控制部件交互的诸如数据记录提交、存储和检索的操作可以在本文被称为“数据平面”操作。动态提供组的计算、存储和网络资源在一些这种实施方案中可用来例如基于分区策略实现所述服务，所述分区策略允许流管理工作量以可度量的方式在许多服务部件中进行分配，如下文进一步详细描述的。缩写SMS本文中可用来是指流管理服务，并且还指包括用于实现流管理服务的虚拟和/或物理资源的采集的流管理系统。

[0038] 在各种实施方案中，SMS的一些消费者可以开发直接调用SMS编程接口的应用程序。然而，在至少一些实施方案中，除了SMS接口，可为消费者提供更高等级的抽象或应用程序等级的处理框架，所述处理框架可简化用于那些不希望使用由SMS直接支持的较低等级的流管理功能来开发应用程序的客户端的流处理的各个方面。这种框架可提供其自身的(例如在SMS接口的顶部上建立的)编程接口，从而使得消费者能够与较低等级的流管理操

作相比更加关注有待使用流记录实现的商业逻辑。较高等级的框架可实现为在一些实施方案中具有其自身的控制平面和数据平面部件的流处理服务(SPS),所述流处理服务可提供高级功能,诸如供应用于流处理的自动化资源、处理节点的自动化故障转移、构建任意流处理工作流图表的能力、支持短暂流、基于工作量变化或其他触发条件的动态重新分区等。在至少一些实施方案中,流管理服务、流处理服务或者两种服务可以被实现为虚拟化环境中的多租户管理的网络可访问服务。也就是说,在这种实施方案中各种物理资源(诸如计算机服务器或主机、存储装置、网络装置等)可以至少在一些情况下在不同消费者的流中共享,而不需要使消费者准确意识到资源是如何共享的或者甚至根本不需要使消费者意识到给定资源正在被共享。管理的多租户流管理和/或处理的控制部件管理的服务可以被动态添加、移除或者基于各种可应用策略来重新配置被用于特定流的节点或资源,一些策略可以是客户端可选择的。此外,控制部件还可负责明显地实现各种类型的安全协议(例如,来确保一名客户的流应用程序无法访问另一名客户的数据,即使至少一些硬件或软件可由这两名客户共享)、监测资源使用用于计费、产生可用于审计或调试的记录信息等。从管理的多租户服务的客户角度来看,由所述服务实现的控制/管理功能可消除支持大规模流应用程序中所涉及的许多复杂性。在一些情境中,这种多租户服务的消费者可以能够指示对于至少一些类型的流相关操作他们不希望共享资源,在这种情况下,对于那些类型的操作可至少将一些物理资源暂时地指定为是单租户的(即,限制为代表单个消费者或客户而执行的操作)。

[0039] 可以采取在各种实施方案中实现SMS和/或SPS控制平面和数据平面操作的多种不同方法。例如,关于控制平面操作,在一些实现方式中,可以设置冗余组的控制服务器或节点。冗余组可包括多个控制服务器,所述多个控制服务器中的一个服务器被指定为主服务器,所述主服务器负责响应关于各种流的管理请求,而另一个服务器可以被指定来在诸如当前的主服务器处产生故障(或者失去到主服务器的连接性)的触发条件的情况下接任主服务器。在另一种实现方式中,在网络可访问数据库服务处创建的一个或多个表可用来存储用于各种流的控制平面元数据(诸如分区映射),并且各种摄取、存储或检索节点可以能够视获取数据平面操作所需的元数据的子集的需要来访问所述表。下文提供了关于在不同实施方案中的SPS和SMS数据平面和控制平面功能的各个方面的细节。应注意,在实现流管理服务的一些实施方案中,可能不需要实现提供较高等级基元的流处理服务。在其他实施方案中,可仅仅将流处理服务的高等级编程接口暴露给消费者,并且由所用的较低等级的流管理接口对于客户可能是不可用的。

[0040] 根据一些实施方案,流管理系统可包括多个独立可配置的子系统,所述子系统包括主要负责获取或采集数据记录的记录摄取子系统、主要负责根据适用的持久性或耐用性策略保存数据记录内容的记录存储子系统、和主要负责响应于针对存储记录的读取请求的记录检索子系统。在一些实施方案中也可实现控制子系统,所述控制子系统包括一个或多个管理或控制部件,所述一个或多个管理或控制部件负责通过例如动态确定和/或初始化用于在选定资源(诸如虚拟或物理服务器)处的摄取、存储和检索子系统中的每一个的节点的所需数量来配置其余子系统。摄取、存储、检索和控制子系统中的每一个可使用相应多个硬件和/或软件部件来实现,所述多个硬件和/或软件部件可以共同称为子系统的“节点”或“服务器”。因此,SMS的各种资源可以从逻辑上说属于四种功能种类中的一种:摄取、存储、

检索和控制。在一些实现方式中，可建立相应组控制部件用于其他子系统中的每一个，例如可实现独立摄取控制子系统、存储控制子系统和/或检索控制子系统。这种控制子系统可以各自负责识别用于对应子系统的其他节点的资源和/或负责响应于来自客户端或来自其他子系统的管理查询。在一些实现方式中，能够执行各种类型的SMS和/或SPS功能的节点池可以提前进行设置，并且那些节点池的选择构件可以根据需要被分配给新的流或者新的处理阶段。

[0041] 在至少一些实施方案中可实现流分区策略和相关联的映射，以例如在不同组摄取、存储、检索和/或控制节点之间分配数据记录的子集。例如，基于选择用于特定数据流的分区策略并且基于其他因素(诸如记录摄取率和/或检索率的期望值)，控制部件可确定最初(即在流创建时间处)应当建立多少节点(例如，进程或线程)用来摄取、存储和检索，以及这些节点应当如何被映射成虚拟机和/或物理机。随着时间的推移，与给定流相关联的工作量可增加或减少，这(除了其他触发条件)可能导致流的重新分区。这种重新分区可以涉及各种参数的变化，诸如用于确定记录的分区的功能，使用的分区键，分区的总数，摄取节点、存储节点或检索节点的数量，或者不同物理或虚拟资源上的节点的放置。在至少一些实施方案中，重新分区可以在没有中断数据记录流动的情况下使用下文中进一步详细描述的技术动态地实现。不同的分区方案和重新分区触发标准在一些实施方案中可例如基于客户端提供的参数或者基于SMS控制节点的启发而用于不同的数据流。在一些实施方案中，也许有可能例如基于客户偏好、流的预期寿命或其他因素限制重新分区的数量和/或频率。

[0042] 可在不同的实施方案中实现许多不同的记录摄取策略和接口。例如，在一些实施方案中，客户端(例如被配置来代表SMS的消费者调用SMS的编程接口的可执行部件或模块)可利用在线提交接口或参考提交接口。对于在线提交，数据记录的内容或本体在这种实施方案中可以被包括为提交请求的部分。相比之下，在参考提交请求中，可提供地址(诸如存储装置地址、数据库记录地址或者URL(统一资源定位符))，数据记录的内容或本体可从所述地址获取。在一些实现方式中，也可以或者替代地支持混合提交接口，其中数据记录的前N个字节可以被包括在线，而剩余字节(如果有的话)被提供作为参考。在这种情境下，较短的记录(其本体小于N字节长)可完全由提交请求指定，而较长记录的部分可能不得不从对应地址获取。

[0043] 除了在摄取期间对于指定记录内容的不同替代物，在一些实施方案中，也可以实现与摄取策略有关的各种确认或去重复。例如，对于一些流应用程序，客户端可能希望确保每个和所有的数据记录由SMS可靠地摄取。在大型分布式流管理环境中，数据包可能丢失，或者沿着在数据生产商与摄取节点之间的路径可能不时地产生各种故障，这可能潜在地导致一些提交的数据丢失。因此，在一些实施方案中，SMS可实现至少一次摄取策略，根据所述策略记录提交者可一次或多次提交相同的记录直到从摄取子系统接收到肯定的确认。在正常操作条件下，记录可提交一次，并且提交者在接收的摄取节点已获取和存储记录之后可接收到确认。如果确认丢失或延迟，或者如果记录提交请求自身丢失，那么提交者可一次或多次重新提交相同的数据记录，直到最终接收到确认。摄取节点可以例如基于如果已由提交者接收到确认那么所述记录将不重新提交的期望来产生针对每个提交的确认，不管所述提交是否是重复的。然而，摄取节点在至少一些实施方案中可负责识别出相同的数据记录已提交多次，并且负责避免不必要地存储重复数据的新副本。在一个实施方案中，可支持至

少两个版本的至少一次摄取策略—一个版本(可称为“至少一次摄取、无重复”)其中SMS负责去重复数据记录(即仅响应于一组两个或更多个提交中的一个提交来确保数据被存储在SMS存储子系统处),以及一个版本,其中允许通过SMS的数据记录存储的重复(可称为“至少一次、允许重复”)。所述至少一次、允许重复的方法对于流应用程序可能是有用的,其中存在很少或者没有数据记录重复的负面后果,和/或对于执行它们自己的重复消除的流应用程序可能是有用的。也可以支持其他摄取策略,诸如尽力摄取策略,其中不需要针对所有提交的数据记录的确认。如果尽力摄取策略在至少一些实施方案中生效,那么少量数据记录的丢失是可接受的。客户端可以在各种实施方案中针对各种流来选择它们希望使用哪一种摄取策略。

[0044] 关于流记录的存储,在至少一些实施方案中也可支持许多替代性策略。例如,客户端可以能够从由SMS支持的若干策略中选择持久性策略,所述策略控制记录存储的这些方面如:待存储的给定数据记录的副本的数量、待用于副本的存储技术的类型(例如易失性或非易失性RAM、基于旋转磁盘的存储装置、固态装置(SSD)、网络附属存储装置等)等。例如,如果客户端为基于磁盘的存储装置选择N次复制持久性策略,那么数据记录提交可能不认为是完整的直到已将记录的N个副本安全写入N个相应磁盘装置。在其中使用基于磁盘的存储装置的至少一些实施方案中,SMS存储子系统可以尝试随后将给定分区的输入的数据记录写入磁盘,例如以避免磁盘寻道的性能影响。可使用如下文所述的各种技术来产生序列号用于(并存储有)数据记录,包括例如能够基于摄取时间来进行有序记录检索的基于时间戳的技术。在至少一些实施方案中,给定分区的数据记录可以被存储在一起,例如在磁盘上连续地并且与其他分区的数据记录分开地存储。在一些实现方式中,根据保留策略(由客户端选择或由SMS选择)或者去重复时间窗口策略(指示在提交任何给定数据记录之后的时期,在所述时期内,SMS可能需要确保没有将所述给定数据记录的副本存储在SMS存储子系统中,即使提交了一些副本),至少一些数据记录可存档到不同类型的存储装置和/或在一段时期后从SMS删除。这种移除操作在本文中可称为流“修整”。在一些实施方案中,客户端可提交流修整请求,例如通知SMS不再需要指定的数据记录并且因此从提交修整请求的客户端的角度来看可删除所述指定的数据记录,或者明确地请求删除指定的数据记录。在可能存在消费给定流的数据记录的多个客户端的情境下,SMS可负责确保给定记录在其已由所有感兴趣的消费者访问之前不被过早地删除或修整。在一些实现方式中,如果存在给定流的N个数据消费者,那么在删除所述流的给定记录R之前,SMS可等待直到已确定所有N个数据消费者已读取或处理过R。例如,SMS可基于来自消费者的相应修整请求或者基于数据消费者在所述流内已进展到什么程度的相应指示来确定R已由所有消费者读取。在一些实施方案中,一些类型的数据消费者(诸如测试相关应用程序)可接受在已访问至少数据记录的小子集之前将它们删除。因此,在至少一些实施方案中,应用程序在检索之前可以能够通知SMS关于数据删除的可接受性,并且SMS可根据通知来安排删除。在一些实施方案中,存档策略可以例如被实现为指示例如流数据记录应当被复制到的存储装置的类型的数据保留策略以及用于这种副本的安排策略的部分。

[0045] 在至少一些实施方案中,还可支持多个编程接口用于记录检索。在一个实施方案中,可以使用基于迭代器的方法,其中一个编程接口(例如获取迭代器(getIterator))可用来在流的分区内的指定逻辑偏移处(例如基于序列号或时间戳)实例化和定位迭代器或指

针。不同的编程接口(诸如获取下个记录(getNextRecords))可以随后被用来从迭代器的当前位置开始顺序地读取特定数量的数据记录。迭代器的实例化可以实际上允许客户端在流分区内指定用于记录检索的任意或随机的开始位置。在这种实施方案中,如果客户端希望以随机访问模式读取数据记录,那么客户端可能不得不重复地创建新的迭代器。在基于旋转磁盘的存储系统中,频繁的随机访问所需的磁盘寻道可能显著地影响I/O响应时间。因此,在至少一些实施方案中,当客户端的动机是顺序地而非随机地读取流数据记录时,与被应用到顺序读取访问不同的(例如更高的)计费率可应用至随机读取访问。因此,例如,在一些实现方式中,客户端可以每个获取迭代器呼叫被计费X个货币单位,并且通过获取下个记录检索的每个记录被计费Y个货币单位,其中X>Y。当替代性客户端接口被支持用于其他操作种类(诸如摄取)时,在至少一些实施方案中,针对替代物的计费率或价格也可能不同,例如客户端可能对于参考提交请求比对于在线提交请求要价更多,正如客户端可能对于随机读取比对于顺序读取要价更多。在各种实施方案中,其他因素也可能影响计费,诸如数据记录的大小、随时间的写对读请求的分配、所选的持久性策略等。

[0046] 根据一些实施方案,流处理服务(SPS)可允许客户端指定包括很多处理阶段的任意复杂的处理工作流程,其中在给定阶段处执行的处理的输出可用作对于零阶段或更多其他阶段的输入。在一些实施方案中,(类似于针对用于摄取、存储和检索数据记录的SMS所描述的那些)分区策略可用来在各个阶段处的多个工作节点中划分处理工作量。在一个这种实施方案中,可实现编程SPS接口从而使得客户端能够指定针对任何给定阶段的各种配置设置,包括例如用于所述阶段(例如,数据记录有待从其检索的一个或多个流连同用于所述流的分区策略)的输入数据源、有待在所述阶段处执行的处理操作、以及用于来自所述阶段的输出或结果分配的描述符或规范(例如,输出是否以不同的流的形式被保存到存储位置、发送至网络端点或者馈送至一个或多个其他处理阶段中)。在至少一些实施方案中,指定用于SPS阶段的处理操作可以是幂等的:即,如果给定处理操作在相同的输入数据上被执行了多次,那么操作结果不会不同于如果操作只被执行一次将已获得的结果。如果处理操作是幂等的,那么从故障(例如在SPS阶段处的工作节点故障)中恢复可以被简化,如下文进一步详细描述的。根据一些实施方案,在一些或所有SPS阶段处可允许非幂等的处理操作。

[0047] 至少部分基于配置信息,诸如输入流分区策略和随后通过SPS编程接口接收的处理操作的性质,在各种实施方案中SPS控制服务器可确定最初有多少工作节点有待设置用于处理工作流程的各个阶段。当确定工作节点的初始数量和放置时,也可以考虑有待用于工作节点(例如,正在使用的虚拟机或物理机)的资源的执行能力。可以实例化所选择数量的工作节点(所述工作节点可以在一些实现方式中每个包括可执行线程或可执行过程)。每个工作节点可以被配置,例如以从适当的输入资源(例如从一个或多个流分区的检索节点)获取数据记录,在数据记录上执行指定的处理操作、并将处理结果传输至指定的输出目的地。此外,在至少一些实施方案中,可以实现检验点方案,根据所述检验点方案,给定的工作节点可以被配置来存储进度记录或者指示分区的在那个工作节点处已被处理的部分的检验点,其中假设分区记录被顺序地处理。在一些实现方式中,工作节点可以例如将进度记录定期(例如,每隔N秒或者每隔R个已处理的数据记录)写入永久性存储装置和/或响应来自SPS控制服务器的检验点请求。

[0048] 在一些实施方案中,进度记录可用于从工作节点故障的快速恢复。例如,SPS控制

服务器可例如使用心跳机构和/或通过监测资源利用水平(诸如CPU利用率、I/O装置利用率或网络利用率水平)来随时间监测各个工作节点的健康状况。响应于由SPS控制服务器做出的特定工作节点呈不需要或者不健康状态(例如,如果其是无响应或过载的)的确定,替换工作节点可以被实例化以接任特定工作节点的责任。替换工作节点可访问由替换工作节点存储的最近的进度记录,以识别替换工作节点应当处理的所述组数据记录。在处理操作是幂等的实施方案中,即使一些操作被重复(例如,由于最近的进度记录在替换工作节点的实例化之前某时被写入),所述处理的总体结果将不会受到故障和替换的影响。在一些实现方式中,除了存储指示给定流或分区的已由其处理过的子集的进度记录,工作节点还可以被配置来存储累积的应用程序状态信息。例如,如果流处理工作流程负责基于分析指示服务使用指标的流数据记录来确定针对特定服务的客户端计费总额,那么工作节点可定期存储针对各种客户端确定的累积的计费总额。

[0049] 在至少一些实施方案中,SPS控制服务器还可以被配置来通过开始其他行动来响应于各种其他触发,诸如改变工作量水平或检测到的工作量失衡(例如如果针对一个分区的摄取率不成比例地高于其他分区的那些摄取率),所述其他行动诸如为各个阶段请求输入流的动态重新分区、在给定阶段处改变分配至给定分区的工作节点的数量、为一些阶段分配更高性能的工作节点或者将工作节点从一个物理资源转移至具有不同性能能力的另一个物理资源。在一些实施方案中,例如,响应于由SPS控制服务器做出的有待针对给定阶段完成尽力恢复策略(而不是基于检验点的恢复策略)的确定,上文所述类型的进度记录可以不由至少一些SPS阶段的工作节点存储。在这种尽力恢复策略的一些实现方式中,替换工作节点可以当接收新的数据记录时对它们进行简单处理,而不需要访问进度记录。在一些实施方案中,如果客户端希望在SPS阶段处实现尽力恢复策略,那么在所述阶段处执行的流处理操作不一定需要是幂等的。在有待在SPS阶段处在流记录上执行的非幂等处理操作的实施方案中,可能不支持基于检验点的恢复,并且可使用诸如尽力恢复的不同的恢复方案。在至少一个实施方案中,在SPS阶段处可以仅允许幂等的流处理操作。

[0050] 一些流的数据记录可以包含敏感或机密信息,或者在SPS阶段处执行的处理操作可包括专有算法的使用,所述专有算法若由竞争对手发现可能是有问题的。客户端可能因此关心流管理和处理操作的各个种类的安全性,尤其是如果使用位于不完全由客户端自身控制的供应商网络数据中心处的资源来执行所述操作。由诸如公司或公共部门组织的实体建立以提供通过互联网和/或其他网络到分布组的客户端可访问的一种或多种网络可访问服务(诸如各种类型的基于云的数据库、计算或存储服务)的网络在本文中可以被称为供应商网络。在一些实现方式中,客户端可以能够从针对它们的数据流的多种安全相关的选项中进行选择。如上所述,组合的SPS和SMS配置可包括属于多种不同的功能种类的节点,诸如用于SMS和/或SPS的控制节点、SMS摄取节点、SMS存储节点、SMS检索节点以及SPS处理或工作节点。在一些实施方案中,做出的可用于客户端的基于安全性的选择可包括对于各种类型的节点的放置和地点的选项。例如,在一个实施方案中,客户端可以能够请求在位于客户端所有的设施上的计算装置处实现用于流工作流程的一个或多个处理阶段的SPS工作节点,即使流记录最初是使用位于供应商网络处的资源来采集和/或存储的。响应于这种放置请求,用于给定流的不同的功能种类的节点可以在具有不用的安全特性或特征的相应资源集合处被实例化。

[0051] 在不同的实施方案中，所述资源集合可以在各种安全相关的特性上不同于彼此，包括例如物理地点、正在使用的物理安全协议(例如其具有对资源的物理访问)、网络隔离等级(例如资源的网络地址对各种实体的可见程度)、多租户对单租户等。在一些实施方案中，客户端可以能够在供应商网络内建立隔离的虚拟网络(IVN)，其中给定客户端被赋予对包括在那个客户端的IVN内的各种装置的网络配置的实质性控制权。具体地说，客户端可以能够限制对分配给它们的IVN内的各个服务器或计算实例的网络地址(例如，互联网协议或IP地址)的访问。在这种实施方案中，客户端可以请求它们的SMS或SPS节点中的某些子集在指定IVN内被实例化。在诸如虚拟化实例主机(其可以通常被配置成多租户主机)的供应商网络资源被用于各个种类的SMS或SPS节点的实施方案中，客户端可以请求在实例主机上实例化一些组节点，所述实例主机被限制来实现仅属于客户端的实例(即一些SMS或SPS节点可在被配置成单租户主机的实例主机处实现)。

[0052] 在一些实施方案中，作为另一种安全相关的选项，客户端可以请求特定流的数据记录在将它们在网络链接上传输之前进行加密，例如在SMS处、在摄取子系统与存储子系统之间、在存储子系统与检索子系统之间、在检索子系统与SPS工作节点之间和/或在工作节点与SPS输出目的地之间摄取之前进行加密。在一些实施方案中，客户端可指定待使用的加密算法。在一个实施方案中，诸如TLS(传输层安全)协议或SSL(安全套接层)协议的安全网络协议可以被用于数据记录传输和/或用于传输SPS处理结果。

[0053] 数据流概念和概述

[0054] 图1提供了根据至少一些实施方案的数据流概念的简化概述。如图所示，流100可以包括多个数据记录(DR)110，诸如DR 110A、110B、110C、110D和110E。诸如数据生产商120A和120B的一个或多个数据生产商120(也可称为数据源)可执行写操作151以产生流100的数据记录的内容。许多不同类型的数据生产商在不同的实施方案中可产生数据流，例如像移动电话或平板电脑应用程序、传感器阵列、社交媒体平台、记录应用程序或系统记录部件、不同种类的监控代理等。一个或多个数据消费者130(诸如数据消费者130A和130B)可执行读操作152以访问由数据生产商120产生的数据记录的内容。在一些实施方案中，数据消费者130可包括例如流处理阶段的工作节点。

[0055] 在至少一些实施方案中，如存储在SMS中的给定的数据记录110可包括数据部分101(例如分别是DR 110A、110B、110C、110D和110E的数据部分101A、101B、101C、101D和101E)和序列号SN 102(例如分别是DR 110A、110B、110C、110D和110E的SN 102A、102B、102C、102D和102E)。在描绘的实施方案中，序列号102可指示将DR接收在流管理系统处(或者在流管理系统的特定节点处)的顺序。在一些实现方式中，数据部分101可包括不变的未解释的字节序列：也就是说，一旦完成写操作152，由于写入所产生的DR的内容可不由SMS改变，并且通常SMS可能不清楚数据的语义。在一些实现方式中，给定流100的不同的数据记录可包括不同的数据量，而在其他实现方式中，给定流的所有数据记录可具有相同大小。在至少一些实现方式中，SMS的节点(例如摄取子系统节点和/或存储子系统节点)可以负责产生SN 102。如下文进一步详细描述的，数据记录的序列号不需要一直是连续的。在一种实现方式中，作为写入请求的部分，客户端或数据生产商120可提供最小序列号有待用于对应的数据记录的指示。在一些实施方案中，数据生产商120可例如通过提供存储装置地址(诸如装置名称和装置内的偏移量)或可从其获得数据部分的网络地址(诸如URL)来提交包含到数

据记录的数据部分的指针(或其地址)的写入请求。

[0056] 流管理服务可负责接收来自数据生产商120的数据、存储所述数据、并且使数据消费者130能够在各种实施方案中以一种或多种访问模式访问所述数据。在至少一些实施方案中,流100可以被分区或“碎片化”以分配接收、存储和检索数据记录的工作量。在这种实施方案中,分区或碎片可以被基于数据记录的一个或多个属性来选择用于传入的数据记录110,并且待摄取、存储或检索数据记录的特定节点可基于所述分区来识别。在一些实现方式中,数据生产商120可提供具有各自的写操作的可用作分区属性的明确的分区键,并且这种键可以被映射至分区标识符。在其他实现方式中,SMS可基于如数据生产商120的身份、数据生产商的IP地址的这类因素或者甚至基于提交的数据内容来推断分区ID。在将数据流分区的一些实现方式中,序列号可在按分区的基础上进行分配,例如尽管序列号可指示接收特定分区的数据记录的顺序,在两个不同分区中的数据记录DR1和DR2的序列号可能未必指示接收DR1和DR2的相对顺序。在其他实现方式中,序列号可在流宽而不是在按分区的基础上分配,以使得如果分配给数据记录DR1的序列号SN1小于分配给数据记录DR2的序列号SN2,这将意味着DR1与DR2相比由SMS更早地接收,不管DR1和DR2属于哪个分区。

[0057] 由SMS支持的检索和读取接口可允许数据消费者130在各种实施方案中继续地和/或以随机顺序访问数据记录。在一个实施方案中,可支持基于迭代器的读取应用程序编程接口(API)组。数据消费者130可提交请求以获取用于数据流的迭代器,其中所述迭代器的初始位置由指定序列号和/或分区标识符指示。在将引发器实例化之后,数据消费者可提交请求以从所述流或分区内的所述初始位置开始按相继顺序读取数据记录。在这种实施方案中,如果数据消费者希望以一些随机顺序读取数据记录,那么新的迭代器可能不得不针对每次读取而实例化。在至少一些实现方式中,可以通常使用避免磁盘寻道的顺序写操作来将给定分区或流的数据记录以序列号顺序写入基于磁盘的存储装置。顺序读操作也可避免磁盘寻道的开销。因此,在一些实施方案中,数据消费者可以被使用价格激励来鼓励执行比随机读取更多的顺序读取:例如,诸如迭代器实例化的随机访问读取操作可具有比顺序访问读取操作更高的相关联的计费率。

[0058] 示例性系统环境

[0059] 图2提供了根据至少一些实施方案的在流管理系统(SMS)和包括流处理阶段的采集的流处理系统(SPS)的各种子部件之中的数据流的概述。如图所示,SMS 280可包括摄取子系统204、存储子系统206、检索子系统208和SMS控制子系统210。如下文所述,SMS子系统中的每一个可包括例如使用在供应商网络(或者客户端所有的或第三方设施)的各种资源处实例化的相应的可执行线程或进程实现的一个或多个节点或部件。摄取子系统204的节点可以基于用于所述流的分区策略来被(例如,由SMS控制子系统210的节点)配置来获取来自数据生产商120(诸如120A、120B和120C)的特定数据流的数据记录,并且每个摄取节点可将接收的数据记录传递至存储子系统206的对应节点。存储子系统节点可根据选择用于所述流的持久性策略将数据记录保存在任何不同类型的存储装置上。检索子系统208的节点可响应于来自数据消费者的读取请求,诸如SPS 290的工作节点。可以借助于SPS控制子系统220建立流处理阶段215,诸如SPS 290的阶段215A、215B、1215C和215D。每个阶段215可包括由SPS控制子系统220配置以在接收的数据记录上执行一组处理操作的一个或多个工作节点。如图所示,一些阶段215(诸如215A和215B)可直接从SMS280获取数据记录,而其他的

阶段(诸如215C和215D)可从其他阶段接收它们的输入。在一些实施方案中,多个SPS阶段215可并行操作,例如在阶段215A和215B处,不同的处理操作可在从相同的流检索到的数据记录上同时执行。应注意,类似于图2中示出的用于特定流的那些的相应的子系统和处理阶段也可针对其他流而实例化。

[0060] 在至少一些实施方案中,图2中示出的子系统和处理阶段的至少一些节点可使用供应商网络资源来实现。如之前指出的,由诸如公司或公共部门组织的实体建立以提供通过互联网和/或其他网络到分布组的客户端可访问的一种或多种网络可访问服务(诸如各种类型的基于云的数据库、计算或存储服务)的网络在本文中可以被称为供应商网络。一些服务可用来构建更高水平的服务:例如计算、存储或数据库服务可用作用于流管理服务或流处理服务的构建模块。供应商网络的至少一些核心服务可以在称为"实例"的服务单元中被打包用于客户端使用:例如,由虚拟化计算服务实例化的虚拟机可代表"计算实例",并且诸如由存储装置实例化的块级体积的存储装置可以被称为"存储实例",或者数据库管理服务器可以被称为"数据库实例"。计算装置在本文中可以被称为"实例主机"或者更简单地称为"主机",所述计算装置诸如在其处可实现供应商网络的各种网络可访问服务的这种单元的服务器。在一些实施方案中,摄取子系统204、存储子系统206、检索子系统208、SMS控制系统210、处理阶段215和/或SPS控制子系统220的节点可包括在多个实例主机上的各个计算实例处执行的线程或进程。给定的实例主机可包括若干计算实例,并且在特定实例主机处的计算实例的采集可以被用来实现用于一个或多个客户端的各种不同的流的节点。在一些实施方案中,存储实例可以被用于存储各种流的数据记录,或者用作流处理阶段的结果的目的地。如下文参考图15和图16描述的,随着时间的变化,控制子系统节点可响应于各种触发条件动态地修改其他子系统的群体,例如通过添加或去除节点、改变节点到处理实例或计算实例或实例主机的映射或者重新分区给定流同时仍然继续接收、存储和处理数据记录。

[0061] 在供应商网络资源被用于流相关操作的实施方案的内容中,术语"客户端"当用作给定通信的来源或目的地时可以是指由实体(诸如组织、具有多个用户或单个用户的组)所有、管理或分派给所述实体的计算装置、进程、硬件模块或软件模块中的任何一个,所述实体能够访问和利用供应商网络的至少一个网络可访问服务。一种服务的客户端可以自身使用另一种服务的资源来实现,例如流数据消费者(流管理装置的客户端)可以包括计算实例(由虚拟化计算服务提供的资源)。

[0062] 给定的供应商网络可包括托管各种资源池的很多数据中心(其可以遍及不同的地理区域来分配),诸如物理和/或虚拟计算机服务器、每个具有一个或多个存储装置的存储服务器、网络设备等的采集需要实现、配置和分配由供应商提供的基础结构和服务。在这种实施方案中,许多不同的硬件和/或软件部件可共同用来实现所述服务中的每一种,所述硬件和/或软件部件中的一些可在不同的数据中心处或者在不同的地理区域中被实例化或执行。客户端可与供应商网络处的资源和服务交互,所述资源和服务来自供应商网络外部的位于客户端所有的或客户端管理的处所或数据中心的装置和/或来自供应商网络内的装置。应注意,尽管供应商网络用作其中可实现本文所述的许多流管理和处理技术的一种示例性内容,那些技术还可应用至除了供应商网络以外的其他类型的分布式系统,例如应用至由单个企业实体针对其自身的应用程序而操作的大规模分布式环境。

[0063] 编程接口实施例

[0064] 如上文所指示的，在至少一些实施方案中，SPS可利用SMS编程接口来构建更高水平的功能，所述功能可以更容易地由SPS客户端使用以实现用于各种基于流的应用程序的所需商业逻辑。当考虑到SPS功能与SMS功能之间的差异，类推可能是有用的。SPS功能可以大体上与呈更高水平的语言(诸如C++)的编程语言结构进行比较，而SMS功能可以大体上与汇编语言指令进行比较，编程语言结构通过编译器转化成汇编语言指令。也许有可能直接使用汇编语言指令来实现相同操作，但是呈更高水平语言的编程可能通常更容易用于许多种类的消费者或用户。类似地，也许有可能使用由SMS提供的基元来实现各种应用程序，但是通过使用SPS特征这可能更容易完成。SPS处理操作(诸如数据记录上执行的幂等的处理操作)可以在流记录的数据内容上实现，而SMS操作被执行以获取、存储和检索记录自身，而通常不考虑所述记录的内容。图3示出根据至少一些实施方案的在SMS SPS处可实现的相应组编程接口的实例。通过实例，许多不同的应用程序编程接口(API)被指示用于SMS和SPS。示出的API不意图是在任何给定实现方式中所支持的那些API的详尽列表，并且示出的API中的一些在给定实现方式中可能不支持。

[0065] 如由箭头350所指示的，SPS客户端375可调用SPS编程接口305以配置处理阶段。各种类型的SPS编程接口305可在不同的实施方案中实现。例如，创建流处理阶段(createStreamProcessingStage)API可以使得客户端能够请求用于指定输入流的新的处理阶段215的配置，以使得所述阶段的工作节点的每一个被配置来执行在接口调用中所指定的一组幂等操作，并且将结果分配至由输出分配描述符或策略指示的目的地。在创建流处理阶段API或其等效物的一些版本中，客户端也可以请求创建输入流，而在其他版本中，输入流可能不得不在产生处理阶段之前创建。恢复策略可以被指定用于工作节点，从而例如指示基于检验点的恢复技术是否有待使用或者尽力恢复技术是否是优选的。在一些实施方案中，可支持初始化工作节点(initializeWorkerNode)API，以在指定阶段处请求工作节点的明确实例化。在实现基于检验点的恢复的实施方案中，可支持保存检验点(saveCheckpoint)API，以允许客户端请求由工作节点产生进度记录。

[0066] 在不同的实施方案中可支持各种类型的SPS输出管理API，诸如设置输出分配(setOutputDistribution)API，客户端通过所述设置输出分配API可指示使用在指定阶段处执行的处理操作的结果以及将用于新创建的流的特定分区策略而待创建的一个或多个流。一些处理阶段可以被主要配置用于重新分区，例如基于记录属性集A1将数据记录映射至N1分区的一种分区功能PF1可以用于输入流S1，并且处理阶段可以用来实现不同的分区功能PF2以(使用不同的属性集A2或者相同的属性集A1)将那些相同的数据记录映射至N2分区。诸如链接阶段(linkStage)的一些SPS API可以被用来配置包括多个阶段的任意图形(例如有向无环图)。在一些实施方案中，可支持到第三方或开源流处理框架或服务的连接器。在一个这种实施方案中，SPS阶段可用于(例如通过在所述阶段处执行的处理操作的适当格式化的结果)准备数据记录，用于由存在的第三方或开源系统的消费。在描述的实施方案中，诸如创建第三方连接器(createThirdPartyConnector)的API可用于建立这种连接器，并且可通过一个或多个连接器模块执行SPS阶段的结果到与第三方系统兼容的格式的适当转化，所述一个或多个连接器模块作为创建第三方连接器调用的结果被实例化。

[0067] SPS可调用SMS API 307以执行其功能中的至少一些，如由箭头352所指示的。在描

绘的实施方案中，SMS API 307可包括例如创建流(createStream)和删除流(deleteStream)(以分别创建和删除流)以及获得流信息(getStreamInfo)(以获取用于流的元数据，诸如负责给定分区的各种类型节点的网络地址)。放置记录(putRecord)接口可用来写入数据记录，而获得迭代器(getIterator)和获得下个记录(getNextRecord)接口可分别用于非顺序和顺序的读取。在一些实施方案中，重新分区流接口可用于请求指定流的动态重新分区。希望这样做的客户端370可直接调用SMS API 307，如由箭头354所指示的。如之前指示的，在其他实施方案中也可实现各种其他SMS和/或SPS API，并且在一些实施方案中可以不实现图3中列出的API中的一些。

[0068] 在各种实施方案中，除了API之外的编程接口也可以或代替地被实现用于SPS或SMS。这种接口可包括图形用户接口、网页或网站、命令行接口等。在一些情况下，基于网络的接口或GUI可使用API作为构建模块，例如基于网络的交互可以在SMS或SPS的控制部件处产生一个或多个API的调用。图4示出根据至少一些实施方案的示例性基于网络的接口，所述接口可实现为使得SPS客户端能够产生流处理阶段的图形。如图所示，接口包括具有消息区402的网页400、图形菜单区404和图形设计区403。

[0069] 用户可以被提供关于消息区402中的流处理图形的构建的常规指令，以及能够用于学习更多关于流概念和基元的链接。许多图形图标可在菜单区404中被提供为流处理图形工具集的部分。例如，客户端作为各个SPS处理阶段的输入或输出可以被允许指示持续流451、短暂流452或到第三方处理环境的连接器453。关于基于网络的接口被实现用于其的SPS/SMS，持续流451可以被限定为其数据记录被存储在持久性存储装置上的流，所述持久性存储装置诸如磁盘、非易失性RAM或SSD，并且短暂流452可以被限定为其数据记录不需要被存储在持久性存储装置处的一种流。短暂流可以例如从SPS阶段的输出产生，所述输出被预期由待实现尽力恢复策略的不同的SPS阶段作为输入消费。

[0070] 在示例性SPS图形构建网页400中支持两种类型的处理阶段：阶段455，其中使用基于检验点的工作节点恢复(例如，每个工作节点不时地保存进度记录，并且在特定工作节点故障的情况下，替换节点参考故障的节点的进度记录以确定开始处理哪一个数据记录)；以及阶段456，其中使用尽力恢复(例如，替换工作节点不参考进度记录，但是仅仅当接收到新的数据记录时开始对其进行处理)。关于有待在每个阶段处执行的处理操作的细节可以通过在图形构建区403中的对应图标上点击而进入，如由消息区402中的指令所指示的。除了用于流、连接器和处理阶段的图标，菜单区404还包括指示第三方或外部流处理系统的图标类型459，以及指示可以在供应商网络处实现的存储装置的节点的图标类型460，所述供应商网络的资源正用于所述处理阶段。

[0071] 在图4中示出的示例性情境下，客户端已构建图形405，所述图形405在图形设计区403内包括三个处理阶段412、415和416。被配置成使用基于检验点的恢复的处理阶段412使用持续流411作为输入。阶段412处的处理的输出或结果被发送至两个目的地：呈形成阶段415的输入的不同的持续流413的形式；以及呈形成阶段416的输入的短暂流414的形式。阶段415和416都为它们的工作节点使用尽力恢复策略。阶段415的输出被以短暂流的形式发送至存储服务节点419。阶段415的输出被通过连接器417发送至第三方处理系统418。“保存图形”按钮420可以被用来例如以任何适当格式保存处理阶段图形的表示，所述格式诸如JSON(JavaScript对象标记法)、XML(可延伸标记语言)或者YAML。在各种实施方案中，任意

复杂的处理工作流程可使用类似于图4中示出的那些工具来构建。使用这种工具创建的工作流程可随后被激活，并且这种激活可引起SMS API的调用，例如以获取用于处理阶段(诸如图4的阶段412)的数据记录，获得迭代器接口和/或获得下个记录接口可在流411上调用。

[0072] 图5示出根据至少一些实施方案的在SMS处可实现的编程记录提交接口和记录检索接口的实例。在描绘的实施方案中，诸如示出的DR 110K和110Q的数据记录可通过各种类型的编程摄取接口510提交给SMS。在一些实施方案中，DR 110可包括四种类型的元件：流标识符，诸如501A(用于流“S1”)或501B(用于流“S2”)；记录的数据或本体的指示；任选的分区键504(诸如504A或504B)；以及任选的排序偏好指示符506(诸如排序偏好指示符506A和506B)。在一些数据记录中，数据本身可在线提供(例如DR 110K的在线数据502)，而对于其他数据记录可提供指针或地址503，从而为SMS指示网络可访问地点(或者不需要网络传输的本地装置处的地址)。在一些实施方案中，给定流可支持在线数据记录提交和参考(基于地址的)数据记录提交。在其他实施方案中，给定流可能需要数据生产商供应所有的在线数据或者所有的参考数据。在一些实现方式中，数据记录提交可包括待用于所述记录的分区标识符。

[0073] 在描绘的实施方案中，传入的数据记录110可基于分区策略被引导至相应的摄取和/或存储节点。类似地，记录检索也可以是基于分区的，例如一个或多个检索节点可以被指定用于响应于针对给定分区的记录的读取请求。对于一些流，数据生产商可能需要提供具有各自的数据记录写入请求的明确的分区键。对于其他流，SMS可以能够根据分区方案来分配数据记录，所述分区方案依赖于元数据或除了明确供应的分区键之外的属性，例如，与提交的数据生产商有关的识别信息可用作分区键，或者可使用提交的数据生产商的IP地址的部分或所有，或者可使用正在提交的数据的部分。在一些实现方式中，例如，可将散列函数应用至分区键以获取一定大小的整数值，诸如128位整数。所述大小(例如从0至2^128-1)的正整数的全部范围可以被划分成N个连续的子区间，其中每个子区间代表相应的分区。因此，在这种实例中，确定或供应用于数据记录的任何给定的分区键将被散列成对应的128位整数，并且所述整数从属的128位整数的连续的子区间可指示所述数据记录从属的区间。关于分区策略和它们的使用的另外细节在下文中参考图15来提供。

[0074] 负责摄取或接受特定分区的数据记录、存储所述数据记录并且响应于针对所述特定分区的读取请求的所述组节点共同称为用于图5中的分区的ISR(摄取、存储和检索)节点。标记Sj-Pk被用来指示流Si的第k个分区。在示出的实施方案中，ISR节点520A被配置用于摄取、存储和检索分区S1-P1的记录，ISR节点520B被建立用于分区S1-P2的记录，ISR节点520C被建立用于分区S1-P3的记录，ISR节点520K被建立用于分区S2-P1的记录，并且ISR节点520L被建立用于分区S2-P2的记录。在一些实施方案中，摄取子系统、存储子系统或者检索子系统的给定节点可以被配置来处理超过一个分区(或者超过一个流的超过一个分区)的数据记录。在一些实施方案中，给定流的单个分区的记录可以由超过一个节点摄取、存储或检索。指定用于给定分区Sj-Pk的摄取节点的数量可以在至少一些情况下不同于指定用于不同分区Sj-P1的摄取节点的数量，并且还可以不同于指定用于Sj-Pk的存储节点的数量和/或指定用于Sj-Pk的检索节点的数量。在一些实施方案中，关于摄取和/或检索，SMS控制节点可实现API(诸如获得流信息)，以允许客户端确定哪些节点负责哪些分区。数据记录与分区之间以及分区与ISR节点(或控制节点)之间配置的映射可以随时间而修改，如下文中

在关于动态重新分区的讨论中所描述的。

[0075] 在一些实施方案中，若干不同的编程接口580可实现用于从给定分区检索或读取流数据记录。如图5中所示，一些检索接口581可实现用于非顺序的访问，诸如获得迭代器接口(以在具有指定序列号的数据记录处或其后实例化迭代器或读取指针)或获得记录(getRecord)接口(以读取具有指定序列号的数据记录)。其他检索接口582可实现用于顺序检索，诸如获得下个记录接口(其是请求从迭代器的当前位置按照增加序列号的顺序读取N个记录的接口)。在基于旋转磁盘的存储系统中，如之前提及的，顺序I/O在许多情况下可比随机I/O有效得多，因为在平均每个I/O上所需的磁盘头寻找的数量对于顺序I/O来说可通常比随机I/O低得多。在许多实施方案中，给定分区的数据记录可以序列号顺序写入，并且因此基于序列号顺序的顺序读取请求(例如使用获得下个记录接口或类似接口)可比随机读取请求有效得多。在至少一些实施方案中，因此，不同的计费率可以被设置用于顺序对非顺序的检索接口，例如对于非顺序读取客户端可能被收取更多费用。

[0076] 摄取子系统

[0077] 图6示出根据至少一些实施方案的SMS的摄取子系统204的示例性元件。在描绘的实施方案中，摄取操作被逻辑上划分为前端功能和后端功能，其中前端功能涉及与数据生产商120(例如120A、120B或120C)的交互，并且后端功能涉及与SMS存储子系统的交互。这种前端/后端分离可具有若干优点，诸如加强了存储子系统的安全性并且避免了不得不向数据生产商提供分区策略的细节。SMS客户端库602可提供用于在各种数据生产商120处的安装，并且数据生产商可调用库602中包括的编程接口以提交数据来摄取。例如，在一个实施方案中，数据生产商120可包括在供应商网络的成百上千个物理和/或虚拟服务器处实例化的记录或监测代理。这种代理可以在它们相应的服务器处采集各种日志消息和/或指标，并且定期将采集的消息或指标提交给由SMS的一个或多个摄取控制节点660实例化的前端负载分配器604端点。在一些实施方案中，一个或多个虚拟IP地址(VIP)可建立用于负载分配器，数据生产商可将流数据提交给负载分配器。在一种实现方式中，循环DNS(域名系统)技术可用于VIP，以从若干同等配置的负载分配器中选择特定负载分配器，数据有待由数据生产商120发送至所述负载分配器。

[0078] 在描绘的实施方案中，可将接收的数据记录引导至若干前端节点606(例如606A、606B或606C)中的任何一个。在至少一些实施方案中，负载分配器604可能不了解用于数据记录的分区策略650，并且前端节点606可因此通过使用循环负载均衡(或一些其他通用的负载均衡算法)而不是基于分区的负载均衡而被选择用于给定数据记录。前端节点606可了解用于各种流的分区策略650，并且可与摄取控制节点660交互以获取指定的后端摄取节点608(例如608A、608B或608C)的身份，所述后端摄取节点608被配置用于给定分区的数据记录。因此，在描绘的实施方案中，前端节点604可各自基于数据记录所从属的相应分区来将数据记录传输至多个后端节点606。如之前指出的，数据记录所从属的分区可基于各种因素的任何组合来确定，所述因素诸如由数据生产商供应的分区键、诸如数据生产商的身份或地址的一种或多种其他属性或者数据的内容。

[0079] 后端节点606可各自接收从属于一个或多个流的一个或多个分区的数据记录，并将所述数据记录传输至存储子系统的一个或多个节点。在一些实施方案中，后端节点可以被称为“放置(PUT)服务器”，其中数据被通过HTTP(超文本传输协议)“放置”网络服务API提

交。给定后端节点可确定存储子系统节点集，其数据记录有待通过向控制节点660提交查询而传输至所述存储子系统节点集(在其中用于不同子系统的控制功能是由分开的节点集处理的实施方案中，这转而可对存储子系统的控制节点提交对应的查询)。

[0080] 在至少一些实施方案中，可支持许多不同的摄取确认策略652，诸如至少一次摄取策略或尽力摄取策略。在至少一次策略中，数据生产商120可能需要对每个提交的数据记录的积极确认，并且可重复地提交相同的数据记录(如果未接收到第一次提交的确认)直到最终接收到确认。在尽力摄取策略中，对于提交的至少一些数据记录可能不需要积极确认(尽管摄取子系统可能仍提供偶尔的确认，或者可响应于来自数据生产商的对确认的明确请求)。在其中摄取子系统204需要为数据生产商提供确认的一些实施方案中，在产生确认之前，负责给定数据记录的后端摄取节点608可等待直到在存储子系统处已成功创建数据记录的所需数量的副本(例如，根据建立用于所述流的持久性策略)。在各种实施方案中，序列号可由摄取子系统产生用于接收的每个数据记录，例如指示所述记录被相对于相同的分区或流的其他记录摄取的顺序，并且这种序列号可作为确认或作为确认的部分返回给数据生产商。关于序列号的另外细节在下文中参考图13a和图13b来提供。在一些实现方式中，确认和/或序列号可通过前端节点606传输回数据生产商。在至少一种实现方式中，至少一次策略可在摄取子系统自身的前端节点与后端节点之间实现，例如给定的前端节点606可为适当的后端节点608重复地提交数据记录，直到后端节点提供确认。

[0081] 摄取控制节点660除了其他功能可负责：实例化前端节点和后端节点、监测节点的健康和工作量水平、根据需要协调故障转移、提供对关于哪个节点负责给定分区的查询的响应或者对策略相关的查询的响应，用于来源于流的动态重新分区的摄取相关的配置操作。在一些实施方案中，指定用于一个或多个流的给定集的摄取控制节点的数量自身可随时间而改变，例如一个或多个主控制节点可负责根据需要来重新配置控制节点池。在其中冗余组被建立用于摄取前端节点或后端节点的一些实施方案中，如下文中关于图9和图10进一步详细描述的，控制节点660可负责跟踪哪一个节点是基元并且哪一个是非基元，用于检测针对故障转移的触发条件并且当需要故障转移时用于选择替换物。应注意，在一些实施方案中可以不实现图6中示出的多层摄取子系统架构，例如在一些情境下可以仅配置单组摄取节点。

[0082] 存储子系统

[0083] 图7示出根据至少一些实施方案的SMS的存储子系统的示例性元件。如图所示，摄取节点608(例如在其中前端和后端摄取责任是由不同组节点处理的实施方案中的后端摄取节点)可将流的一个或多个分区的数据记录传输至配置用于那些分区的相应的存储节点702。例如，分区S1-P1的数据记录110A被发送至存储节点702A，分区S2-P3的数据记录110B被发送至存储节点702B和702C，分区S3-P7的数据记录110C被发送至存储节点702D，并且分区S4-P5的数据记录110D被最初发送至存储节点702E。存储控制节点780可负责：实施被应用至不同流的数据记录的持久性策略750、根据需要配置和重新配置存储节点、监测存储节点状态、管理故障转移、响应于存储配置查询或存储策略查询、以及在描绘的实施方案中的各种其他管理任务。

[0084] 在不同的实施方案中，持久性策略750可以以不同的方式不同于彼此。例如，应用至流Sj的持久性策略P1可以在以下方面中不同于应用至流Sk的策略P2：(a)待存储的每个

数据记录的副本的数量;(b)所述副本有待存储到其上的存储装置或系统的类型(例如副本是否待存储到易失性存储器、非易失性高速缓存、基于旋转磁盘的存储装置、固态驱动器(SSD)、各类存储器具、各类RAID(廉价磁盘的冗余阵列)中,是否待存储到数据库管理系统中、在由供应商网络实现的存储服务的节点处等);(c)所述副本的地理分布(例如通过在不同数据中心中放置副本,流数据对于大规模故障或某种类型的灾难是否是可复原的);(d)写入确认协议(例如如果有待存储N个副本,那么在应将确认提供给摄取节点之前必须成功写入所述N个副本中的多少个副本);和/或(e)在有待存储数据记录的多个副本的情况下,所述副本是否应当并行或顺序地进行创建。在有待存储数据记录的多个副本的一些情况下,如在数据记录110D的情况下,给定的存储节点可将数据记录传输至另一个存储节点(例如存储节点702E将用于进一步复制的数据记录110D发送至存储节点702F,并且存储节点702F将其继续发送至存储节点702G)。在使用多副本持久性策略的其他情况下,如在针对其有待存储两个存储器中的副本的数据记录110B的情况下,摄取节点可并行地开始多次复制。在至少一些实施方案中,客户端的选择的持久性策略可以不指定有待用于流数据记录的存储地点的类型;相反,SMS可基于各种标准来选择适当类型的存储技术和/或地点,所述标准诸如成本、性能、到数据源的接近度、耐用性需求等。在一个实施方案中,客户端或SMS可决定使用用于给定流的不同分区或者用于不同流的不同的存储技术或存储地点类型。

[0085] 在图7中示出的实例中,应用至流S1(或者至少流S1的分区S1-P1)的持久性策略是单个副本在存储器中的策略,而为流S2应用的是两个并行副本在存储器中的策略。因此,数据记录110A的存储器中的副本704A在存储节点702A处创建,而对应于数据记录110B的两个存储器中的副本705A和705B在存储节点702B和702C处并行地创建。对于流S3的数据记录110C,创建单个磁盘上的副本706A。对于流S4,可应用顺序的三个副本在磁盘上的策略,并且因此在存储节点702E、702F和702G处顺序地创建相应的磁盘上的副本707A、707B和707C。在不同的实施方案中,可将各种其他类型的持久性策略应用至数据流。检索子系统的节点响应于各种类型的检索API由数据消费者的调用可从适当的存储节点获取数据记录。

[0086] 检索子系统和处理阶段

[0087] 图8示出根据至少一些实施方案的SMS的检索子系统的示例性元件和检索子系统与SPS的交互的实例。如图所示,检索子系统206可包括多个检索节点802,诸如检索节点802A、802B和802C,以及检索控制节点880的集合。检索节点802中的每一个可以被配置来响应于来自各种客户端或数据消费者130的流数据检索请求,诸如如下文所述的SPS的工作节点840。在不同的实施方案中,多种编程检索接口802可由检索节点实现,诸如之前描述的非顺序和顺序的检索接口。在一些实施方案中,诸如HTTP获得(GET)请求的网络服务API可用于数据记录检索,并且检索节点802可因此被称为获得服务器。在描绘的实施方案中,给定的检索节点802可例如由检索控制节点880配置,以从适当组存储子系统节点702(诸如存储节点702A和702B)获取一个或多个流分区的数据记录。

[0088] 在描绘的实施方案中,检索节点802可与一个或多个存储节点702交互,并且还响应于从一个或多个SPS工作节点840接收的检索请求。例如,分区S4-P5的数据记录(例如数据记录110K)和分区S5-P8的数据记录(例如数据记录110L)被由检索节点802A从存储节点702A读取,并且分别被提供给工作节点840A和840K。分区S6-P7的数据记录(诸如110M)被由检索节点802B从存储节点702A读取,并提供给工作节点840K。分区S4-P7的数据记录被由检

索节点802C从存储节点702B读取，并提供给工作节点840B，并且还提供给其他数据消费者130(例如，直接调用SMS检索API而不是通过SPS与SMS交互的数据消费者)。

[0089] 在至少一些实施方案中，检索节点802中的一些或所有可实现相应的高速缓存804(诸如检索节点802A处的高速缓存804A、检索节点802B处的高速缓存804B和检索节点802C处的高速缓存804C)，其中各个分区的数据记录预期到将来的检索请求可暂时地保留。检索控制节点880可负责实现许多检索策略882，包括例如高速缓存策略(例如高速缓存应当被配置多大用于给定分区、数据记录应当被高速缓存多久)、存储节点选择策略(例如在存储多个副本的数据记录的情境中应当最先接触哪一个特定存储节点以获取给定数据记录)等。此外，检索控制节点可负责：实例化和监测检索节点802、响应关于哪些检索节点负责哪些分区的查询、开始或响应于重新分区操作等。

[0090] 在示出的实例中，SPS 290包括两个处理阶段：215A和215B。SPS控制节点885可负责在各个处理阶段215处实例化工作节点804，诸如处理分区S4-P5的记录的工作节点840A、处理分区S4-P7的记录的工作节点840B和处理分区S5-P8和S6-P7的记录的工作节点840K。SPS控制节点885可实现编程接口(诸如图3和图4中示出的那些接口)，以使得SPS客户端能够设计处理工作流程。各种检验点策略850可实现用于不同的处理阶段或工作流程，从而指示工作节点何时或者是否有待存储进度记录，所述进度记录指示所述工作节点在处理它们相应的分区中达到什么程度、有待用于进度记录的存储装置的类型等。故障转移/恢复策略852可指示将导致利用不同的节点来替换工作节点的触发条件或阈值，以及尽力恢复是否有待使用或者基于检验点的恢复是否有待用于给定的处理阶段。在至少一些实施方案中，SPS控制节点885可与各种类型的SMS控制节点交互，例如以识别有待从其获取给定流的数据记录的检索节点、建立对于特定处理工作流程可能需要的新的短暂流或持续流等。在至少一个实施方案中，客户端可与SPS控制节点交互以实例化流，例如一些客户端可能希望仅调用较高水平的SPS接口而不是利用SMS控制接口。应注意，尽管图6、图7和图8中示出分开的控制节点集用于SMS摄取、存储和检索子系统，并且对于SPS阶段，在至少一些实施方案中给定的控制节点可用于若干所述子系统和/或SPS。

[0091] 节点冗余组

[0092] 在至少一些实施方案中，节点的冗余组可以被配置用于SMS的一个或多个子系统。也就是说，代替例如配置一个检索节点用于为流分区Sj-Pk检索数据记录，可建立两个或更多个节点用于这种检索，其中一个节点在给定点处被适时地授予“主要”或积极的角色，而其他一个节点或多个节点被指定为“非主要”节点。当前的主要节点可负责响应工作请求，例如从客户端或者从其他子系统的节点接收的请求。非主要的一个节点或多个节点可保持休止，直到例如由于故障、到主要节点的连接性的丢失或者其他触发条件而触发了故障转移，届时选择的非主要节点可由控制节点通知来接任先前的主要节点的责任。在故障转移期间，主要角色可以因此被从当前现任的主要节点撤回，并被授予当前的非主要节点。在一些实施方案中，当做出将发生故障转移的确定时(例如可能不需要明确的通知)，非主要节点自身可接任为主要节点。在各种实施方案中，这种节点的冗余组可以在SMS处被建立用于摄取、存储、检索和/或控制功能，并且在至少一些实施方案中，在SPS处也可以采取类似的方法用于工作节点。在一些实施方案中，用于给定功能的包括至少一个主要节点和至少一个非主要节点的这种组可以被称为“冗余组”或“复制组”。应注意，存储节点的冗余组可独

立地实现存储的数据记录的物理副本的数量，例如有待存储数据记录的副本的数量可由持久性策略来确定，而被配置用于对应分区的存储节点的数量可基于冗余组策略来确定。

[0093] 图9示出根据至少一些实施方案的可建立用于SMS或SPS的节点的冗余组的实例。在描绘的实施方案中，对于给定流分区Sj-Pk，相应的冗余组(RG)905、915、925和935被建立用于摄取节点、存储节点、检索节点和控制节点。在示出的实施方案中实现了用于控制节点的共用的RG 935，尽管在一些实施方案中可实现用于摄取控制节点、存储控制节点或检索控制节点的分开的RG。每个RG包括主要节点(例如主要摄取节点910A、主要存储节点920A、主要检索节点930A以及主要控制节点940A)和至少一个非主要节点(例如非主要摄取节点910B、非主要存储节点920B、非主要检索节点920C以及非主要检索节点920D)。根据相应的故障转移策略912(用于摄取节点)、922(用于存储节点)、932(用于检索节点)以及942(用于控制节点)，主要角色可以被撤回并授予当前的非主要节点。故障转移策略可以例如管理：将导致主要节点状态变化的触发条件、是否且如何监测主要节点或非主要节点的健康状态、在给定的冗余组中有待配置的非主要节点的数量等。在至少一些实施方案中，可建立单个RG用于多个分区，例如RG 905可负责处理分区Sj-Pk和Sp-Pq的记录的摄取。在一些实现方式中，被指定为用于一个分区的主要节点的节点可同时被指定为用于另一个分区的非主要节点。在一个实施方案中，多个节点可同时被指定为给定的RG内的主要节点，例如给定分区的摄取相关的工作量可在两个主要节点中进行分配，其中一个节点在任何一个主要节点处发生故障的情况下被指定为非主要节点。给定RG中实例化的节点的数量可取决于对应功能(例如在所述组意图能够承受多少并发或重叠的故障上)所需的可用性或复原性水平。在一些实施方案中，除了或者代替被用于SMS节点，冗余组可以被建立用于SPS处理阶段的工作节点。给定RG的构件可有时在地理上分布，例如遍及若干个数据中心，如图10中所示。在一些实施方案中，选择的控制节点可以被配置来例如使用心跳机制或其他健康监测技术检测故障转移触发的条件，并且这种控制节点可通过选择适当的非主要节点作为对故障的主要节点的替换物、通知/启动选择的替换节点等来协调故障转移。

[0094] 在一些实施方案中，供应商网络可以被组织成多个地理区域，并且每个区域可包括本文中也可以被称为“可用性区”的一个或多个可用性容器。可用性容器转而可包括一个或多个不同地点或数据中心，所述可用性容器是以这样一种方式工程化的(例如，通过独立的基础结构部件，诸如功率相关的设备、冷却设备、物理安全部件)：给定的可用性容器中的资源与其他可用性容器中的故障隔离。一个可用性容器中的故障可能不期望在任何其他可用性容器中产生故障，因此，资源实例或控制服务器的可用性配置文件意图独立于在不同的可用性容器中的资源实例或控制服务器的可用性配置文件。可以通过在相应的可用性容器中启动多个应用程序实例或者(在一些SMS和SPS的情况下)将给定的冗余组的节点遍及多个可用性容器分布来在单个地点处保护各种类型的应用程序免受故障。同时，在一些实现方式中，在存在于相同的地理区域内的资源(诸如用于SMS和SPS节点的主机或计算实例)之间可提供廉价且低延迟的网络连接，并且相同的可用性容器的资源之间的网络传输可以甚至更快。一些客户端可能希望例如，在区域水平、可用性容器水平或者数据中心水平处指定保留和/或实例化它们的流管理或流处理资源的地点，以维持它们的应用程序的各种部件精确地在哪里运行的所需程度的控制。其他客户端可能对于保留或实例化它们的资源的精确地点不太感兴趣，只要所述资源例如对于性能、高可用性等满足客户端需求即可。在一

些实施方案中，位于一个可用性容器(或数据中心)中的控制节点可以能够远程地配置其他可用性容器(或其他数据中心)中的其他SMS或SPS节点，也就是说，特定的可用性容器或数据中心可以不需要具有局部控制节点来管理SMS/SPS节点。

[0095] 图10示出根据至少一些实施方案的供应商网络环境，其中给定冗余组的节点可分布在多个数据中心中。在描绘的实施方案中，供应商网络1002包括三个可用性容器1003A、1003B和1003C。每个可用性容器包括一个或多个数据中心的部分或全部，例如可用性容器1003A包括数据中心1005A和1005B，可用性容器1003B包括数据中心1005C，以及可用性容器1003C包括数据中心1005D。示出了SMS和/或SPS节点的许多不同的冗余组1012。一些RG 1012可在单个数据中心内全部实现，如在位于数据中心1005A内的RG 1012A的情况下。其他RG可使用给定可用性容器内的多个数据中心的资源，诸如RG 1012B，所述RG 1012B跨越可用性容器1003A的数据中心1005A和1005B。然而其他RG可使用遍布不同的可用性容器的资源来实现。例如，RG 1012C分别使用位于可用性容器1003A和1003B的数据中心1005B和1005C中的资源，并且RG 1012D分别利用可用性容器1003A、1003B和1003C中的数据中心1005B、1005C和1005D处的资源。在一种示例性部署中，如果RG 1012包括一个主要节点和两个非主要节点，那么这三个节点中的每一个可位于不同的可用性容器中，因此确保至少一个节点非常可能保持其功能性，即使在两个不同的可用性容器处同时发生大规模故障事件。

[0096] 在描绘的实施方案中，分别与SMS和SPS相关联的控制台服务1078和1076可提供易于使用的基于网络的接口，用于配置供应商网络1002中的流相关的设置。可以使用资源在供应商网络1002中实现许多另外的服务(其至少一些可由SMS和/或SPS使用)，所述资源遍布一个或多个数据中心或者遍及一个或多个可用性容器。例如，可实现虚拟计算服务1072，从而使得客户端能够利用选择数量的打包为各种不同能力水平的计算实例的计算能力，并且这种计算实例可用来实现SMS和/或SPS节点。可实现一种或多种存储服务1070，从而使得客户端能够例如通过块装置体积接口或者通过网络服务接口而存储和访问具有所需数据耐久性水平的数据对象。在一些实施方案中，存储对象可附接到服务1072的计算实例或者可从其访问，并且可以被用来在SMS存储子系统处实现各种流持久性策略。在一个实施方案中，诸如高性能键值(key-value)数据库管理服务1074的一个或多个数据库服务，或者相关的数据库服务可在供应商网络1002处实现，并且这种数据库服务可用来通过SMNS存储子系统存储流数据记录，和/或用来存储控制子系统、摄取子系统、存储子系统、检索子系统或处理阶段的元数据。

[0097] 流安全选项

[0098] 在至少一些实施方案中，SMS和/或SPS的用户可以被提供用于数据流的多种安全相关选项，从而使得客户端能够选择资源的安全配置文件(例如虚拟机或物理机)，所述资源有待用于各种功能种类，诸如摄取、存储、检索、处理和/或控制。这种选项可包括例如关于用于各种节点的资源的物理地点的类型的选择(例如，是否有待使用供应商网络设施，或者是否有待使用客户端所有的设施，哪一种设施可具有与供应商网络设施不同的安全特征)、关于流数据的加密的选择和/或在流处理基础结构的各个部分中的网络隔离选择。一些客户端可能担心侵入者或攻击者的可能性，所述侵入者或攻击者获取到有价值的专有商业逻辑或算法的访问，例如并且可能希望使用客户端所有的处所内的计算装置来实现流处

理工作节点。有待用于实现一组SMS和/或SPS节点的资源的类型在本文中可称为用于那些节点的“放置目的地类型”。图11示出根据至少一些实施方案的可以被选择用于SMS或SPS的节点的多个放置目的地类型。

[0099] 在描绘的实施方案中，放置目的地可在供应商网络1102内被选择用于一些类型的SMS/SPS功能种类(例如，摄取、存储、检索、控制或处理)，并且在用于其他类型的SMS/SPS功能种类的供应商网络1102外部。在供应商网络1102内，可使用多租户实例主机1103实现一些资源，诸如计算实例、存储实例或者数据库实例。这种多租户实例主机可在用于一个或多个客户端的所述SMS或SPS节点中的每一个处被实例化，可形成放置目的地类型的第一种类“A”。为了避免不得不与其他客户端共享物理资源，一些客户端可请求它们的SMS/SPS节点使用局限于单个客户端的实例主机来实现。这种单租户实例主机可形成放置种类类型“B”。出于若干原因，从一些客户端的角度来看，单租户实例主机可能是优选的。由于多租户实例主机可包括从属于其他客户端的计算实例，在多租户实例主机中比在单租户实例主机中可能存在来自另一个客户端的实例的安全攻击的更高可能性。此外，当使用单租户实例主机时，也可以避免其中一个客户端的在多租户主机上运行的计算实例CI1经历工作量的激增并开始消耗大比例的主机的计算周期或其他资源，因此潜在地影响到另一个客户端的在不同的计算实例CI2上运行的应用程序的性能的“嘈杂邻居”现象。

[0100] 在描绘的实施方案中，隔离的虚拟网络(IVN)1106(诸如IVN1106A和1106B)可代表放置目的地类型的另一个种类“C”。在一些实施方案中，可应供应商网络客户端的请求创建IVN 1106作为专用网络的逻辑等效物，但是在网络配置正由客户端在很大程度上控制的情况下可使用供应商网络资源构建IVN 1106。例如，客户端可决定IVN1106内有待使用的IP地址，而不需要担心重复可能已在IVN外部使用的IP地址的可能性。在描绘的实施方案中，在一个或多个IVN中实现各种类型的SMS和SPS节点可为客户端的流数据的管理和/或处理增加额外水平的网络安全性。在一些情况下，给定客户端可能希望在一个IVN 1106中放置一个功能种类的SMS/SPS节点，并且在不同的IVN中放置不同功能种类的SMS/SPS节点。在各种实施方案中，给定IVN 1106可包括单租户实例主机、多租户实例主机或者两种类型的实例主机。在一些实施方案中，使用供应商网络的资源的另一组放置目的地类型选择(或安全配置文件选择)(图11中未示出)对于至少一些客户端可以是可用的。在客户端可从供应商网络的用于流相关操作的虚拟化计算服务获得和使用计算资源的实施方案中，所述计算实例可以被用在两种模式中的一种中。在一种模式中，客户端可为SPS或SMS提供一个可执行程序或多个可执行程序，所述可执行程序有待在被配置为SPS工作节点的计算实例处(或者在摄取、存储或检索节点处)运行，并使SMS或SPS运行所述程序并管理节点。这种第一模式可以被称为使用用于流操作的计算实例的“流服务管理的”模式。在其他模式中，客户端可能希望在来自SPS或SMS较少支持的情况下运行可执行程序并管理计算实例。这种第二模式可以被称为使用用于流操作的计算实例的“客户端管理的”模式。这两种操作模式可因此代表关于客户端可选择的放置目的地类型或安全配置文件的另外的选择。如果例如可执行程序有可能需要调试(包括单步调试)，客户端可选择客户端管理的模式，所述调试可由来自客户端的组织的主题专家最好地执行，而流服务管理的模式对于不太可能需要调试的更成熟的代码可以是合理的选择。在一些实施方案中，不同的价格策略可应用至这两种模式。

[0101] 在图11示出的实施方案中，在供应商网络外部的设施处可支持许多放置选项。例

如，SMS库1171和/或SPS库1172被安装在其上的主机1160可以被用于从客户端设施(例如客户端所有的数据中心或处所)1110A或1110B内的流管理或处理，其中两种类型的客户端设施在它们连接到供应商网络的方式上是不同的。客户端设施1110A被通过至少一些共享的互联网链接1151链接到供应商网络1102(即其他实体的网络流量也可在客户端设施1110A与供应商网络1102之间的一些链接上流动)。相比之下，一些客户端设施(诸如1110B)可以被通过特殊的非共享专用物理链接1106(有时可以被称为“直接连接”链接)链接到供应商网络。在图11中使用的术语中，这两种不同类型的客户端处所分别包括放置目的地选项“D”和“E”。在一些实施方案中，SMS和/或SPS的部分在第三方设施(例如使用的但是不由SMS/SPS的客户端所有或管理的数据中心)处也可以是可实现的，并且这种第三方处所可以被指定为放置目的地类型“F”。在至少一些客户端和/或第三方处所中，SMS和/或SPS库可能必须从供应商网络获取，并且安装在有待用于SMS/SPS节点的主机上。在至少一个实施方案中，可以借助于适当的库在供应商网络外部实现所有不同的功能种类的节点。

[0102] 在不同的实施方案中，不同的放置目的地类型在各种安全相关方面可不同于彼此，诸如实现的网络隔离特征、支持的入侵检测功能、实现的物理安全策略、支持的加密级别等。因此，各种目的地类型中的每一种可以被认为具有相应的安全配置文件，所述安全配置文件可以以一种或多种方式不同于其他放置目的地的安全配置文件。如图12a和图12b中所示，在一些实施方案中，SMS和/或SPS的客户端可以以编程方式(例如通过发送到SMS或SPS的一个或多个控制节点的请求)为不同的子系统或节点集选择相应的放置目的地类型。应注意，在一些实施方案中并且对于某些类型的流应用程序，客户端可能希望控制放置目的地类型，这不仅仅是出于安全原因但是还出于性能和/或功能原因。例如，可以通过使用专用客户端处所资源或单租户实例主机来避免上文所述的嘈杂邻居现象。在一些实施方案中，客户端可具有它们希望用于SPS阶段或SMS节点的专用或专有硬件和/或软件，其中使用这种部件可达到的功能能力或性能水平在供应商网络处无法容易地复制，或者只是在供应商网络处不支持。客户端可能已在外部数据中心处访问具有超级计算机水平处理能力的计算机服务器，例如所述计算机服务器可以能够以比单独使用供应商网络资源将可能获得的速率高得多的速率来执行SPS处理。使客户端能够为各种节点选择放置目的地可允许使用这种专用装置或软件。

[0103] 图12a和图12b分别示出根据至少一些实施方案的可以由SPS客户端和SMS客户端提交的安全选项请求的实例。图12a示出SPS安全选项请求1200，其中客户端指示具有标识符1210的处理阶段、请求用于所述阶段的控制节点(元件1212)的放置目的地类型(PDT)以及请求用于工作节点(元件1214)的PDT中的一个或多个。在至少一个实施方案中，客户端还可以能够提交请求以为它们的流数据记录或流处理结果配置加密设置，例如通过请求在那些数据记录在各种网络链接上传输之前使用指定的算法或协议来将它们进行加密，或者请求将各种控制或管理的交互进行加密。例如，在图12a中，用于所述阶段的加密设置可指示有待应用至阶段处理操作的结果的加密技术，和/或用于所述阶段的控制节点与所述阶段的工作节点之间的通信的加密。

[0104] 类似地，在图12b中，客户端的SMS安全选项请求1250包括许多要素，所述要素指示客户端对于具有指定标识符1252的一个或多个流的安全偏好。对于摄取节点、存储节点和检索节点的放置目的地类型偏好可分别在要素1254、1258和1262中进行指示。对于摄取控

制节点、存储控制节点和检索控制节点的PDT偏好可分别由要素1256、1260和1264进行指示。对于数据记录的加密偏好可通过要素1266来指示，例如当数据记录被从一个种类的节点传输至另一个种类的节点时是否和/或如何为其实现加密。通过使用诸如图12a和图12b中示出的那些的安全选项请求，客户端可以能够(例如在供应商网络内或者在供应商网络外部)选择所述地点，以及用于它们的流管理和处理环境的不同部分的各种其他安全配置文件部件。

[0105] 应注意，在至少一些实施方案中，节点放置目的地的选择可能出于除安全以外的其他原因来提供。例如，出于性能原因(例如为了避免之前指出的“嘈杂邻居”问题，而不是主要出于安全原因)，客户端可能希望具有在单租户主机处实现的一些类型的SMS或SPS节点。在至少一些实施方案中，放置选择在流的使用寿命期间可以改变，例如客户端可最初允许SMS节点在多租户实例主机处实例化，但是后来可能希望将所述节点的至少一些子集移至单租户实例主机。在至少一些实施方案中，不同的价格策略可应用至不同的安全相关的选项，例如，在IVN处实现特定功能种类的SMS节点可能比在IVN外部的多租户实例主机处实现特定功能种类的SMS节点花费更高，或者在单租户实例主机处实现SMS节点可能比在多租户实例主机处实现SMS节点花费更高。

[0106] 流记录的顺序存储和检索

[0107] 对于许多类型的流应用程序，数据记录可在SMS处以非常高的速率从多个数据生产商120接收，并且数据消费者可能通常希望以产生所述记录的顺序来访问存储的数据记录。如之前提及的，尤其是在旋转磁盘被用作用于流数据记录的存储装置的环境中，顺序的I/O访问模式(用于读取和写入)可具有优于随机I/O访问模式的显著的性能优势。在若干实施方案中，流指定或分区指定的序列号可当数据记录由SMS接收时来分配给它们，并且可支持基于序列号的顺序检索操作。图13a示出根据至少一些实施方案的在流数据生产商与SMS的摄取子系统之间的示例性交互。流数据生产商可向摄取子系统提交数据记录110，并且在描绘的实施方案中，摄取子系统可回复已被选择用于提交的记录的序列号102。在至少一些实施方案中，摄取节点可获取来自存储子系统的序列号的部分，例如在这种实施方案中序列号102可继接收的数据记录的存储之后根据可应用的持久性策略来确定，并且存储子系统可产生用于所述数据记录的自己的数字的序列指示符，并提供那个指示符用于包括在由摄取节点分配给数据记录的更大的序列号中。

[0108] 序列号可在各种实施方案中实现以提供数据记录的稳定、一致的排序，并且使得能够由数据消费者在记录上进行可重复的迭代。在至少一些实现方式中，分配给特定分区的数据记录的序列号可随时间单调增加，尽管它们不需要是连续的。在各种实施方案中，序列号可以被指定具有以下语义中的至少一些子集：(a)序列号在流内是独一无二的，即没有给定流的两个数据记录可以被分配相同的序列号；(b)序列号可用作到流的数据记录中的索引，并且可用来在给定流分区内的数据记录上进行迭代；(c)对于任何给定数据生产商，数据生产商成功地提交数据记录的顺序反映在分配给数据记录的序列号中；以及(d)用于具有给定分区键值的数据记录的序列号在整个重新动态分区操作上保持单调增加的语义，例如分配给具有分区键值K1的数据记录的序列号在重新分区后可能各自大于分配给在动态重新分区之前具有那个分区键值K1的数据记录的任何序列号。(下文中参考图16进一步详细地描述动态重新分区。

[0109] 在一些实施方案中，数据生产商可能希望影响被选择用于至少一些数据记录的序列号102的选择。例如，数据生产商120可能希望界定流的分配的序列号内的边界或分隔符，以使得对于所述流的数据消费者来说提交针对流的特定子集的读取请求变得更容易。在一些实现方式中，数据生产商120可提交最小序列号连同记录的指示，并且SMS可根据请求的最小值来选择序列号，所述最小值也符合上文讨论的序列号语义。

[0110] 图13b示出根据至少一些实施方案的可以在SMS处被产生用于摄取的数据记录的序列号的示例性要素。在描绘的实施方案中，序列号可包括四个要素：n1位SMS版本号1302、n2位时间戳或纪元值1304、n3位子序列号1306和n4位分区号1308。在一些实现方式中，可使用128位序列号，例如n1、n2、n3和n4可以分别是4位、44位、64位和16位。版本号1302可仅仅用于避免整个SMS软件版本发布上的混乱，例如以使得较为容易地告知SMS软件的哪一种版本被用来产生序列号。在至少一些实现方式中，版本号1302可能不期望频繁地改变。可例如通过摄取子系统节点从本地时钟源或全球可访问的时钟源获取时间戳值1304(例如，实现获得当前纪元(getCurrentEpoch)或获得当前时间(getCurrentTime)API的供应商网络的状态管理系统)。在至少一些实现方式中，从众所周知的时间点的偏移(例如，已从1970年1月1日的00:00:00AM UTC过去的秒数，其可通过在基于Unix™的操作系统中调用各种时间相关的系统呼叫来获取)可用于时间戳值1304。在一些实施方案中，序列号1036可由存储子系统产生，并且可指示特定分区的数据记录写入存储装置的顺序。因此在许多数据记录在给定秒内接收并且时间戳值1304仅在近似一秒间隔处变化的实现方式中，序列号1306可用作用于数据记录的记录到达(或存储)顺序的指示符，所述数据记录刚好已在相同的秒内到达并且因此被分配相同的时间戳值。在一些实施方案中，分区号1308可唯一识别出给定流内的分区。在序列号时间戳(至少近似地)指示摄取对应的数据记录的时钟时间的至少一些实现方式中，序列号可用于索引机制，所述索引机制用于某些类型的基于时间的检索请求。例如，客户端可能希望检索在特定日子上或者在指定时间范围期间产生或摄取的流记录，并且序列号可用作隐含的次级索引的键以检索适当组的数据记录。因此，在至少一些实施方案中，包含用于有序存储和检索的时间戳的序列号的使用可具有另外的好处，那就是提供了到所述组存储的数据记录中的时间索引。

[0111] 通常通过使用大的顺序写操作可通常将给定分区的数据记录以序列号顺序写入(例如到磁盘)。在一些实施方案中，如之前所指出的，可实现基于迭代器的编程接口，以允许数据消费者以序列号顺序读取数据记录。图14示出根据至少一些实施方案的在SMS处的流数据记录的有序存储和检索的实例。分区Sj-Pk(流Sj的第k个分区)的六个数据记录110A-110F被示出以序列号顺序来存储。如图所示，序列号在至少一些实施方案中可以不是连续的，例如由于将值分配给上文所讨论的时间戳部分1304或序列号1306的方式可能不总是产生用于那些要素的连续的值。

[0112] 在图14中示出的实例中，数据消费者通过指定起始序列号“865”已请求产生迭代器。响应于所述请求，SMS已初始化迭代器1，所述迭代器1被定位在具有最近的序列号的数据中心处，所述最近的序列号高于或等于所请求的起始序列号。在这种情况下，由于下一个更低的序列号(860，被分配给数据记录110B)小于消费者的请求中的起始序列号，具有序列号870的数据记录110C已被选择为迭代器的起始位置。获得迭代器接口可考虑在分区内的所请求的位置处设置指针的请求的逻辑等效物，并且获得下个记录接口可以随后被用来从

所述指针位置开始读取数据记录，例如将指针沿着流以序列号顺序移动。在示出的实例中，数据消费者已调用获得下个记录接口，其中参数“迭代器”设置成迭代器1，并且“最大数记录(maxNumRecord)”(返回的数据记录的最大数)设置成3。因此，SMS检索子系统将数据记录110C、110D和110E以那个顺序返回给数据消费者。迭代器(迭代器1)在完成获得下个记录呼叫之后可移动至新的位置，例如到数据记录110F，并且用于相同的迭代器的随后的获得下个记录调用可返回以110F起始的数据记录。在一些实施方案中，获得迭代器呼叫的语义在一些实施方案中可以是不同的，例如迭代器可以被定位在具有等于或低于所请求的序列号的最高序列号的最近的数据记录处，而不是将所述迭代器定位在具有高于或等于指定序列号的最近序列号的数据记录处。在另一个实施方案中，客户端可能必须在获得迭代器呼叫中指定现有的序列号，例如如果具有所请求的序列号的记录在流中不存在，那么可返回错误。

[0113] 分区映射

[0114] 如之前所述，在各种实施方案中，与给定流的记录的摄取、存储、检索和处理相关的工作量可根据各种分区和重新分区策略被划分和分配在若干节点中。图15示出根据至少一些实施方案的流分区映射1501和可以针对SMS和SPS节点做出的对应配置决策的实例。当创建或初始化特定数据流时，例如响应于客户端的创建流API的调用，分区策略可启动用于所述流，所述分区策略可用来确定分区，流的任何给定数据记录将认为是所述分区的成员。摄取子系统204、存储子系统206、检索子系统208和有待针对给定数据记录执行操作的任何相关SPS阶段215的特定节点可在记录的分区的基础上进行选择。在一个实施方案中，用于给定数据记录的控制节点的至少一个子集也可基于分区进行选择。在至少一些实施方案中，可支持数据流的动态重新分区作为分区策略的部分，例如响应于所述策略中所指出的触发条件或者响应于明确的请求。

[0115] 在各种实施方案中，选择用于给定数据记录的分区可依赖于用于所述记录的分区键，所述分区键的值可由数据生产商直接地(例如作为写入或放置请求的参数)或者间接地(例如，SMS可使用元数据作为分区键，所述元数据诸如数据生产商客户端的标识符或名称、数据生产商的IP地址或者数据记录的实际内容的部分)供应。在图15中示出的实施方案中，一个或多个映射函数1506可应用至数据记录分区键或属性1502，以确定数据记录分区标识符1510。在一种实现方式中，例如，给定分区标识符1510可代表在128位整数值的空间上的连续范围，以使得用于流的所有分区的范围的并集可覆盖128位整数能够假定的所有可能的值。在这种示例性情境中，一个简单的映射函数1506可从数据记录的分区键值或所选的属性值产生128位散列值，并且分区标识符可基于散列值正好位于其内的特定连续范围来确定。在一些实现方式中，连续范围可至少最初大小相等；在其他实现方式中，不同的分区可对应于可能与彼此大小不同的连续范围。在一种实现方式中，重新分区还可产生对范围界限的调节。其他分区函数106可用在不同的实现方式中。

[0116] 如果数据流经历动态重新分区(如下文进一步详细讨论的)，那么具有特定键的记录所映射到的分区可以改变。因此，在至少一些实施方案中，SMS和/或SPS控制节点可能必须在流的使用寿命期间记录应用至流的若干不同的映射。在一些实施方案中，诸如时间戳有效范围1511或序列号有效范围的元数据可由用于每个分区映射的控制节点存储。时间戳有效范围1511可例如指示特定的映射M1从流的创建时间直到时间T1来应用，指示不同的映

射M2从T1到T2来应用等。当响应于在流处引导的读取请求时，检索节点可能必须首先确定将使用哪一种映射(例如取决于读取请求中所指示的序列号)，并随后使用那个映射来识别适当的存储节点。

[0117] 在至少一些实施方案中，SMS和SPS控制节点可负责将分区映射至若干不同间隔尺寸处的资源。例如，如图15的示例性实现方式1599中所示，在一种实现方式中，摄取、存储、检索或处理(工作)节点可各自实现为服务器虚拟机内的执行的相应的进程或相应的线程，诸如Java™虚拟机(JVM)或计算实例，并且JVM或计算实例各自可在特定物理主机处实例化。在一些实施方案中，多个JVM可在单个计算实例内启动，从而增加了另一层资源映射决策。因此，对于给定分区，一个或多个控制节点可选择将使用哪一种特定资源作为摄取节点1515、存储节点1520、检索节点1525或处理阶段工作节点1530(例如分别用于阶段PS1或PS2的节点1530A或1530B)。控制节点还可确定那些节点到服务器(诸如摄取服务器1535、存储服务器1540、检索服务器1545或处理服务器1550)的映射，以及服务器与主机(诸如摄取主机1555、存储主机1560、检索主机1565或SPS主机1570A/1570B)之间的映射。在一些实现方式中，分区映射可以被认为包括在示出的各个资源间隔尺寸(例如节点、服务器和主机间隔尺寸)的每一个处的识别信息(例如，资源标识符)、用作一个函数或多个函数1506的输入的数据记录属性的指示、以及函数1506本身。控制服务器可存储元数据存储中的分区映射的代表，并且在一些实施方案中可暴露出各种API(诸如获得分区信息(getPartitionInfo)API)或其他编程接口，以为数据生产商、数据消费者或者为SMS子系统或SPS的节点提供映射信息。

[0118] 数据记录到分区的映射以及从分区到资源的映射在一些实施方案中通过各种因素可能变得更加复杂，所述因素诸如：(a)在一些实施方案中，给定节点、服务器或主机可以被指定负责多个分区，或者(b)在被分配至给定分区或分区集的新的节点、服务器或主机中可能产生故障或其他触发。此外，如在上文中指出并在下文中描述的，用于给定流的分区映射可随时间动态地修改，而流记录继续由SMS和/或SPS节点处理。因此，在一些实施方案中，若干版本的映射元数据可以至少暂时地保留用于给定流，每个版本对应于不同的时段。

[0119] 动态流重新分区

[0120] 图16示出根据至少一些实施方案的动态流重新分区的实例。在图16中示出的时间轴的时间T1处，创建或初始化流S1。分区映射PM1被创建用于流S1，并在时间间隔T1到T2期间保持有效。由SMS在T1与T2之间接收的三个数据记录通过实例示出。数据记录110A(DR 110A)被提交具有客户端供应的分区键值“Alice”，DR 110B被提交具有客户端供应的分区键值“Bill”，并且DR 110C被提交具有客户端供应的分区键值“Charlie”。在初始映射PM1中，所有三个数据记录110A、110B和110C被映射至具有分区标识符“P1”的相同分区。对于P1数据记录，单个节点I1被配置来处理摄取，单个节点S1被配置来处理存储，单个节点R1被配置来处理检索，并且单个工作节点W1被配置来处理SPS处理。用于映射PM1的有效性范围的起始时间戳被设置成T1。

[0121] 在图16的示例性时间轴中，在时间T2处，流S1被动态重新分区。在描绘的实施方案中，数据记录继续到达并由SMS和SPS处理，而不考虑何时发生重新分区；无论是SMS还是SPS都不需要脱机。重新分区可由于许多因素中的任何一种而开始，例如，响应于在摄取、存储、检索或处理节点处的过载状态的检测、响应于在各种子系统的不同主机处的工作量水平之

间的倾斜或不均衡的检测、或者响应于来自数据消费者或数据生产商客户端的请求。在描绘的实施方案中，新的映射PM2在时间T2处(或者在T2不久以后)起作用，如由示出用于PM2的有效性范围起始时间戳设置所指示的。在至少一些实现方式中，不同组数据记录属性除了在重新分区之前使用之外可用于对数据记录分区。在一些情况下，另外的分区属性可(例如在SMS的请求处)由数据生产商提交，而在其他情况下，所述另外的属性可由SMS摄取节点产生。这种另外的属性可称为“加盐”属性，并且使用用于重新分区的另外的属性的技术可称为“加盐”。在一种示例性实现方式中，过载摄取服务器可向数据生产商(例如向正在由数据生产商执行的SMS客户端库代码)指示：对于重新分区，除了先前使用的分区键之外提供了随机选择的较小的整数值。原始分区键和加盐的另外的整数的组合可随后被用来在不同组摄取节点中分配摄取工作量。在一些实施方案中，检索节点和/或数据消费者可能必须被告知关于用于重新分区的另外的属性。在至少一些实现方式中，这种另外的属性可不用于重新分区。

[0122] 在图16中示出的实施方案中，相对于在T2之前被选择用于相同的键的分区，新的分区映射产生被选择用于在T2之后接收的至少一些数据记录的不同分区。DR 110P在T2之后被提交具有分区键值“Alice”，DR 110Q在T2之后被提交具有分区键值“Bill”，并且DR 110R在T2之后被提交具有分区键值“Charlie”。在示出的示例性情境中，通过使用PM2映射，DR 110P被指定分区“P4”的成员，DR 110Q被指定分区“P5”的成员，而DR 110R被指定分区“P6”的成员。在描绘的实施方案中，没有一个示出为在T2之后接收的实例性数据记录被指定为先前使用的分区“P1”的成员，相反，在重新分区后可使用全新的分区。在一些实施方案中，至少一些先前使用的分区在重新分区后可继续使用。对于新的分区P4、P5和P6中的每一个，不同的节点可以被指定用于摄取、存储、检索和/或处理。例如，节点I4、S4、R4和W4可以被配置用于分区P4，节点I5、S5、R5和P5可以被配置用于分区P5，以及节点I6、S6、R6和P6可以被配置用于分区P6。在一些实施方案中，如在重新分区之前用于这种记录，在重新分区之后相同的存储节点可用于具有特定分区键或属性的记录、但是在所述节点内不同的存储地点(例如，在重新分区之后可使用不同的磁盘、不同的磁盘分区或者不同的SSD)。

[0123] 在T2处的动态重新分区之后的至少一些时期期间，检索请求可继续被检索用于数据记录，所述数据记录在重新分区之前由SMS摄取和/或存储子系统处理。在至少一些情况下，请求的数据记录可能必须基于PM1映射来检索，所述PM1映射在摄取数据记录的时候是有效的。因此，如图16中指示的，出于数据检索的目的，PM1和PM2可继续在T2之后的一段时间使用。在至少一些实现方式中，数据记录当它们老化时可最终从流中删除，并且旧的分区映射也可最终弃用，例如当所有对应的数据记录自身已被删除时。在一些实施方案中，代替被删除(或者在删除之前)，流记录可存档(例如基于客户端选择的存档策略)至不同组的存储地点或装置，以使得由SMS使用的分区映射可仍然是可用来在存档之后检索记录。在这种实施方案中，诸如PM1和PM2的分区映射可以保留，只要它们需要支持针对存档存储装置的检索请求。在一些存档的实现方式中，可使用不需要保留流分区映射的不同检索方法(例如新的索引可创建用于存档的数据记录)。在一些实施方案中，诸如P2的在重新分区之前正在使用的但是在重新分区之后写入将不再引导至其的分区可在重新分区之后的某一时刻被“关闭”用于读取，例如响应于检索请求可提供“到达的分区终点”错误消息的等效物。

[0124] 在一些实现方式中，给定数据流可以被划分成许多(例如成百上千)个分区。考虑

一种示例性情况，其中流S1被最初划分成1000个分区，P1、P2、……、P1000。在对应于一个分区的过载状态的情况下，比如P7被删除，这对于改变数据记录到P7的最初映射可能是值得的，但是其他分区的映射不需要改变。在一种方法中，可通过重新分区操作来创建两个新的分区P1001和P1002。在重新分区之后接收的记录可以在重新分区之后被映射至P1001或P1002，因此在两个分区中分配P7的工作量，所述记录的属性将最初(即在原始映射的基础上)在P7中已产生它们的成员关系。例如P1-P6和P8-P1000的剩余的分区可以不需要修改。当仅仅分区的较小子集受到这种重新分区的影响时，在至少一些实施方案中，可产生和存储组合的数据结构，诸如分区条目(或分区条目的树)的有向无环图。每个条目可指示分区函数输出范围，以及有效性时间范围(在那期间条目的分区信息被认为有效的时期)。在上文的实例中，假设涉及P7的重新分区在时间T2处执行，而流S1(及其初始映射)在时间T1处创建。在这种情境下，用于关于P7的条目的有效性时期将是“T1至T2”，用于P1001和P1002的有效性时期将是“T2向前”，并且用于剩余分区的有效性时期将是“T1向前”。在至少一些实现方式中，使用这种组合的数据结构可引起用于分区映射元数据的存储器或存储装置的数量的实质性减少。在上文的实例中，讨论了将分区P7分离成两个新的分区。在至少一些实现方式中，分区在重新分区期间还可以被合并，例如为其接收相对少的检索请求或者提交相对少的记录的两个邻近的分区可合并成单个分区。对于任何时间点，可使用分区函数和有效性时间范围信息来明白地确定数据记录所从属的分区。随着时间的变化，组合的数据结构可发展为更多的分离部分和/或执行合并，但是对于分区元数据所需的总的空间可能(当然取决于分离多久一次发生以及平均有多少分区受到分离的影响)不明显地增加。相比之下，在不同的实现方式中，每当发生重新分区，用于流的整组不变的元数据可以被复制并且与用于受到重新分区的影响的分区的条目进行组合。在后面一种实现方式中，对于分区映射元数据的存储装置和存储器的需求可能以快得多的速率增加，尤其是如果旧的映射如上文所述在重新分区之后可能必须被保留至少一段时间。

[0125] 在使用包括时间戳值(诸如图13b中示出的时间戳值1304)的序列号的至少一些实施方案中，指定类型的序列号转变可以被实现用于动态重新分区。通过实例假设类似于图13b中示出的方案的基于时间戳的序列号方案正用于流S1，其中每秒产生新的时间戳值来包括在所述序列号中。在支持动态重新分区的至少一些实现方式中，除了在动态重新分区之前使用之外，在动态重新分区之后分配的序列号可全部使用不同组的时间戳值(以对应于重新分区事件的所选的初始时间戳值开始)。例如，如果在递交(即使其生效)动态重新分区的时间处所使用的时间戳值是Tk，那么在所述递交之后所发出的任何新的序列号可能需要向前来使用时间戳值Tk+1。由于序列号值在图13b中所使用的方案中在时间戳值的高阶位的至少一些中对它们进行编码，确保重新分区事件对应于如前所述的时间戳界限可转而简化在响应于检索请求来识别有待使用的映射中所涉及的簿记。因此，在这种实现方式中，当接收到指定特定的序列号的检索请求时，可从所述序列号中提取时间戳值，并且可容易地确定是否应当使用重新分区后的映射，或者是否应当使用重新分区前的映射。如果提取的时间戳值低于被选择用于重新分区的初始时间戳，那么可使用重新分区前的映射，并且如果提取的时间戳值等于或高于被选择用于重新分区的初始时间戳值，那么可使用重新分区后的映射。

[0126] 用于流管理和处理的方法

[0127] 图17是根据至少一些实施方案的示出可执行来支持用于数据记录摄取和数据记录检索的相应组编程接口的操作方面的流程图。如要素1701中所示，可例如从SMS客户端或数据生产商客户端接收对于创建或初始化数据流的请求。可确定(要素1704)有待用于所述流的初始的分区映射，例如可基于分区策略来识别有待用于识别特定数据记录所从属的分区的函数和有待用于所述函数的输入参数。如之前提及的，在各种实施方案中，SMS的控制部件可负责接收和响应流创建请求。实现流创建和初始化(以及其他控制平面操作)的方式从一个实施方案到另一个实施方案可以不同。在一个实施方案中，例如，可建立控制服务器的冗余组，并且所述冗余组的主要控制服务器可以通过在持久性存储地点中产生和存储用于新的流的适当的元数据(例如，初始的分区映射，初始的摄取、存储和检索节点集等)来响应于流创建请求。可以通过使用存储的元数据的主要控制服务器来产生对随后的关于所述流的查询(例如来自前端摄取节点的关于负责给定分区的后端节点的请求)的响应。在SMS控制平面功能的另一种实现方式中，可将流配置元数据存储在数据库中，所述数据库由摄取、存储或检索子系统的至少一些节点是直接可访问的。在已创建和初始化流之后，通常在没有与控制部件的另外的交互的情况下，可开始诸如记录提交、存储和检索的数据平面操作，并且可由对应子系统的相应部件处理所述数据平面操作。

[0128] 在一些实施方案中，数据生产商可能需要提交具有写入请求的明确的分区键，而在其他实施方案中，可基于与写入请求相关联的元数据(诸如数据生产商的身份、从其接收数据记录的IP地址)来确定有待用于分区函数的输入，或者从数据记录本身的内容来确定有待用于分区函数的输入。在至少一种实现方式中，客户端可任选地在数据记录提交中供应分区标识符，并且在这种实现方式中，可能不需要另外的分区函数。

[0129] 当为所述流确定或配置用于摄取、存储和检索函数的初始的节点集(要素1707)时，可以考虑许多不同的因素。例如，分区映射自身(其可确定将所述流划分成多少个分区，以及所述分区的相对预期的大小)、关于预期的摄取速率和/或检索速率的信息(如果这种信息有用)、对于流数据记录的耐久性/持久性需求和/或对于各种子系统的高可用性需求(其可引起类似于图9和图10中示出的那些的冗余组的建立)可影响不同子系统的节点的数量和放置。此外，在客户端可指示针对各种种类的节点(如图11、图12a和图12b中所示)的放置目的地类型偏好的实施方案中，这种偏好还可在确定有待用于SMS和/或SPS节点的资源中起作用。在至少一些实施方案中，可提前建立能够执行摄取、存储和/或检索功能的节点的相应的池，并且控制部件可将这种池的选择成员分配给创建的每个新的流。在其他实施方案中，至少在一些情况下，当创建或初始化流时，可能必须实例化新的摄取、存储或检索节点。

[0130] 在描绘的实施方案中的摄取节点处，可通过被实现用于数据记录提交的任何一组编程接口来接收记录(要素1710)，包括例如在线提交接口(其中，将数据包括在提交请求中)和参考提交接口(其中，在提交请求中提供地址，可由SMS摄取节点或SMS存储节点例如使用网络服务请求或其他接口从所述提交请求检索数据)。可在不同的实施方案中针对提交记录的方法中的每一种来提供任何数量的不同类型的编程接口，例如可支持相应的应用程序编程接口(API)用于在线对参考提交，可建立网页或网址，可实现图形用户界面或者可开发命令行工具。在至少一些实施方案中，SMS可为每个摄取的记录分配序列号，例如指示摄取或存储记录的顺序，并且序列号可以被用于通过数据消费者的检索请求。在检索子系

统节点处，可通过任何一组实现的编程检索接口来接收记录检索请求，并且作为响应可提供请求的数据记录的内容(要素1713)。对于非顺序访问，接口可包括例如获得迭代器(基于获得迭代器调用中所指示的序列号来请求在分区内选择的位置处有待实例化的迭代器)或者获得具有序列号的记录(getRecordWithSequenceNumber)(以获取具有指定序列号的数据记录)。对于顺序访问，可实现诸如获得下个记录的接口(从迭代器的当前位置开始或者从指定的序列号开始按顺序请求多个记录)。在至少一些实施方案中，不同的检索接口可具有与它们相关联的不同的计费率，例如可将用于顺序检索的按记录的计费率设置成低于用于非顺序检索的每个记录的计费率。在一些实施方案中，不同的提交接口也可具有不同的计费率，例如参考提交可比在线提交每个记录花费更高。

[0131] 随着时间的变化，控制节点或专用的计费服务器可采集用于在流管理服务的各个子系统处实现的不同编程接口的使用指标(要素1716)。所述指标可包括例如：不同编程接口的调用计数、摄取或检索的记录的总数(其可不同于用于至少一些接口的调用计数，所述接口诸如可以被用来通过单个调用来检索多个记录的获得下个记录接口)、摄取或检索的数据的总数等。可任选地至少部分基于使用指标和与编程接口相关联的相应的计费率来产生有待向拥有所述流的客户端或者生产和/或消费来自流的数据的客户端收取的计费额(要素1719)。在至少一些实施方案中，计费活动相对于流摄取/检索操作可以是异步的，例如可基于在这个月期间所采集到的指标来在每月计费期的结束时产生计费。

[0132] 图18a是根据至少一些实施方案的示出可执行来配置流处理(SPS)阶段的操作方面的流程图。如要素1801中所示，可实现编程接口以使得客户端能够为流数据记录配置许多处理阶段。为了配置特定阶段，例如客户端可指示在所述阶段有待在分区的流数据记录上执行的处理操作、用于处理操作的输出的分配策略、以及其他参数，诸如待处理的数据将从其获取的输入流的身份。在一些实施方案中，SPS阶段的处理操作可能需要是幂等的。在其他实施方案中，对于至少一些阶段也可支持非幂等操作。在一些实施方案中，如果在给定阶段处待执行的处理是非幂等的，通过配置工作节点来定期地为一些持久性外部地点清除操作的输出从而记录下清除操作何时相对于记录检索顺序来执行，并且随后在恢复期间配置替换工作节点来重演所述清除操作，客户端可仍然能够获取幂等性的恢复相关的益处。在至少一些实施方案中，客户端可以能够利用流数据上并行操作的若干不同的状态和被用作用于其他阶段的输入流的一些阶段的结果来配置有向无环图(DAG)或其他处理阶段的图。在一些实施方案中，可在不同阶段之间创建一个或多个短暂流而不是持续流，例如不必须将来自一个阶段的数据记录输出在将其作为输入馈送至不同阶段之前存储在持久性存储装置上。

[0133] 在一些实施方案中，可实现任何数量的不同的恢复策略用于SPS阶段，包括例如基于检验点的恢复策略或尽力恢复策略。在一个实施方案中，客户端可使用编程接口来选择用于不同SPS阶段的恢复策略。在使用基于检验点的恢复策略的阶段处，可将工作节点配置来间断地存储进度记录或检验点，从而指示在它们已到达的流分区中达到什么程度(例如，可将最近处理的记录的序列号存储为所述进度的指示符)。如下文参考图19所描述的，在故障后，可随后在恢复操作期间使用进度记录。在尽力恢复策略中，不需要存储进度记录，并且响应于故障而配置的替换工作节点当接收到新的数据记录时可对其进行简单处理。在给定SPS阶段图或工作流程内，在一些实施方案中，可将不同的恢复策略应用至不同的阶段。

[0134] SPS控制服务器可例如通过要素1801中所指示的编程接口中的一个来接收幂等操作Op1的指示，所述幂等操作Op1根据分区策略PPol1有待在流S1的特定阶段PS1处执行，其中所述处理的结果有待根据输出分配描述符DDesc1来进行分配(要素1804)。可以例如基于各种因素来确定有待配置用于阶段PS1的工作节点的数量和对于节点所需的虚拟或物理资源(要素1807)，所述因素诸如Ppol1、幂等操作Op1的复杂性以及有待用于工作节点的资源的性能能力。

[0135] 可随后实例化和配置工作节点(要素1810)，例如作为所选的虚拟或物理机资源处的进程或线程。在一种简单的实现方式中，例如，可最初分配一个工作节点用于S1的每个分区。可配置给定的工作节点来：(a)从S1的检索节点的适当的子集接收数据记录，(b)在接收的数据记录上执行Op1，(c)任选地，例如基于用于PS1的恢复策略来存储指示已处理哪一组分区记录的进度记录/检验点，以及(d)将输出传输至由DDesc1指示的目的地(例如作为到中间持续流或短暂流或者直接到其他处理阶段或存储系统的输入)。应注意，至少在一些实施方案中，SPS处理可能不一定产生在前进基础上在别处已传输的任何输出。例如，一些SPS应用程序可简单地用作数据记录的临时资源库，和/或可以实现使得用户能够查看数据记录的查询接口。这种应用程序可管理其自身的输出，例如可响应于接收的查询并且不根据分配描述符来产生输出。记录相关的SPS应用程序可保持从大规模分布式系统所采集的最后一天的日志记录，例如使得客户端能够出于调试或分析的目的来查看记录数据。因此，在一些实施方案中，不需要将输出分配描述符指定用于SPS的至少一些阶段、用于至少一些流或者用于至少一些分区。工作节点可随后按照它们相应的配置设置开始检索和处理数据记录(要素1813)。在至少一些实施方案中，SPS控制节点可(例如使用诸如心跳协议的响应性检查)监测工作节点的健康状态，以及各种其他指标，诸如在被用于工作节点的资源处的资源利用水平(要素1816)。例如如果应当替换工作节点并如下文所述地实现恢复策略，那么从工作节点采集的信息可用来确定是否需要故障转移。

[0136] 在一些实施方案中，可安装的SPS客户端库可提供给希望在客户端所有的处所处和/或在供应商网络的客户端选择的资源处实现SPS工作节点的那些客户端。客户端库还可允许SPS客户端选择它们希望使用SPS管理的服务的各种控制平面特征的程度，诸如健康监测功能、自动工作量监测和均衡、安全管理、动态重新分区等。图18b是根据至少一些实施方案的示出响应于用于流处理工作节点的配置的客户端库的部件调用可执行的操作方面的流程图。如要素1851中所示，可提供SPS客户端库(例如通过从可配置来执行图18a中示出的各种操作的多租户SPS管理的服务的网址下载)。所述库可包括许多可执行部件和/或可链接到客户端应用程序的部件。一些库部件可使得客户端能够选择SPS管理服务、向SPS管理服务注册、或指定各种工作节点的所需特性，在所述工作节点处有待执行一个或多个SPS阶段的流处理操作。例如，一个客户端可能希望使用在用于工作节点的供应商网络的虚拟计算服务处实现的它们自己的计算实例集，而另一个客户端可能希望使用位于客户端自己的用于处理流记录的数据中心处的计算装置(诸如不由供应商网络支持的专用装置)。客户端可以在如果需要的基础上在它们自己的处所处在线地或者如果需要来使用虚拟计算服务的计算实例带来工作节点。除了或者代替工作节点的这种按需的实例化，在一些实施方案中，客户端可预先配置当需要时可部署的潜在可重用的工作节点池。在一些实现方式中，可执行或调用库部件以允许客户端向SPS管理的服务注册，可通过SPS管理服务来处理由客户

端实例化为指定阶段的工作节点的后续控制平面操作用于其的特定进程或线程。在一个实施方案中，客户端还可以能够从有待由用于工作节点的SPS管理服务处理的不同级别的控制平面责任中进行选择，例如，一个客户端可能希望使用它们自己的定制模块来监测工作节点健康，而另一个客户端可能希望利用SPS管理服务来用于监测工作节点健康并且如果检测到故障而采取适当行动。

[0137] SPS管理服务可接收特定客户端希望使用客户端库的指示(要素1854)，所述客户端库用于配置特定SPS阶段PS1的工作节点和/或控制平面操作。(PS1自身可使用所述库中所包括的编程接口来设计，或者使用由SPS管理的服务所暴露的类似于图4中示出的基于网络的接口的编程接口来设计。)客户端也可指示流，所述流的数据有待检索用于用作通过PS1的输入。任选地，在至少一些实施方案中，客户端可指示针对PS1的控制平面设置，例如客户端是否希望使用对于节点的服务的健康监测能力，或者是否愿意使用定制的健康监测工具(要素1857)。取决于由客户端所指示的偏好，可确定有待配置用于客户端的使用的SMS和/或SPS的一个或多个节点(要素1860)。可在客户端的工作节点之间建立到SMS/SPS节点的网络连接性，和/或可执行其他配置操作以使得能够如所需要地得到数据记录流和处理结果。当接收到检索请求时，可将数据记录提供给SP1工作节点，并且可根据需要执行所需的控制平面操作(若有由客户端请求的话)。应注意，至少在一些实施方案中，也可以或者替代地实现类似的方法，所述方法使得客户端能够控制它们希望使用SMS管理服务的各种子系统的控制平面功能的程度。

[0138] 图19是根据至少一些实施方案的示出可执行来实现用于流处理的一种或多种恢复策略的操作方面的流程图。如要素1901中所示，SPS控制节点可确定已符合用于替换特定工作节点的触发标准，例如工作节点可能已变成无响应或者不健康的，当前节点的工作量水平可能已达到故障转移的阈值，在工作节点处检测到的错误的数量可能已超过阈值，或者可识别工作节点的一些其他意想不到的状态。可识别或实例化替换的工作节点(要素1904)。在一些实施方案中，可建立可用工作线程的池，从其中可选择一个工作线程作为替换物，例如或者可启动新的线程或进程。

[0139] 如果在SPS阶段(在所述SPS阶段处特定工作节点是有效的)处有待使用尽力恢复策略(如要素1907中所确定的)，那么替换的工作节点可仅仅当另外的数据记录变得可用时开始对它们进行处理(要素1916)，例如没有替换的工作节点的进度的记录需要检查。如果待使用基于检验点的恢复策略，那么可提供地点的指示(例如存储装置地址或URL)(要素1910)，在所述地点处替换工作节点可访问由替换的工作节点所存储的进度记录。替换工作节点可检索由替换的节点所存储的最近的进度记录，并使用所述进度记录来确定所述组数据记录(要素1913)，在所述组数据记录上，替换工作节点应当执行所述阶段的幂等操作。在这种基于检验点的恢复策略中，取决于在最后的进度记录与实例化替换工作节点的时间之间的持续时间，以及取决于替换的工作节点继正在存储的进度记录之后已处理另外的记录的速率，可不止一次处理一些数量的数据记录。在至少一些实施方案中，如果正在执行的操作是幂等的，那么这种重复操作可以不具有负面效应。在替换工作节点已基于之前存储的进度记录执行重复恢复操作之后，在至少一些实施方案中，替换工作节点可存储其自身的指示完成恢复的进度记录，并且可在最新接收的数据记录上开始正常的工作线程操作(要素1916)。

[0140] 图20是根据至少一些实施方案的示出可执行来实现用于数据流的多种安全选项的操作方面的流程图。如要素2001中所示，可实现一个或多个编程接口，所述一个或多个编程接口使得客户端能够从针对数据流管理和处理的各种安全选项中进行选择，所述选项包括例如针对不同功能种类的节点(例如，摄取、存储、检索、处理或控制节点)的放置目的地类型选项。放置目的地类型在它们的安全配置文件的各个方面中可不同于彼此。在一些实施方案中，有待用于SMS或SPS节点的资源的物理地点可从一种目的地类型到另一种目的地类型而不同。例如，诸如位于供应商网络数据中心处的实例主机的资源可用于所述节点，或者可使用在客户端所有的设施处的资源，或者可使用第三方资源。在至少一些实施方案中，网络隔离级别或其他网络特征可从一种目的地类型到另一种目的地类型而不同，例如可在隔离的虚拟网络内实例化一些SMS或SPS节点，或者在客户端所有的设施处通过专用的隔离物理链接将一些SMS或SPS节点连接到供应商网络。在一个实施方案中，客户端可指示在供应商网络的单租户实例主机处有待建立某些类型的SMS或SPS节点，而不是使用也可以是可用的多租户实例主机来建立。在至少一些实施方案中，各种类型的加密选项通过安全相关的编程接口也可以是可选择的。

[0141] 可通过安全相关的编程接口来接收客户端的关于用于流S1的一个或多个功能种类的节点的安全配置文件选择或偏好。例如，客户端可选择用于功能种类FC1的节点的一个安全配置文件(例如客户端可能希望在客户端所有的处所处实现SPS工作节点)和用于不同功能种类FC2的节点的不同的安全配置文件(例如，客户端可能愿意在供应商网络数据中心处实现SMS摄取节点或存储节点)(要素2004)。在一些情况下，客户端可决定建立具有相同的安全配置文件的所有不同功能种类的节点。在一些实施方案中，SMS和/或SPS可限定针对各种功能种类的默认的放置目的地类型，例如，除非客户端另外指示在供应商网络的隔离的虚拟网络内可建立所有功能种类的节点。

[0142] 可随后基于客户端对于安全配置文件和/或地点的偏好(或者基于对于客户端不对其提供偏好的功能种类的默认设置)来配置不同功能种类的节点(要素2007)。所述配置可涉及例如选择适当的物理主机或物理机，并实例化用于不同功能种类的节点的适当的计算实例、虚拟机、进程和/或线程，并在节点之间建立适当的网络连接。在一些实施方案中，可提供用于不同的流管理和处理功能的可执行的库部件(作为所述配置的部分)用于在供应商网络外部的主机处安装。

[0143] 根据至少一些实施方案，可以例如根据客户端所表达的加密偏好或者基于默认的加密设置而在一个或多个种类的节点处启动加密模块(要素2010)。可以随后启动各种功能种类的节点，以使得如由客户端所期望地摄取、存储、检索和/或处理流数据(要素2013)。

[0144] 图21是根据至少一些实施方案的示出可执行来实现用于数据流的分区策略的操作方面的流程图。如要素2101中所示，可为数据流确定分区策略。所述策略可包括例如数据记录到分区的初始映射，所述初始映射基于由数据生产商供应的键或者基于提交的数据记录的各种属性以及对于重新分区数据流的一种或多种触发标准。在一些实施方案中，例如，可将散列函数应用至一个分区键或多个分区键，从而产生128位整数的散列值。可将可能的128位整数的范围划分成N个连续的子区间，每个子区间代表流的N个分区中的一个。在一些实施方案中，分区的数量和/或子区间的相对大小可从一个流到另一个流而变化。在至少一些实施方案中，为其利益而配置流的客户端可提供关于有待使用的分区方案的输入，例如

所需分区的数量或者有待使用的分区函数的所需特征。在至少一个实施方案中，客户端可为提交的数据记录的一些子集或者全部提交的数据记录提供分区标识符或名称。

[0145] 当接收流的数据记录时，可基于供应的键和/或其他属性确定它们相应的分区，并且可选择适当组的摄取、存储和检索节点用于识别的分区(要素2104)。在至少一些实施方案中，可为数据记录产生相应的序列号，例如指示接收给定分区的记录的顺序(要素2107)。在一些实现方式中，序列号可包括许多要素，诸如时间戳值(例如，从诸如1970年1月1日00:00:00UTC的已知的纪元过去的秒数)、从存储子系统获取的子序列值、SMS软件的版本号和/或分区标识符。在一些实施方案中，可将序列号提供给数据生产商，例如以确认提交的数据记录的成功摄取。在一些实施方案中，序列号还可由数据消费者使用，来以摄取顺序检索流或分区的数据记录。

[0146] 在至少一些实施方案中，可以序列号顺序将数据记录存储在存储节点处，基于分区策略将数据记录引导至所述存储节点(要素2110)。在使用旋转磁盘存储装置的实施方案中，可通常使用顺序写入来将接收的数据记录保存至磁盘，从而避免磁盘寻道等待时间。在至少一些实现方式中，在将记录存储至磁盘之前可将非易失性缓存用作写入高速缓存，例如以进一步减少磁盘寻道的可能性。响应于对读取按照序列号排序的多个数据记录的请求(例如，获得下个记录或类似接口的调用)，可随后使用顺序读取从存储装置读取数据记录(要素2113)。

[0147] 图22是根据至少一些实施方案的示出可执行来实现数据流的动态重新分区的操作方面的流程图。如要素2201中所示，(例如，在SMS或SPS的控制部件处)可做出流有待进行动态重新分区的确定。许多不同的触发条件可产生重新分区流的决定，诸如在摄取、存储、检索、处理或控制节点的一个或多个处的过载的检测，或者在不同节点的工作量水平上的不均衡的检测，或者可从客户端(例如数据生产商或数据消费者)接收的重新分区的请求。在一些实现方式中，客户端重新分区的请求可包括请求的重新分区的具体细节，诸如有待产生的修改的映射的各种参数(例如，有待添加或去除的分区的数量等，应当组合或分离所述指定分区)。在一种实现方式中，客户端重新分区请求可指示客户端希望解决的问题状态(诸如负载不均衡)，并且SMS或SPS可负责将问题状态的描述转化成适当的重新分区操作。在一些情况下，代替请求重新分区或描述问题状态，客户端可指定有待用于重新分区的触发标准。在一些实施方案中，数据流的数据持久性需求的改变的确定可触发重新分区，这可例如产生用于流记录的不同组存储装置或不同的存储技术的选择。在一些情况下，数据流的使用模式(例如，生产或消费数据记录的速率)的改变的检测也可引起重新分区，并且还可引起更适合用于改变的使用模式的不同的存储技术或不同组存储装置的使用。例如，重新分区的决定可基于对期望用于给定分区或全部流的读取和写入的速率的确定，SSD可以是比旋转磁盘更适合的存储技术。在一个实施方案中，安排的或即将产生的软件和/或硬件版本改变可触发重新分区。在一些情况下，当客户端指示出通过使用不同的分区方法或者不同的存储方法可以更有效地符合的预算限制时，定价或计费问题可触发重新分区。在至少一些实施方案中，改变的性能目标也可触发重新分区。在图22中描绘的实施方案中，可选择有待用于在重新分区后分配的序列号的初始的时间戳值(诸如从1970年1月1日00:00:00UTC的秒的偏移、通过若干操作系统中的系统呼叫通常可获得的纪元值)(要素2204)。在一些实现方式中，在供应商网络处实现的全局状态管理器可支持获得纪元值

(getEpochValue)API，例如从而使得SMS和/或SPS的各种部件能够获取有待用于序列号产生的一致的时间戳值。在其他实现方式中，可使用其他时间源，例如可指定SMS或SPS控制节点来向其他部件提供一致地排序的时间戳值，或者可使用局部系统呼叫调用。在一些实施方案中，时间戳值在任何特定主机处不必须对应于挂钟时间，例如可简单地使用单调增加的整数计数器值。

[0148] 可为所述流产生不同于在重新分区决定时使用的映射的修改的分区映射(要素2207)。在至少一些实施方案中，在重新分区之前，改变的映射可将具有特定分区键的数据记录映射至与将具有相同键的数据记录映射至的分区不同的分区。可取决于对于重新分区的触发条件和/或取决于遵守的工作量指标来分离一些分区(通常大量使用的分区)，而可合并其他(通常少量使用的)分区。在一些实施方案中，在重新分区后可使用与重新分区之前不同的分区函数，例如不同的散列函数，或者可使用将散列函数结果划分成分区的不同的方法。在例如分区对应于128位整数的连续范围的一些实现方式中，在重新分区后可将128位整数空间划分成不同组的子区间。在至少一些实施方案中，可将摄取、存储、检索、处理或控制节点的新的组分配给新创建的分区。在一些实现方式中，可将空间有效组合的数据结构用来代表初始映射和修改的映射(要素2208)。例如，可存储有向无环图或树形结构，其中每个条目包含分区函数输出范围(例如，对应于给定分区的分区散列函数的结果的范围)和有效性时间范围的指示，以使得由于重新分区仅仅需要改变对应于修改的分区的记录。在重新分区期间保持不变的用于分区的条目可以不需要在数据结构中进行修改。可配置新的节点来实现修改的分区映射(要素2210)。在至少一些实施方案中，由于在至少一段时间内可继续接收对在之前映射基础上存储的数据记录的检索请求，可保留先前的节点和先前的映射一段时间。当接收到指定特定的序列号或时间戳的读取请求时(要素2213)，可以(例如，在控制节点处或者在检索节点处)做出关于读取请求是否有待通过使用新的分区映射或先前的分区映射来满足的确定。可随后使用所选的映射来识别有待从其获取请求的数据的适当的存储节点。

[0149] 图23是根据至少一些实施方案的示出可执行来实现用于数据流记录的至少一次记录摄取策略的操作方面的流程图。如要素2301中所示，可实现一个或多个编程接口以使得客户端能够从若干摄取策略选项中选择用于数据流的记录摄取策略，所述记录摄取策略选项包括例如(a)至少一次策略，根据所述至少一次策略，记录提交者将一次或多次地提交记录直到接收到肯定的确认，或者(b)尽力摄取策略，根据所述尽力摄取策略，不对至少一些记录提交提供确认。一些数据生产客户端可能不像其他的数据生产客户端那样担心它们的记录的一小部分的潜在丢失，并且可能因此选择尽力摄取方法。在一些实现方式中，即使对于配置用于尽力摄取的流，SMS可仍然提供对于数据记录的一些子集的确认，或者可能甚至尝试提供对于所有数据记录的确认，即使尽力策略不需要对于每个数据记录的确认。

[0150] 可通过一个编程接口接收请求，所述请求指示有待用于指定流的特定摄取策略(要素2304)。可根据对所述流有效的摄取策略来实例化摄取节点(要素2307)。当在摄取节点处接收到相同数据记录的一个或多个提交时(要素2310)，可取决于有效的摄取策略来采取不同的动作。如果正在使用至少一次摄取策略(如要素2313中所确定的)，那么可针对一个或多个提交中的每一个将确认发送给数据生产商，但是可在存储子系统处仅保存一次数据记录(2316)。(应注意，根据对于流有效的持久性策略，在一些情况下可存储给定记录的N

个副本，但是如果提交给定数据记录M次，可能仅仅为一个提交产生副本，即存储的记录副本的总数将仍然是N，而非NxM。)如果尽力摄取策略有效(如要素2313中还检测到的)，可在存储装置处仍然保存一次数据记录，但是不需要将确认发送给数据生产商(要素2319)。在至少一些实施方案中，可至少部分基于所选的摄取策略来任选地确定客户端计费额(要素2322)。如之前指出的，在一些实施方案中，可支持两个版本的至少一次摄取策略。在一个版本中，类似于图23中示出的那个版本，SMS可负责去重复数据记录(即，确保仅响应于一组两个或更多个提交中的一个而将数据存储在SMS存储子系统处)。在不同版本的至少一次摄取中，可允许通过SMS的数据记录的重复。后一种方法对于流应用程序可能是有用的，其中存在很少或者没有数据记录重复的负面后果，和/或对于执行它们自己的重复消除的流应用程序可能是有用的。

[0151] 图24是根据至少一些实施方案的示出可执行来实现用于数据流的多种持久性策略的操作方面的流程图。如要素2401中所示，可实现使得客户端能够从多个持久性策略中选择用于流数据记录的持久性策略的一个或多个编程接口。持久性策略可在各个方面的任何一个上不同于彼此：例如(a)待保存的副本的数量可以不同(例如，可支持N个副本对2个副本对单个副本策略)，(b)待使用的存储地点/装置类型可以不同(例如，旋转磁盘对SSD对RAM对数据库服务或多租户存储装置)和/或(c)所述策略在对大规模故障的恢复性的预期程度上可以不同(例如，可支持多数据中心对单数据中心策略)。可接收请求，所述请求指示客户端的用于指定流的特定持久性策略的选择(要素2404)。在一些实施方案中，由客户端选择的持久性策略可引起用于给定流的相应分区的不同存储地点类型或装置类型的使用。在一个实施方案中，SMS(而不是客户端)可在流等级上或者在分区等级上选择存储地点类型或装置类型。在一些实施方案中，当选择在一些实施方案中的持久性策略时，客户端可指示数据耐久性目标和/或性能目标(诸如所需的读取或写入生产量或延迟)，并且这些目标可由SMS使用来选择适当的存储装置类型或地点。例如，如果需要较少的延迟，那么可使用SSD代替旋转磁盘来存储一个或多个分区或流的数据记录。

[0152] 可确定或配置一组摄取节点以从数据生产商接收所选的流的数据记录，并且可配置一组存储节点以实现所选的持久性策略(要素2407)。当在摄取节点处接收到数据记录时(要素2410)，可基于所选的持久性策略在所选的存储装置处通过存储节点存储数据记录的一个或多个副本，所述存储节点负责数据记录所从属的分区(要素2413)。在至少一些实现方式中，可基于由客户端选择的指定的持久性策略任选地(和/或异步地)确定计费额(要素2416)。

[0153] 用于流处理的分散的工作量管理

[0154] 在一些实施方案中，可以分散的方式例如通过给定SPS阶段内的工作节点来实现SPS的控制平面功能的很大一部分或者所有，所述给定SPS阶段通过诸如数据库表的共享的数据结构来协调各种控制操作(诸如到工作节点的分区分配、响应于动态重新分区、健康监测和/或负载均衡)。给定的工作节点W1可检查共享的数据结构内的条目，以确定例如当前未处理所述阶段的哪一些分区的输入流(若有的话)。如果发现这种分区P1，那么W1可更新共享的数据结构中的条目，以指示W1将在P1的记录上执行所述阶段的处理操作。其他工作节点可获知W1被分配来处理P1记录，并且可能因此分配不同的分区到它们自己。工作节点可定期或偶尔向SMS控制平面提交查询，以确定用于输入流的当前有效的分区映射，并且必

要时更新共享的数据结构以指示映射改变(例如由于重新分区)。在各种实施方案中,负载均衡和其他操作也可通过共享的数据结构来协调,如下文所述。在一些这种分散的实现方式中,可能不需要专用的控制节点用于SPS,从而减少了实现SPS工作流程所需的开销。这种分散的SPS控制平面实现方式可能特别受关心预算的消费者的欢迎,所述消费者利用SPS客户端库来例如在分配给消费者的供应商网络内的计算实例处或者在供应商网络外部的地点处实现流处理的各个方面。例如当用于SMS和SPS的所有资源在供应商网络内进行配置时,分散的SPS控制平面技术也可用在未使用客户端库的实施方案中。工作节点在其处实现用于至少一些处理阶段的一些或所有SPS控制平面功能的SPS可以在本文中被称为"分散控制SPS"。

[0155] 图25示出根据至少一些实施方案的流处理系统的实例,其中处理阶段的工作节点使用数据库表来协调它们的工作量。在分散控制SPS 2590内,限定了两个阶段215A和215B,每个阶段具有相应组工作节点。阶段215A包括工作节点2540A和2540B,而阶段415B包括工作节点2540K和2540L。对于阶段215A和215B中的每一个,在数据库服务2520处创建了对应的分区分配(PA)表2550,诸如用于阶段215A的PA表2550A、用于阶段215B的PA表2550B。在一些实施方案中,例如响应于客户端库部件或函数的调用,可在阶段初始化期间创建用于给定阶段的PA表2550。每个PA表2550可填入代表所述阶段的输入流的未分配分区的条目或行的初始集(即没有工作节点当前所分配至的分区)。PA表条目的示例性列或属性在图26中示出并在下文中进行描述。启动用于所述阶段的工作节点2540(例如,在计算实例或其他服务器处启动的进程或线程)可以被赋予到所述阶段的PA表的读取/写入访问。来自工作节点的针对PA表的读取和写入在图25中由分别用于工作节点2540A、2540B、2540K和2540L的箭头2564A、2564B、2564K和2564L来表示。

[0156] 给定的工作节点2540可以被配置来通过检查PA表中的条目而选择在其上执行所述阶段的处理操作的特定分区。在一种实现方式中,工作节点2540A可扫描PA表2550A中的条目,直到其发现未分配的分区Pk的条目,并且可尝试通过更新所述条目来将分区Pk分配给其自身,例如通过将工作节点的标识符插入到所述条目的一个列中。这种插入可认为类似于通过工作节点来锁定分区。取决于正在使用的数据库服务的类型,可使用(例如,通过刚好几乎相同的时间识别出未分配的分区的两个或更多个工作节点)管理到PA表条目的潜在并发的写入的不同方法。

[0157] 在一个实施方案中,可使用供应商网络的非相关的多租户数据库服务,所述多租户数据库服务在不必须支持相关的数据库事务语义的情况下支持强烈一致性和条件性的写操作。在这种情况下,可由工作节点使用条件性写操作来更新。如果没有工作节点分配给分区,那么考虑其中列"工作节点ID"被用来指示分配给PA表中的分区的特定工作节点的标识符的实例,并且所述列的值被设置成"空值"。在这种情境下,具有标识符WID1的工作节点可请求以下的逻辑等效物:"如果在用于分区Pk的条目中,工作节点ID是空值,那么将用于那个条目的工作节点ID设置成WID1"。如果这种条件性写请求成功,那么具有标识符WID1的工作节点可假定分区Pk被分配给它。工作节点可随后开始例如使用SMS检索子系统206的记录检索接口来检索分区Pk的数据记录,如由箭头2554(例如分别用于工作节点2540A、2540B、2540K和2540L的箭头2554A、2554B、2554K和2554L)所指示的,并在检索的记录上执行处理操作。如果条件性写入失败,那么工作节点可重新开始搜索不同的未分配的分区。在

其他实施方案中，可使用支持事务的数据库服务(诸如相关的数据库)，并且事务功能可用来实现条件性写操作的等效物，例如以确保仅仅将分区分配给工作节点的多个并发(或接近并发)的尝试中的一个成功，并且在这种并发尝试中所涉及的工作节点被可靠地通知它们的成功或失败。在一些实施方案中，可使用既不依赖于条件性写入也不依赖于事务支持的同步技术。在一些实现方式中，可以不使用数据库服务；相反可由工作节点使用锁定服务以获得排他性访问，用于对类似于PA表的持久性数据结构中的条目的更新。

[0158] 其他工作节点2540可检查PA表中的条目，确定哪些分区是未分配的，并且可最终成功地将一个或多个分区分配给自己。以这种方式，用于所述阶段的一个输入流或多个输入流的分区的处理的工作量可最终由所述阶段的工作节点在它们中进行分配。

[0159] 任何给定流的初始的分区映射可随时间而改变，例如由于之前描述的动态重新分区操作而改变。因此，在图25中描绘的实施方案中，工作节点2540中的一个或多个可偶尔(或者响应于如下文所述的触发条件)向它们的阶段的输入流的SMS控制子系统210提交请求，以获取当前的分区元数据。在一些实现方式中，这种请求可包括SMS控制平面API的调用，诸如由箭头2544A、2544B、2544K和2544L指示的获得流信息API的调用。SMS控制子系统可例如回复所述流的分区的最新的列表和/或其他细节，诸如分区的有效性时期。如果由SMS控制子系统210提供的分区信息不匹配PA表中的条目，那么PA表可由工作节点进行修改，例如通过为一个或多个分区插入或删除条目来进行修改。在至少一些实施方案中，到SMS控制子系统的这种请求2554可通常比记录检索请求2554(和/或数据库读取或写入操作2564)频率低得多，如由箭头2554A的标签“不频繁”所指示的。例如，一旦工作节点分配有分区，其可通常保持检索和处理那个分区的数据记录直到分区数据被完全消费掉(例如，如果所述流的所有者关闭所述流，或者如果由于动态重新分区而关闭所述分区)，或者直到遇到一些其他低可能性的状况(例如，如下文所讨论的，如果不同的工作节点由于检测到的负载不均衡而请求分区的转移)。因此，在各种实施方案中，与调用获得流信息或类似的API相关联的开销可能通常是相当小的，即使响应于任何给定调用而提供有大量的信息(如果成百上千个分区被限定用于阶段的输入流，那么可能会是这种情况)。

[0160] 在图25中描绘的实施方案中，分散控制SPS环境的一些关键的工作量管理操作可因此被概括如下：(a)由流处理阶段的第一工作节点至少部分基于访问数据库表来选择流处理阶段的输入数据流的特定分区，在所述流处理阶段上实现限定用于那个阶段的一组处理操作；(b)将特定分区到第一工作节点的分配的指示符写入到存储在表中的特定条目中；(c)由第一工作节点使用在多租户流管理服务处实现的编程记录检索接口检索特定分区的记录；(d)由第一工作节点在特定分区的记录上实现所述组处理操作；(e)由第二工作节点至少部分基于特定数据库表中的特定条目确定分配第一工作节点以在特定分区上执行所述组处理操作；以及(f)由第二工作节点选择不同的分区，在所述不同的分区上执行所述组处理操作。如果并且当工作节点确定没有更多的记录保留在被分配给其的分区中，那么工作节点可请求来自SMS控制子系统的输入流上的元数据，并且如果元数据指示出差异则可更新PA表。

[0161] 图26示出根据至少一些实施方案的可存储在用于工作量协调的分区分配表2550中的示例性条目。如图所示，表2550可包括四列：分区标识符列2614、分配的工作节点标识符列2618、工作节点健康指示符列2620以及工作量水平指示符列2622。在其他实现方式中，

可实现其他的列设置，例如在一些实施方案中可使用指示分区创建时间或分区功能输出值范围的列，或者可以不使用工作量水平指示符列。

[0162] 应注意，在一些实施方案中，由SMS控制子系统维持的分区列表2650(例如作为分区条目树、图形或之前所描述的其他组合的数据结构的部分)至少在一些时间点处可包括比被包括在PA表2550中更多的分区。在描绘的实例中，分区列表2650包括分区P1、P2、P3、P4和P5，其中P1和P4被示出为由于重新分区的关闭状态，而P2、P3和P5被示出为是有效的(即其数据记录当前正在被检索和处理的分区)。在描绘的实施方案中，PA表2650包括用于有效分区的条目，并且不包括用于关闭的分区的条目(例如，当工作节点在重新分区发生之后获取对获得流信息调用的响应时，所述条目可能已由工作节点删除)。至少在一些实现方式中，并非流的所有当前打开的分区在给定的时间点处可必须具有PA表中的相应的条目；相反，例如可仅仅呈现当前分配或正在处理的那些分区的子集。

[0163] 在图26中示出的示例性情境中，分区P1和P2被分配给分别具有标识符W7和W3的工作节点，而P5当前是未分配的。在不同的实现方式中，健康指示符列2620可存储不同类型的值。在一些实现方式中，工作节点可负责定期(例如，每隔N秒，或者根据基于一些组直观推断的安排)更新它们所分配的分区的PA条目中的健康指示符列的内容，以指示工作节点是有效的并且能够继续它们的检索和处理操作。在图26中，可存储用于所述条目的工作节点更新健康指示符列的最近时间(“最后修改时间”)的指示，例如工作节点W7被示出为在2013年12月1日的02:24:54和53秒处已修改了条目。在一些实施方案中，其他工作节点可使用最后修改时间值来确定分配的工作节点是否健康，例如如果已过去X秒或X分钟，如在用于所述阶段的故障转移策略中所限定的，那么分配的工作节点可能会被假定为不健康或不可访问的，并且可对所述分区进行重新分配。在其他实现方式中，可将计数器用作健康指示符(例如，如果计数器值在Y秒中未改变，那么分配的工作节点可以被视为用于故障转移的候选)，或者可使用指示分配的工作节点何时最后一次读取条目的“最后读取时间”值。

[0164] 在至少一些实施方案中，工作量水平指示符值2622可例如通过分配的工作节点存储在条目中，诸如在一些最近时间间隔期间(例如，在最后修改时间之前的五分钟内)所处理的记录的数量、工作节点的最近的性能相关的指标，诸如CPU利用率、存储器利用率、存储装置利用率等。在一些实施方案中，可由工作节点使用这种工作量水平指示符值，以确定是否存在负载不均衡，如下文关于图29所描述的，并且响应于检测到的不均衡而采取行动。例如，工作节点Wk可确定其工作量水平超过了平均的工作量水平，并且可以不分配其分区中的一个，或者可以请求动态重新分区；可替代地，工作节点Wk可确定其工作量相对于其他工作节点的工作量来说过低，并且可以为其自身分配另外的分区。因此，在描绘的实施方案中，通过使用图26中所指示的PA表的列，工作节点可执行相同类型的控制平面功能中的一些，所述控制平面功能在集中控制的SPS实现方式中可通常由专用的SPS控制节点执行

[0165] 图27示出根据至少一些实施方案的可由流处理阶段的工作节点执行来选择在其上执行处理操作的分区的操作方面。如要素2701中所示，可在用于分散控制的SPS处理阶段SP1的数据库服务处初始化PA表PAT1。可例如当例如从客户端设施处的主机或者从供应商网络数据中心处的计算实例调用SPS客户端库部件时来创建所述表。客户端库可用于各种目的：例如以提供用于有待在SPS阶段处实现的特定处理操作的可执行部件，诸如JAR(Java™存档)文件，以指示标签(诸如程序名称、进程名称或者计算实例名称)，所述标签可

用来识别工作节点、用来指示有待用作用于所述阶段的输入的流、用来指示所述阶段的输出目的地(若有的话)等。在一些实施方案中,可最初为PAT1填入用于被限定用于所述阶段的输入流的分区{P1、P2、……}的至少一个子集的条目或行。在一些实现方式中,可最初保持表格空置,并且一个或多个工作节点可例如由于从SMS控制子系统获取分区元数据而为表格填入用于未分配的分区的行。可例如在供应商网络内的各个计算实例处或者在客户端所有的计算装置处启动工作节点{W1、W2、……}的初始集(要素2704)。在描绘的实施方案中,可赋予工作节点到PAT1的读取和写入访问。

[0166] 当工作节点在线出现时,它们可以各自访问PAT1以试图发现未分配的分区。例如,工作节点W1可检查PAT1并且发现分区P1是未分配的(要素2707)。W1可随后例如取决于正在使用的数据库服务的类型来通过使用条件性的写入请求或事务性的更新请求而在PAT1中更新P1的条目,以指示将P1分配给W1(要素2710)。通过已更新所述表,W1可通过使用SMS检索子系统接口开始P1的数据记录的检索(要素2713),并且可在检索的记录上执行阶段PS1的处理操作。

[0167] 同时,在一些时间点处,不同的工作节点W2可以其自己的尝试来访问PAT1,以发现未分配的分区(要素2716)。W2可基于W1的之前的更新来确定已分配P1,但是未分配不同的分区P2。在一些实施方案中,由W2(例如基于P2的条目中的健康指示符列)作出的P2的当前分配的工作节点不健康或不活动的确定也可引导W2选择P2。因此,在至少一些实施方案中,未分配状态或者当前工作节点的不健康状态的确定可用来选择用于重新分配(或初始分配)的给定分区。W2可随后尝试更新PAT1以将P2分配给自己(要素2719)。如果更新成功,那么W2可开始使用SMS检索接口来检索P2记录(要素2722),并执行被限定用于所述阶段的适当的处理操作。

[0168] 如之前提及的,分散控制的SPS中的工作节点可(通常非频繁地)从SMS获取分区映射信息,并且如果需要来使用这种信息更新PA表。图28示出根据至少一些实施方案的可由流处理阶段的工作节点执行来基于从流管理服务控制子系统获取的信息更新分区分配表的操作方面。如要素2801中所示,在工作节点初始化期间或者响应于各种触发条件(诸如分配给其的分区中的一个的关闭),工作节点W1可向SMS控制子系统提交请求以获取最近的或当前的分区列表或有效的分区列表。在一些实现方式中,出于这个目的可调用获得流信息或类似的API。在一些实施方案中,可使用其他触发条件:例如,在随机量的时间之后或者响应于工作量水平的未预期的减少或增加,工作节点可各自被配置来获取新的分区列表。可将由SMS返回的分区列表与用于所述分区的PA表中的条目进行比较(要素2807)。在描述的实施方案中,如果发现差异(例如,如果在最新获取的分区列表中存在不在PA表中的一些分区,或者如果在PA表中存在不在SMS的列表中的条目),那么工作节点可在PA表中插入或删除条目,以解决所述差异(要素2810)。(如果在一些实现方式中,以删除为目标的条目当前具有分配的工作节点,那么可能需要另外的协调,例如可直接地或通过PA表自身通知分配的工作节点。

[0169] 在调整所述差异之后,或者如果没有检测到差异,那么工作节点W1可选择一组分区,在所述组分区上工作节点W1应当执行所述阶段的处理操作(要素2813),并且因此可更新PA表。在一些情况下,取决于引起检索分区列表的触发条件,W1可能已经具有分配给其的一个或多个分区,并且可能不需要对它的分配做出改变或者更新PA表。W1可随后在没有必

须与SMS控制子系统进行交互或者改变PA表中的条目的数量的情况下，继续检索其所分配的一个分区或多个分区的数据记录，并处理所述记录(要素2816)。最终，当检测到触发条件时(例如当“到达的分区终点”响应的等效物被接收到检索请求，从而指示分区是关闭的)，为了最新的分区信息W1可再次向SMS控制子系统发送请求，并且可重复要素2801向前的操作。

[0170] 图29示出根据至少一些实施方案的可由流处理阶段的工作节点执行的负载均衡操作的方面。如要素2901中所示，工作节点W1可确定当检测到各种触发条件中的任何一种，诸如检测到高资源利用率水平时或者基于可配置的安排而有待在其阶段上执行负载均衡分析。W1可检查PA表中的条目(要素2904)，以确定用于所述阶段的各种工作量指标。这种指标可包括分配给工作节点的分区的平均数、工作节点的或者不同分区的平均工作量水平(在将工作量水平指示符保存在表中的实施方案中)、每个工作节点工作量的范围或分配等。

[0171] W1可随后(例如基于分配给W1的分区的数量和/或每个分区工作量水平指示符)将其自身的工作量与所述指标的一些或所有进行比较。一般来说，可获得三种类型的结论中的任何一种：W1是过载的、W1是欠载的、或者W1的工作量既不太高也不太低。可通过由在一些实施方案中为其利益而配置所述阶段的客户端所选择的策略或者在其他实施方案中通过使用一些默认设置的直观推断来限定“过高”或“过低”的工作量水平。如果W1确定其工作量过低(要素2907)，例如低于一些最小负载阈值T1，那么可识别更忙或者更高负载的工作节点Wk(要素2910)。W1可随后例如通过尝试修改PA表中的Pm条目、请求这种修改(这可能产生正在生成用于Wk的通知)或者通过直接请求Wk来开始将一个或多个分区Pm从Wk转移至W1的过程(要素2913)。

[0172] 如果W1确定其工作量过高(要素2916)，例如超过最大阈值T2，那么W1可识别一个或多个其所分配的分区Pn以放弃(即，释放来由其他工作节点进行分配)(要素2919)。W1可随后例如通过从用于Pn的条目的分配的列移除其标识符来修改PA表中的适当的条目(要素2922)。如果W1的工作量既不太高也不太低或者在W1后已采取上文所述的多种行动来增加或减少其工作量，那么W1可重新开始处理其被分配至的分区的记录(要素2925)。当并且如果符合触发另一个负载均衡分析的条件时，可重复对应于要素2901的向前的操作。应注意，在图29中示出的操作中，W1被示出为仅仅当其相对于其自身的工作量而检测到不均衡时才开始工作量的变化。在其他实施方案中，如果W1除自身之外在其他工作节点中检测到不均衡时，例如如果W1确定W2具有比W3低得多的工作量水平时，那么W1可开始重新均衡行动。在一些实现方式中，如果并且当W1检测到工作量不均衡时，那么W1可(例如通过调用诸如图3中示出的重新分区流(repartitionStream)SMS API及其等效物)请求或开始动态重新分区。在一些实施方案中，可由最近配置的工作节点来执行图29中示出的许多种操作，例如当在阶段已在操作中一段时间后将新的节点添加至所述阶段时，新的节点可通过请求来自重负载的现有节点的分区的重新分配来间接通知现有节点它们的存在。在一些实施方案中，也可在一个或多个SMS子系统处使用或者替代地使用类似于上文描述的那些的用于SPS工作节点的分散的控制技术，例如摄取、存储或检索子系统的节点可使用类似于PA表的共享的数据结构来协调它们的工作量。

[0173] 应注意，在各种实施方案中，可使用除了在图17-图24以及图27-图29的流程图中

所示出的那些操作的操作，以实现上文所述的流管理服务和/或流处理功能。在一些实施方案中，示出的操作中的一些可以不实现，或者可以不同的顺序来实现，或者并行而非顺序地实现。还应注意，关于在各种实施方案中编程接口所支持的SMS和SPS功能中的每一个，一种或多种技术的任何组合可用来实现所述接口，包括网页、网址、网络服务API、其他API、命令列工具、图形用户界面、移动应用程序(app)、平板电脑app等。

[0174] 使用案例

[0175] 建立用于流数据记录的采集、存储、检索和阶段性处理的可度量的基于分区的动态可配置管理的多租户服务的上文所述的技术在许多种情境中可能是有用的。例如，大型供应商网络可包括数以千计的实例主机，从而实现同时用于数以万计的客户端的许多不同的多租户或单租户服务的服务实例。各种实例和主机上所安装的监测和/或计费代理可快速产生数以千计的指标记录，可能需要存储和分析所述指标记录以产生精确的计费记录，来确定用于供应商网络的数据中心的有效的供应计划，来检测网络攻击等。监测的记录可形成到用于可度量的摄取和存储的SMS的输入流，并且可实现描述的SPS技术用于采集的指标的分析。类似地，采集和分析来自许多日志源(例如，来自分布式应用程序的节点的应用程序日志，或者来自数据中心处的主机或计算实例的系统日志)的大量的日志记录的应用程序也可以能够利用SMS和SPS功能。在至少一些环境中，SPS处理操作可包括实时ETL(提取转换加载)处理操作，(即，将接收的数据记录实时地转换用于加载到目的地中的操作，而不是离线地进行所述转换)，或者用于插入到数据仓库中的数据记录的转换。使用用于将数据实时地加载到数据仓库中的SMS/SPS组合在将所述数据插入仓库中用于分析之前可避免对于清洁和整理来自一个或多个数据源的数据通常所需的延迟。

[0176] 许多不同的“大数据”应用程序也可使用SMS和SPS技术来构建。例如，可使用流来有效地执行各种形式的社交媒体交互中的趋势的分析。可将从移动电话或平板电脑采集的数据，诸如用户的地点信息作为流记录来管理。例如从监控摄像机机群采集的音频或视频信息可代表可以可度量方式采集和处理从而潜在地有助于防止各种类型的攻击的流数据集的另一个种类。需要例如从气象卫星、基于海洋的传感器、基于森林的传感器、天文望远镜采集的日益增长的数据集的分析的科学应用程序也可得益于本文所述的流管理和处理能力。基于灵活策略的配置选项和定价选项可帮助不同类型的用户定制适合他们的指定预算和数据持久性/可用性需求的流功能。

[0177] 本公开的实施方案可鉴于以下条款来描述：

[0178] 1.一种系统，其包括：

[0179] 一个或多个计算装置，其被配置来：

[0180] 实现一个或多个编程接口，从而使得多租户流处理服务的客户端能够对应于与指定的数据流相关联的特定的处理阶段来指示：(a)有待根据分区策略在所述指定的数据流的数据记录上执行的处理操作，和(b)用于所述处理操作的结果的输出分配描述符；

[0181] 通过所述一个或多个编程接口从特定的客户端接收有待在所述特定的处理阶段处的特定的数据流的数据记录上执行特定的处理操作的指示，和用于所述特定的处理操作的结果的特定的输出分配描述符；

[0182] 至少部分基于所述分区策略并且至少部分基于有待部署为用于所述处理阶段的工作节点的资源的估计的性能能力来确定用于所述指定的数据流的初始数量的工作节点；

[0183] 配置所述初始数量的工作节点的特定的工作节点来:(a)接收所述特定的数据流的一个或多个分区的数据记录,(b)在接收的数据记录上执行所述特定的处理操作,(c)存储进度记录,所述进度记录指示已在所述工作节点处处理过的所述一个或多个分区的部分以及(d)根据所述特定的输出分配描述符将所述特定的处理操作的结果传递至一个或多个目的地;

[0184] 监测所述特定的工作节点的健康状态;以及

[0185] 响应于所述特定的工作节点正处于不希望的状态中的确定,配置替换工作节点以替换所述特定的工作节点,其中所述替换工作节点访问由所述特定的工作节点存储的进度记录,以识别所述一个或多个分区的至少一个数据记录,在所述一个或多个分区上有待由所述替换工作节点执行所述特定的处理操作。

[0186] 2.如条款1所述的系统,其中所述特定的输出分配描述符指示有待根据不同的分区策略将所述特定的处理操作的结果作为不同的数据流的数据记录分配至或更多个被配置用于所述不同的数据流的摄取节点。

[0187] 3.如条款1所述的系统,其中所述一个或多个计算装置还被配置来:

[0188] 从所述特定的客户端接收另一个处理阶段的指示,有待将所述特定的数据流的数据记录作为输入提供给所述另一个处理阶段,其中有待在所述其他处理阶段处执行不同的处理操作;以及

[0189] 配置用于所述其他处理阶段的另外组的工作节点。

[0190] 4.如条款1所述的系统,其中所述一个或多个计算装置还被配置来:

[0191] 响应于被配置来为另一个处理阶段执行另一个处理操作的不同的工作节点正处于不希望的状态中的确定,配置不同的替换工作节点以在一个或多个随后接收的数据记录上执行所述其他处理操作,而不访问进度记录。

[0192] 5.如条款1所述的系统,其中所述一个或多个计算装置还被配置来:

[0193] 响应于在所述处理阶段的不同工作节点处的工作量水平符合触发标准的确定,实现阶段重新配置操作,所述阶段重新配置操作包括以下各项中的一个或多个:(a)当继续处理所述流的另外的数据记录时所执行的所述特定的数据流的动态重新分区,(b)替代工作节点到在所述不同的工作节点处先前处理的至少一个分区的分配,(c)被配置用于所述处理阶段的多个工作节点的改变或者(d)工作节点从一个服务器到另一个服务器的转移。

[0194] 6.一种方法,其包括:

[0195] 由一个或多个计算装置执行以下各项:

[0196] 在多租户流处理服务处从特定的客户端接收有待在指定的处理阶段处在特定的数据流的数据记录上执行特定操作的指示,和用于所述特定操作的结果的特定的输出分配描述符;

[0197] 至少部分基于所述特定的操作来确定有待配置用于所述指定的处理阶段的工作节点的初始数量;

[0198] 配置所述初始数量的工作节点的特定的工作节点来:(a)在所述特定的数据流的一个或多个分区的接收的数据记录上执行所述特定的操作,(b)存储进度记录,所述进度记录指示已在所述工作节点处处理过的所述一个或多个分区的部分以及(c)根据所述特定的输出分配描述符将所述特定的操作的结果传递至一个或多个目的地;以及

[0199] 响应于所述特定的工作节点正处于不健康的状态中的确定，选择替换工作节点以替换所述特定的工作节点，其中所述替换工作节点访问由所述特定的工作节点存储的进度记录，以识别所述一个或多个分区的至少一个数据记录，在所述一个或多个分区上有待由所述替换工作节点执行所述特定的操作。

[0200] 7.如条款6所述的方法，其还包括由所述一个多个计算装置执行：

[0201] 调用由多租户流管理服务所实现的一个或多个编程数据记录检索接口，以接收所述一个或多个分区的数据记录，包括特定的编程数据记录检索接口，其包括请求的数据记录的分区内的序列号的指示作为参数。

[0202] 8.如条款6所述的方法，其还包括由所述一个多个计算装置执行：

[0203] 实现一个或多个编程接口以使得所述流处理服务的客户端能够指定用于一个或多个数据流的数据记录的处理阶段的有向无环图。

[0204] 9.如条款6所述的方法，其还包括由所述一个多个计算装置执行：

[0205] 从负责所述特定的数据流的所述数据记录的存储的多租户流管理服务获取分区策略正在用于所述特定的数据流的指示；以及

[0206] 至少部分基于所述分区策略来确定所述工作节点的初始数量。

[0207] 10.如条款6所述的方法，其中所述特定的输出分配描述符指示有待根据不同的分区策略将所述特定操作的结果作为不同的数据流的数据记录分配至或更多个被配置用于所述不同的数据流的摄取节点。

[0208] 11.如条款6所述的方法，其还包括由所述一个多个计算装置执行：

[0209] 响应于被配置来为另一个处理阶段执行另一个操作的不同的工作节点正处于不希望的状态中的确定，配置不同的替换工作节点以在一个或多个随后接收的数据记录上执行所述其他操作，而不访问进度记录。

[0210] 12.如条款6所述的方法，其还包括由所述一个多个计算装置执行：

[0211] 响应于在所述处理阶段的不同工作节点处的工作量水平符合触发标准的确定，实现以下各项中的一个或多个：(a)所述特定的数据流的动态重新分区，(b)替代工作节点到在所述不同的工作节点处先前处理的至少一个分区的分配，(c)被配置用于所述处理阶段的多个工作节点的改变或者(d)工作节点从一个服务器到另一个服务器的转移。

[0212] 13.如条款6所述的方法，其中根据输入格式来格式化所述特定操作的结果，所述输入格式与另一个流处理系统兼容，并且其中所述一个或多个目的地的特定的目的地包括所述其他流处理系统的输入节点。

[0213] 14.如条款6所述的方法，其中将所述特定的工作节点配置来将条目存储在持久性数据资源库中，所述条目代表积累的应用程序状态信息，所述状态信息对应于在所述特定的工作节点处处理过的多个数据记录，并且将所述特定的工作节点配置成包括进度记录中的所述条目的指示。

[0214] 15.如条款6所述的方法，其中所述操作包括以下各项中的一个：日志记录分析操作、资源监测指标分析、计费额计算、传感器数据分析、社交媒体交互分析、实时ETL(提取转换加载)处理操作或者在将数据记录插入到数据仓库中之前所述数据记录的转换。

[0215] 16.如条款6所述的方法，其还包括由所述一个多个计算装置执行：

[0216] 响应于通过客户端库部件的调用的流处理配置请求，在所述多租户流处理服务处

注册指定的资源作为用于不同的处理阶段的工作节点。

[0217] 17.如条款6所述的方法，其还包括由所述一个多个计算装置执行：

[0218] 响应于通过客户端库部件的调用的流处理配置请求，在所述多租户流处理服务处确定有待在不同的处理阶段处实现的一种或多种控制平面功能。

[0219] 18.如条款6所述的方法，其中所述操作是幂等操作。

[0220] 19.如条款6所述的方法，其还包括由所述一个多个计算装置执行：

[0221] 在所述多租户流处理服务处从所述特定的客户端接收有待在不同的处理阶段处在不同的数据流的数据记录上执行特定的非幂等操作的指示；以及

[0222] 配置所述不同处理阶段的第一工作节点，以在接收的数据记录上执行所述非幂等操作。

[0223] 20.如条款19所述的方法，其还包括由所述一个多个计算装置执行：

[0224] 配置所述不同处理阶段的所述第一工作节点以：(a)执行清除操作以将所述非幂等操作的结果存储至一个或多个目的地，和(b)将清除操作定时的指示保存在持久性存储地点处；以及

[0225] 使用所述清除操作定时的所述指示来配置替换工作节点，以在所述第一工作节点故障之后的恢复期间重演所述清除操作。

[0226] 21.一种存储程序指令的非暂时性计算机可访问存储介质，所述程序指令当在一个或多个处理器上执行时实现多租户流处理服务的控制节点，其中所述控制节点可操作来：

[0227] 通过编程接口从特定的客户端接收有待在特定的数据流的数据记录上执行的特定操作的指示；

[0228] 至少部分基于与所述特定的数据流相关联的分区策略来确定在处理阶段处用于所述指定的数据流的工作节点的初始数量；

[0229] 配置所述初始数量的工作节点的特定的工作节点，以在所述特定的数据流的一个或多个分区的接收的数据记录上执行所述特定操作；以及

[0230] 响应于所述特定的工作节点正处于不健康的状态中的确定，配置替换工作节点以替换所述特定的工作节点。

[0231] 22.如条款21所述的非暂时性计算机可访问存储介质，其中所述控制节点可操作来：

[0232] 配置冗余组，所述冗余组包括处理不同数据流的不同分区的数据记录的多个工作节点，其中将所述多个工作节点中的至少一个工作节点指定为接收所述不同分区的所述数据记录的主要节点，并且其中将所述多个工作节点中的至少另一个工作节点配置为响应于触发事件来承担主要节点的责任的备用节点。

[0233] 23.如条款21所述的非暂时性计算机可访问存储介质，其中所述控制节点可操作来：

[0234] 响应于在所述处理阶段的不同工作节点处的工作量水平符合触发标准的确定，实现以下各项中的一个或多个：(a)所述特定的数据流的动态重新分区，(b)替代工作节点到在所述不同的工作节点处先前处理的至少一个分区的分配，(c)被配置用于所述处理阶段的多个工作节点的改变或者(d)工作节点从一个服务器到另一个服务器的转移。

[0235] 24.如条款21所述的非暂时性计算机可访问存储介质，其中所述特定的输出分配描述符指示有待根据不同的分区策略将所述特定操作的结果作为短暂数据流的数据记录分配至或更多个被配置用于所述短暂数据流的摄取节点，对于所述短暂数据流来说到持久性存储装置的存储是不需要的。

[0236] 25.如条款21所述的非暂时性计算机可访问存储介质，其中所述控制节点可操作来：

[0237] 实现一个或多个编程接口以使得所述流处理服务的客户端能够指定用于一个或多个数据流的数据记录的处理阶段的有向无环图。

[0238] 说明性计算机系统

[0239] 在至少一些实施方案中，实现本文所述的技术的一种或多种的部分或全部的服务器可包括通用计算机系统，所述通用计算机系统包括或被配置来访问一个或多个计算机可访问介质，本文所述的技术包括实现SMS子系统(例如，摄取、存储、检索和控制子系统)的部件以及SPS工作节点和控制节点的技术。图30示出这种通用计算装置9000。在示出的实施方案中，计算装置9000包括通过输入/输出(I/O)接口9030联接到系统存储器9020的一个或多个处理器9010。计算装置9000还包括联接到I/O接口9030的网络接口9040。

[0240] 在各种实施方案中，计算装置9000可为包括一个处理器9010的单处理器系统，或包括若干处理器9010(例如两个、四个、八个或另一合适数量)的多处理器系统。处理器9010可为能够执行指令的任何处理器。例如，在各种实施方案中，处理器9010可为实施各种指令集架构(ISA)中任何一种架构的通用或嵌入式处理器，所述架构例如x86、PowerPC、SPARC、或MIPS ISA或任何其他合适ISA。在多处理器系统中，每一个处理器9010可通常但不一定实施相同的ISA。在一些实现方式中，可替代常规的处理器或除了常规的处理器之外来使用图形处理单元(GPU)。

[0241] 系统储存器9020可以被配置来存储可由处理器9010访问的指令和数据。在各种实施方案中，系统储存器9020可使用任何合适储存器技术来实施，所述储存器技术例如静态随机存取储存器(SRAM)、同步动态RAM(SDRAM)、非易失性/快闪型储存器或任何其它类型的储存器。在示出的实施方案中，实现一个或多个所需功能的程序指令和数据(诸如上述那些方法、技术和数据)被示出作为代码9025和数据9026存储在系统存储器9020内。

[0242] 在一个实施方案中，I/O接口9030可被配置来协调处理器9010、系统存储器9020和装置中的任何外围装置之间的I/O流量，所述外围装置包括网络接口9040或其他外围接口，诸如用来存储数据对象分区的物理副本的各种类型的持久性和/或易失性存储装置。在一些实施方案中，I/O接口9030可执行任何必需协议、时序或其他数据转换以便将来自一个部件(例如，系统存储器9020)的数据信号转换成适合于由另一个部件(例如，处理器9010)使用的格式。在一些实施方案中，I/O接口9030可包括对于通过各种类型的外围总线附接的装置的支持，所述外围总线例如外围组件互连(PCI)总线标准或通用串行总线(USB)标准的改变形式。在一些实施方案中，I/O接口9030的功能可分成两个或更多个单独的部件中，例如北桥和南桥。另外，在一些实施方案中，I/O接口9030的一些或所有功能，例如至系统储存器9020的接口，可直接并入处理器9010中。

[0243] 网络接口9040可以被配置来允许数据在计算装置9000与附接到一个或多个网络9050的其他装置9060(例如像图1至图29中所示的其他计算机系统或装置)之间进行交换。

在各个实施方案中，网络接口9040可以支持经由任何合适的有线或无线通用数据网络(例如像以太网网络类型)进行通信。另外，网络接口9040可以支持经由电信/电话网络(如模拟语音网络或数字光纤通信网络)、经由存储区域网络(如光纤信道SAN)或经由任何其他合适类型的网络和/或协议进行通信。

[0244] 在一些实施方案中，系统存储器9020可以是如上文针对图1至图29所描述的被配置来存储程序指令和数据的计算机可访问介质的一个实施方案，以用于实现对应的方法和设备的实施方案。然而，在其它实施方案中，可以在不同类型的计算机可访问介质上接收、发送或存储程序指令和/或数据。一般来说，计算机可访问介质可包括非临时性的存储介质或存储器介质，例如磁性介质或光学介质，例如通过I/O接口9030联接到计算装置9000的磁盘或DVD/CD。非暂时性计算机可访问存储介质还可以包括可作为系统存储器9020或另一类型的存储器被包括在计算装置9000的一些实施方案中的任何易失性或非易失性介质，诸如RAM(例如，SDRAM、DDR SDRAM、RDRAM、SRAM等)、ROM等。另外，计算机可访问介质可以包括传输介质或信号，诸如经由通信介质(网络和/或无线链路)传送的电信号、电磁信号或数字信号，例如可以经由网络接口9040来实施。在各种实施方案中，诸如图30中示出的多个计算装置的部分或全部可用来实现所述功能；例如，在各种不同装置上运行的软件部件和服务器可合作来提供所述功能。在一些实施方案中，除了或者代替使用通用计算机系统来实现，所述功能的部分可使用存储装置、网络装置或专用计算机系统来实现。如本文使用的术语“计算装置”是指至少所有这些类型的装置，并且不限于这些类型的装置。

[0245] 结论

[0246] 各个实施方案还可以包括根据前面的描述实现的在计算机可访问介质上接收、发送或存储指令和/或数据。一般来说，计算机可访问介质可以包括存储介质或存储器介质(诸如磁性介质或光学介质，例如磁盘或DVD/CD-ROM)、易失性或非易失性介质(诸如RAM(例如，SDRAM、DDR、RDRAM、SRAM等)、ROM等)以及传输介质或信号(诸如通过通信介质(诸如网络和/或无线链路)传送的信号(诸如电信号、电磁信号或数字信号))。

[0247] 如在图中所示和本文所描述的各种方法表示方法的示例性实施方案。所述方法可以在软件、硬件或其组合中实施。方法的顺序可以改变，并且各个元素可以被添加、重新排序、组合、省略、修改等。

[0248] 受益于本公开的本领域技术人员将清楚可进行各种修改和变化。旨在包含所有这些修改和变化，并且相应地，以上描述应视为具有说明性而非限制性意义。

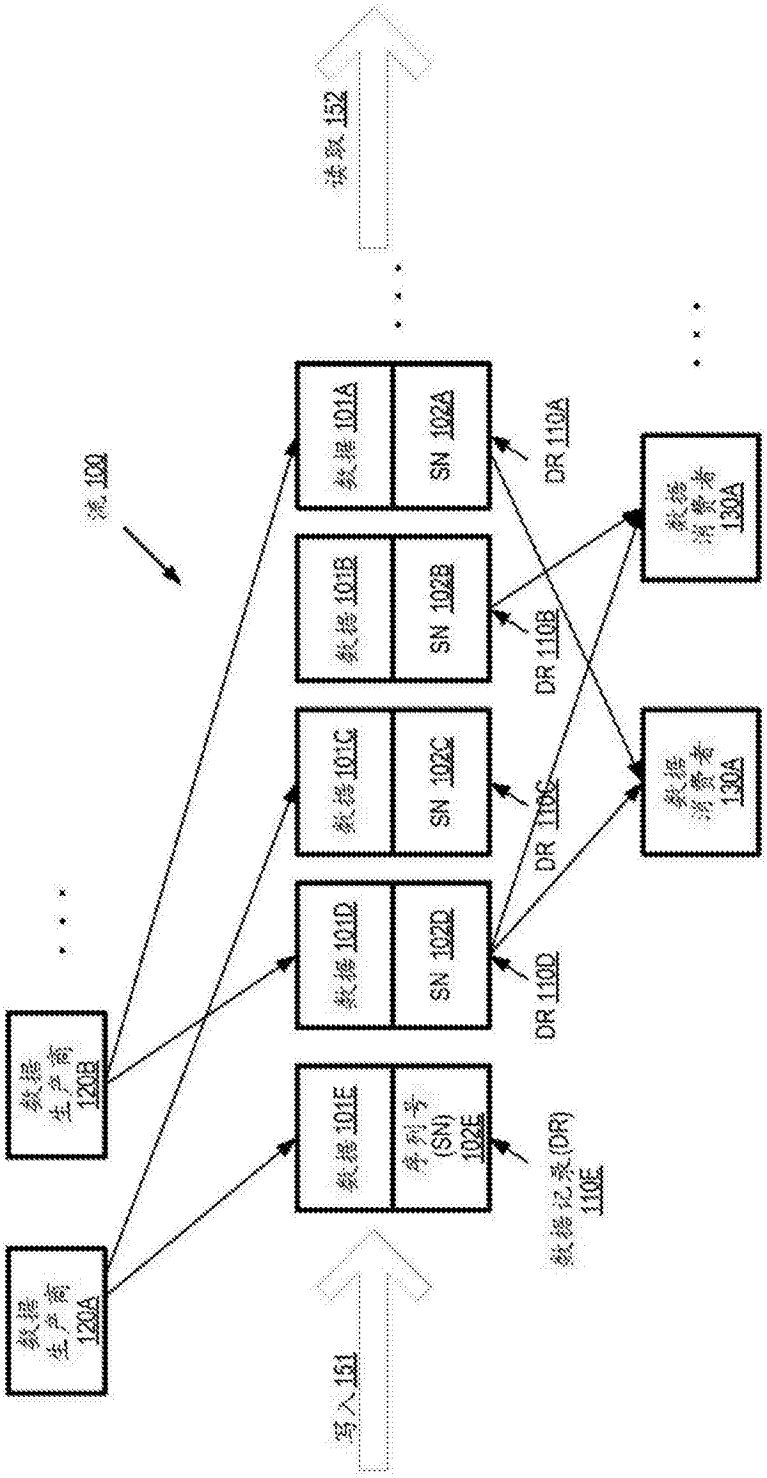

读取 152
流 100
数据 101A
SN 102A
DR 110A
数据 101B
SN 102B
DR 110B
数据 101C
SN 102C
DR 110C
数据 101D
SN 102D
DR 110D
数据 101E
序列号 (SN) 102E
数据记录(DR) 110E
数据生产商 120B
数据生产商 120A
数据消费者 130A
数据消费者 130A
写入 151

图1

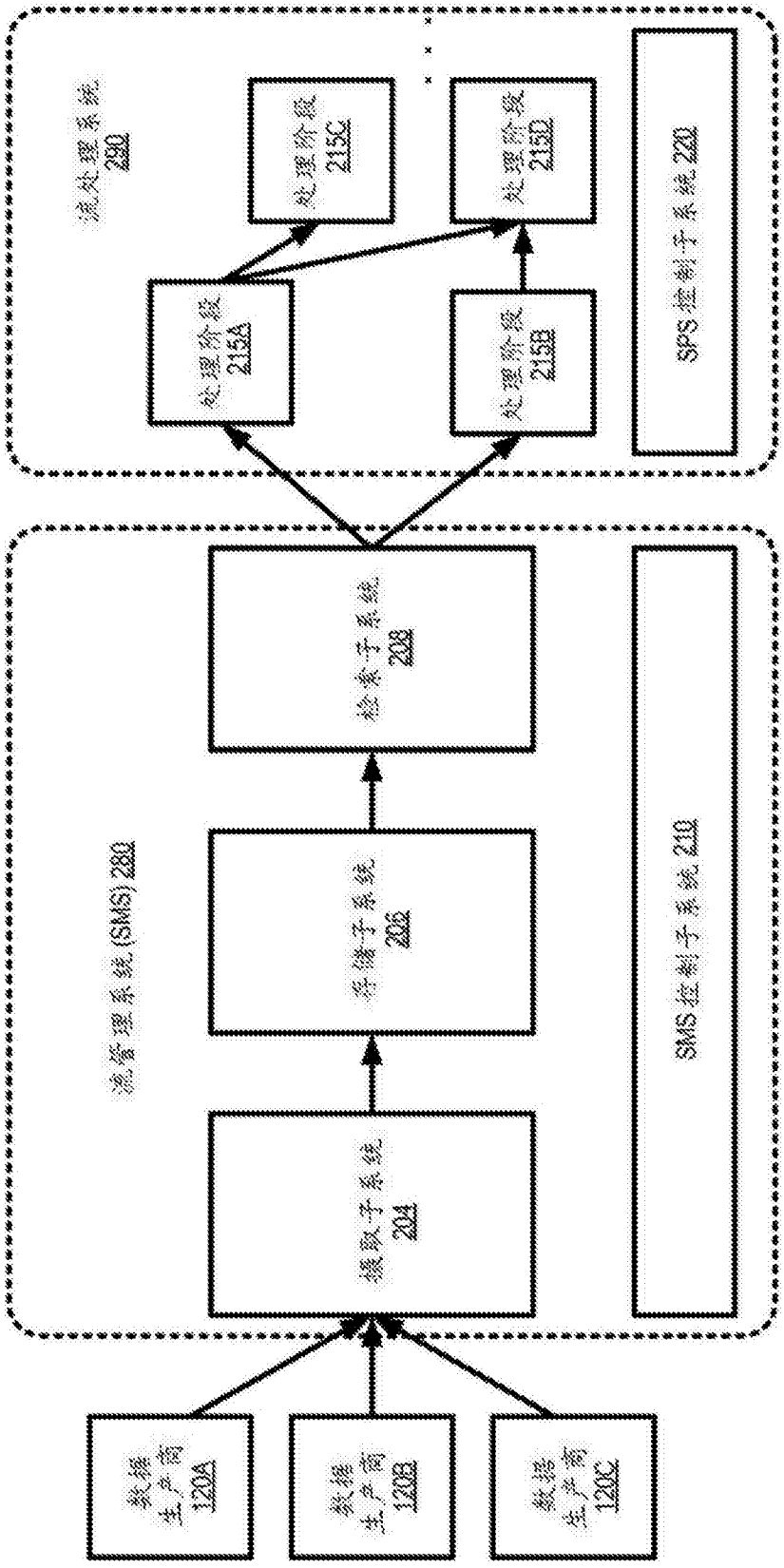

流处理系统
290
处理阶段
215C
处理阶段
215D
SPS 控制子系统 220
处理阶段
215A
处理阶段
215B
流管理系统 (SMS) 280
检索子系统
208
存储子系统
206
SMS 控制子系统 210
获取子系统
204
数据
生产商
120A
数据
生产商
120B
数据
生产商
120C

图2

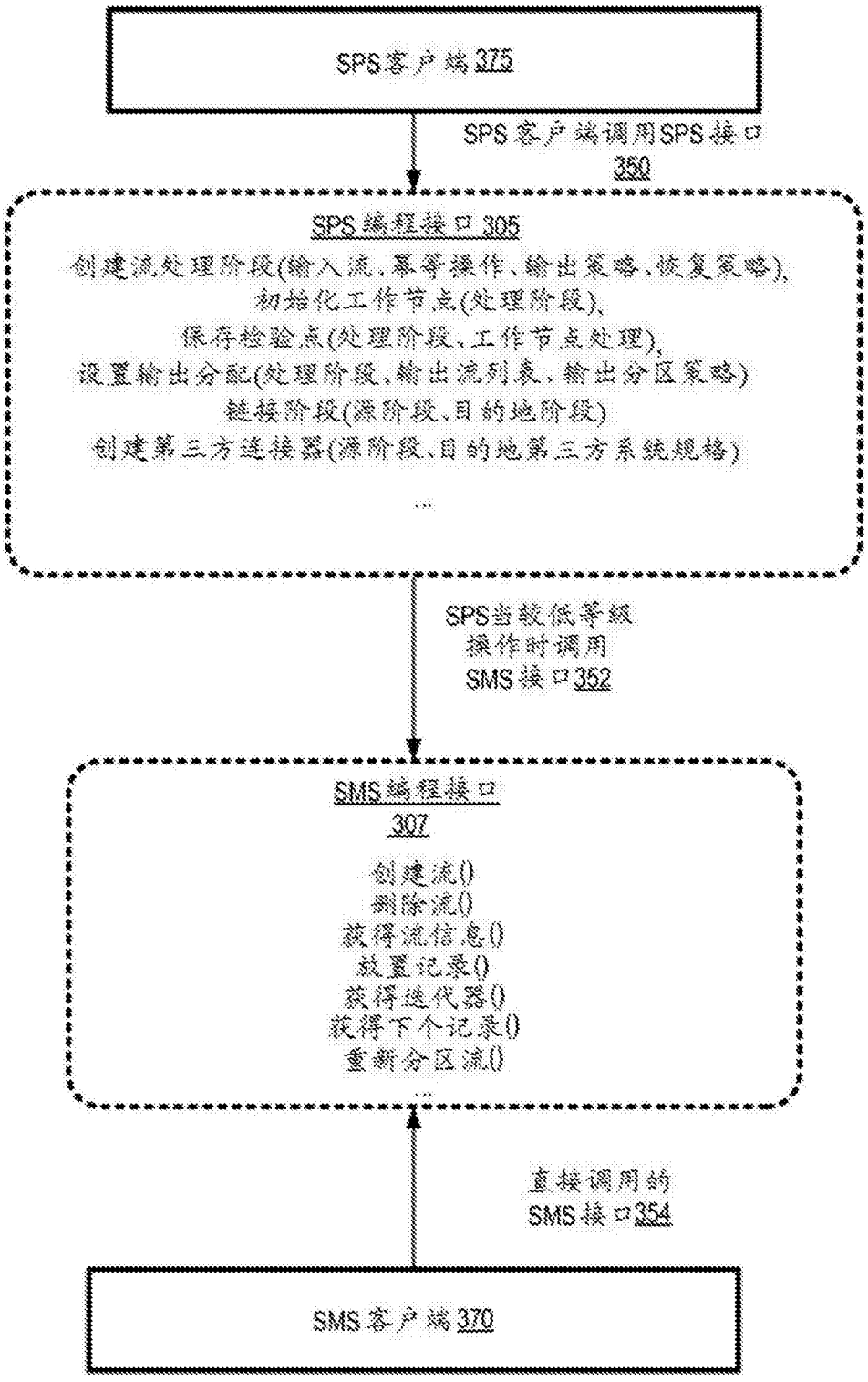

SPS客户端375
SPS客户端调用SPS接口350
SPS编程接口305
创建流处理阶段(输入流、幂等操作、输出策略、恢复策略),
初始化工作节点(处理阶段),
保存检验点(处理阶段、工作节点处理),
设置输出分配(处理阶段、输出流列表、输出分区策略)
链接阶段(源阶段、目的地阶段)
创建第三方连接器(源阶段、目的地第三方系统规格)
...
SPS当较低等级操作时调用SMS接口352
SMS编程接口307
创建流()
删除流()
获得流信息()
放置记录()
获得迭代器()
获得下个记录()
重新分区流()
...
直接调用的SMS接口354
SMS客户端370

图3

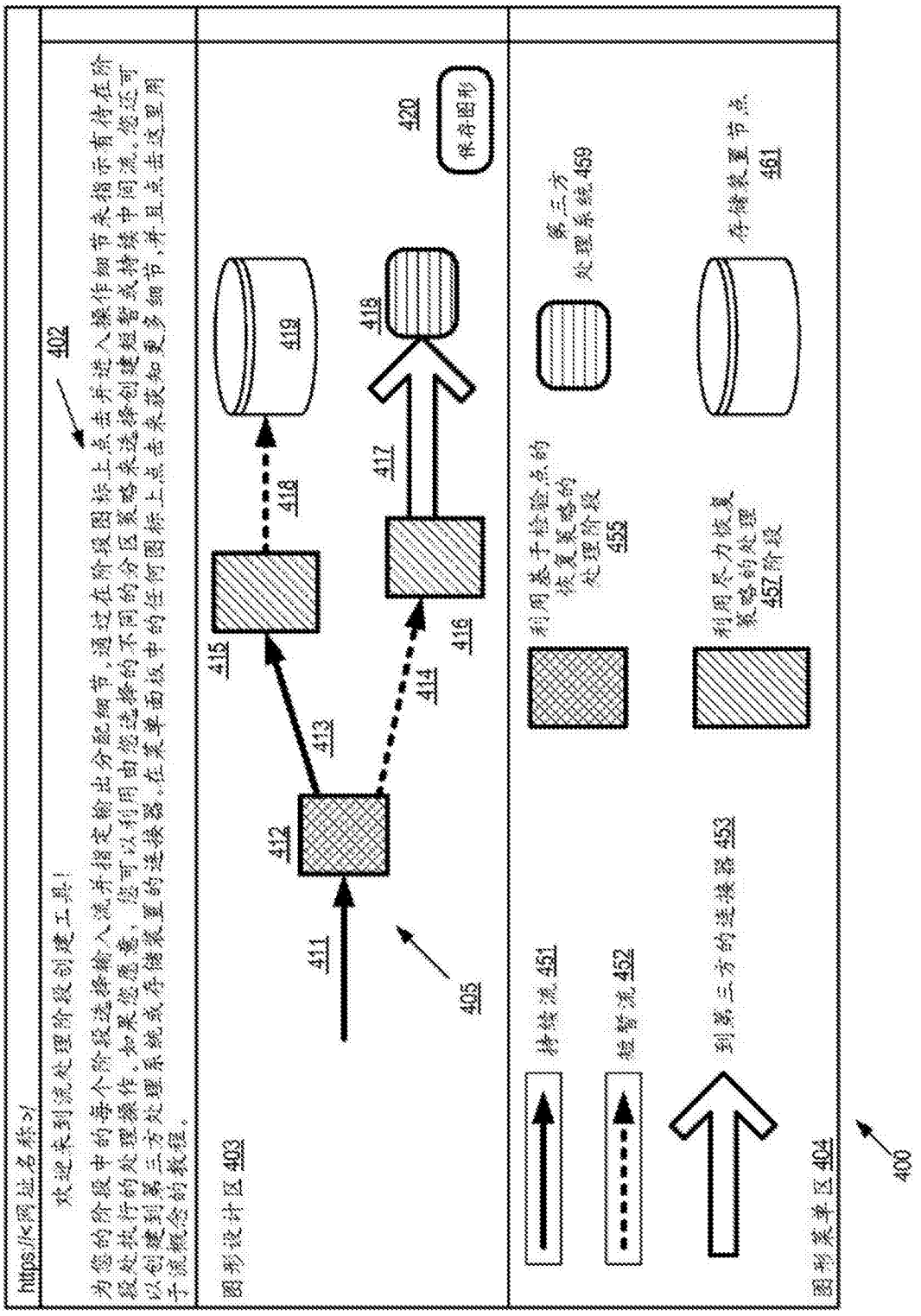

https://网址名称/
欢迎来到流处理阶段创建工具!
为您的阶段中的每个阶段选择输入流并指定输出分配细节,通过在阶段图标上点击并进入操作细节来指示有待在阶段执行的处理操作。如果您愿意,您可以利用由您选择的不同的分区策略来选择创建短暂或持续中间流。您还可以创建到第三方处理系统或存储装置的连接器。在完事面板中的任何图标上点击来获知更多细节,并且点击这里用于流概念的教程。
402
图形设计区 403
411
412
413
414
415
416
417
418
419
420
保存图形
405
持续流 451
短暂流 452
到第三方的连接器 453
利用基于检查点的恢复策略的处理阶段 455
利用尽力恢复策略的处理阶段 457
第三方处理系统 459
存储装置节点 461
图形菜单区 404
400

图4

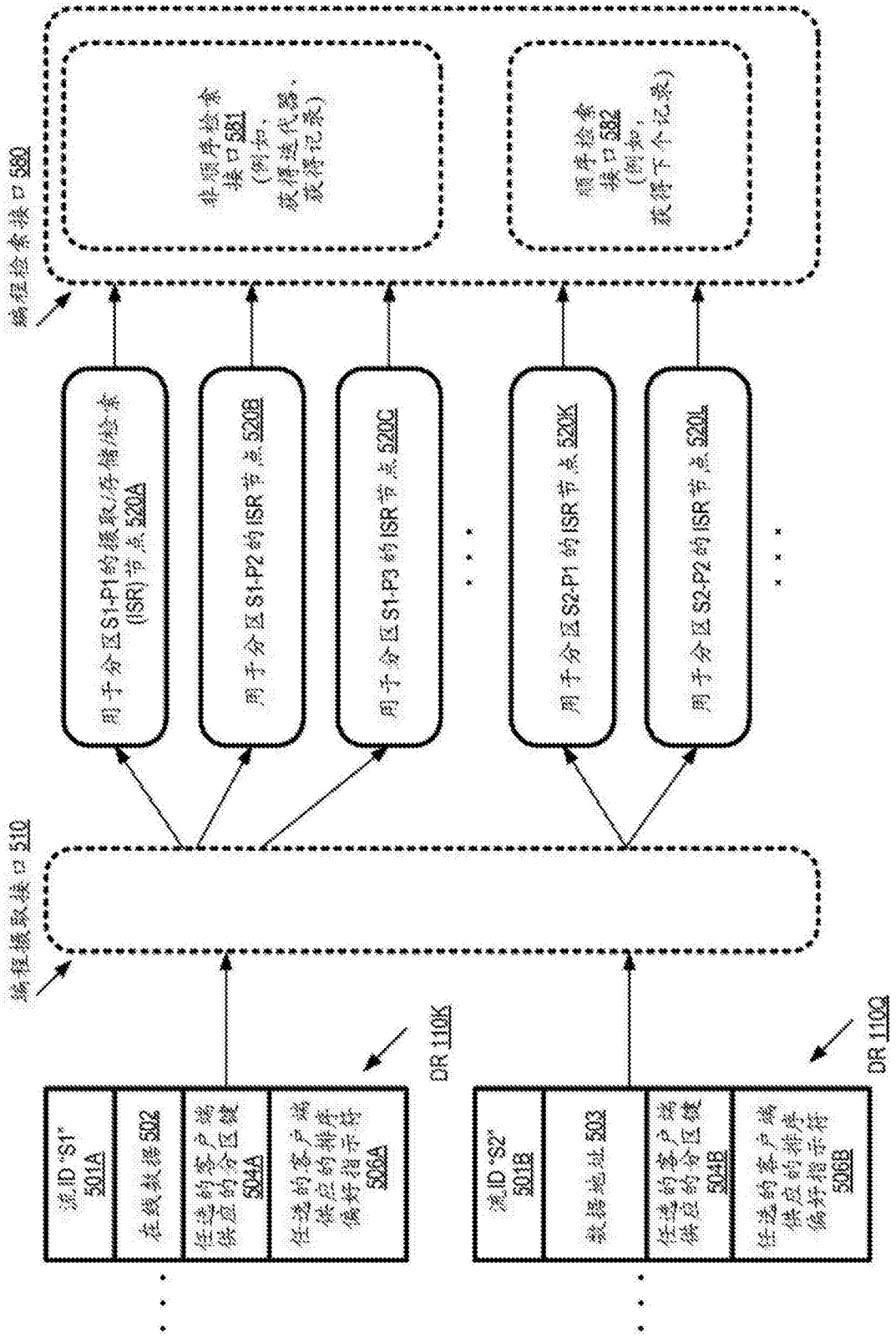

编程检索接口 580
非顺序检索接口 581
(例如，获得迭代器，获得记录)
顺序检索接口 582
(例如，获得下个记录)
用于分区 S1-P1 的级联/存储/检索(ISR)节点 520A
用于分区 S1-P2 的 ISR 节点 520B
用于分区 S1-P3 的 ISR 节点 520C
用于分区 S2-P1 的 ISR 节点 520K
用于分区 S2-P2 的 ISR 节点 520L
编程级联接口 510
DR 110K
DR 110Q
流 ID "S1" 501A
在线数据 502
任选的客户端供应的分区键 504A
任选的客户端供应的排序偏好指示符 506A
流 ID "S2" 501B
数据地址 503
任选的客户端供应的分区键 504B
任选的客户端供应的排序偏好指示符 506B

图5

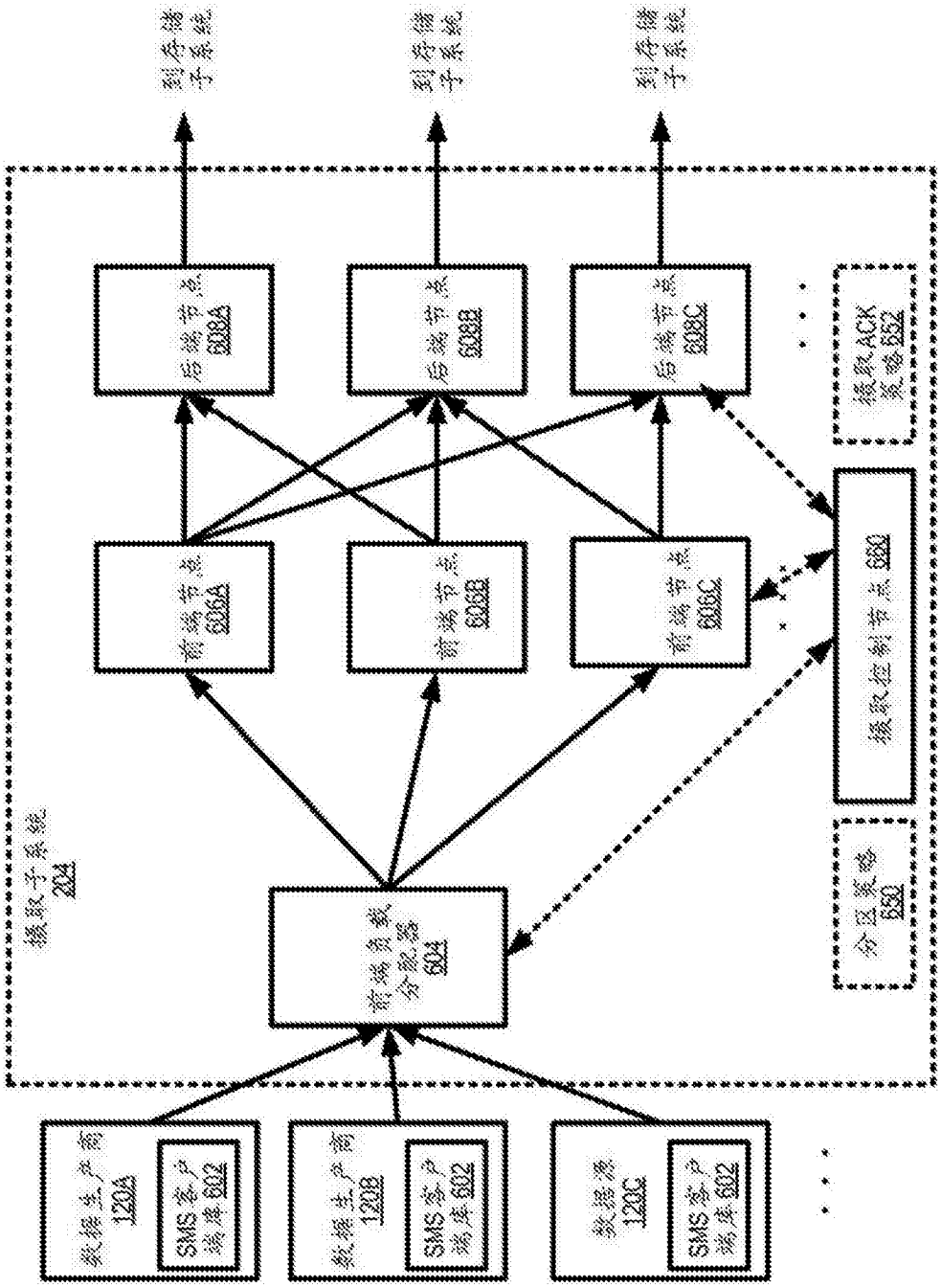

到分销子系统
到分销子系统
到分销子系统
后端节点 608A
后端节点 608B
后端节点 608C
接收ACK 实体 652
前端节点 606A
前端节点 606B
前端节点 606C
接收控制节点 650
接收子系统 204
前端负载分配器 604
分区实体 650
数据生产商 120A
SMS客户端库 602
数据生产商 120B
SMS客户端库 602
数据源 120C
SMS客户端库 602

图6

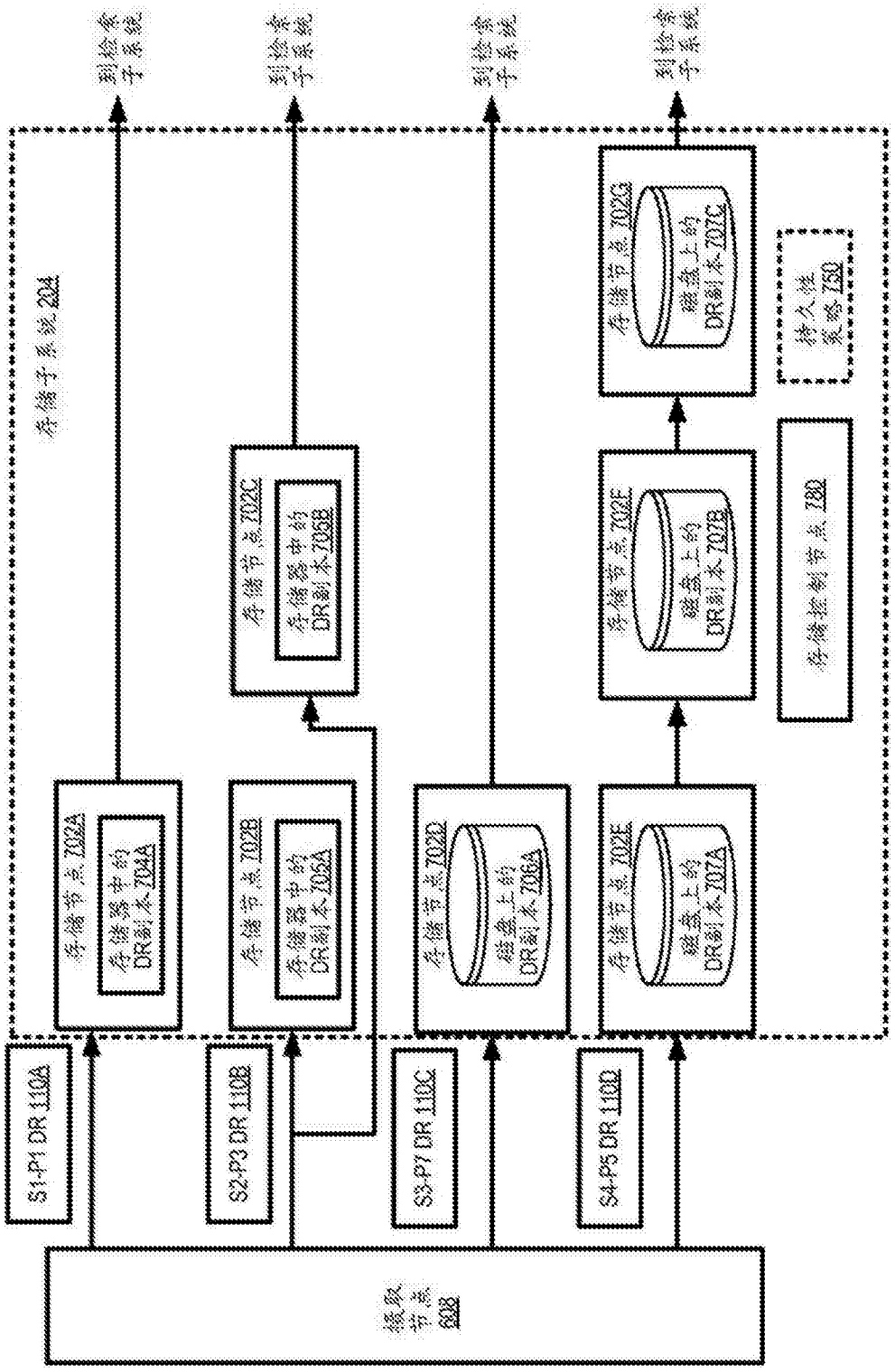

到检索子系统
到检索子系统
到检索子系统
到检索子系统
存储子系统 204
存储节点 702A
存储器中的 DR副本 704A
存储节点 702B
存储器中的 DR副本 706A
存储节点 702C
存储器中的 DR副本 706B
存储节点 702D
磁盘上的 DR副本 706A
存储节点 702E
磁盘上的 DR副本 707A
存储节点 702F
磁盘上的 DR副本 707B
存储节点 702G
磁盘上的 DR副本 707C
持久性策略 750
存储控制节点 780
S1-P1 DR 110A
S2-P3 DR 110B
S3-P7 DR 110C
S4-P5 DR 110D
检索节点 820

图7

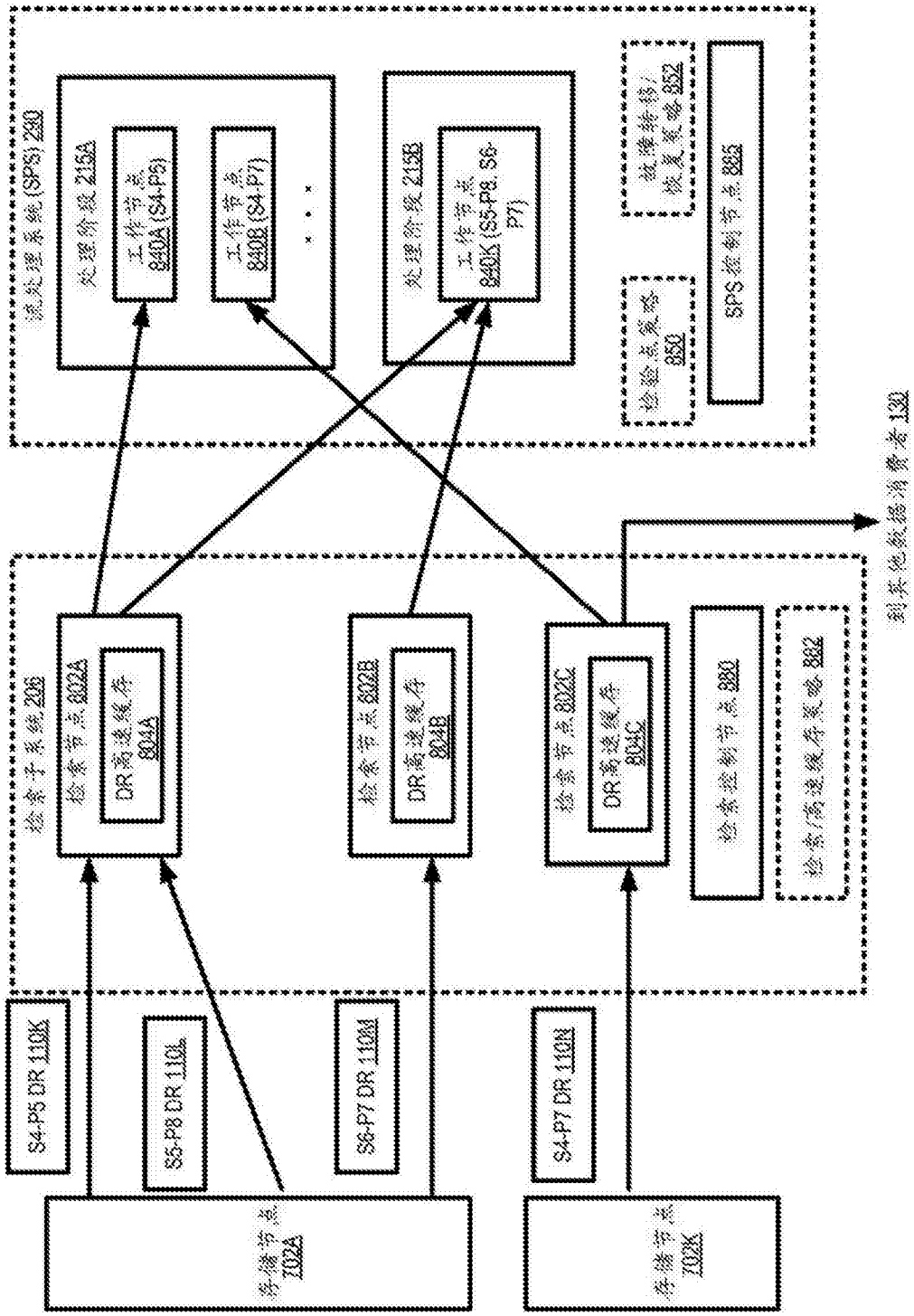

流处理系统(SPS) 230
处理阶段 215A
工作节点 840A (S4-P5)
工作节点 840B (S4-P7)
处理阶段 215B
工作节点 840K (S5-P8, S6-P7)
故障转移/恢复策略 852
检验点策略 850
SPS控制节点 825
检索子系统 206
检索节点 802A
DR高速缓存 804A
检索节点 802B
DR高速缓存 804B
检索节点 802C
DR高速缓存 804C
检索控制节点 880
检索/高速缓存策略 882
到其他数据消费者 130
S4-P5 DR 110K
S5-P8 DR 110L
S6-P7 DR 110M
S4-P7 DR 110N
存储节点 102A
存储节点 102K

图8

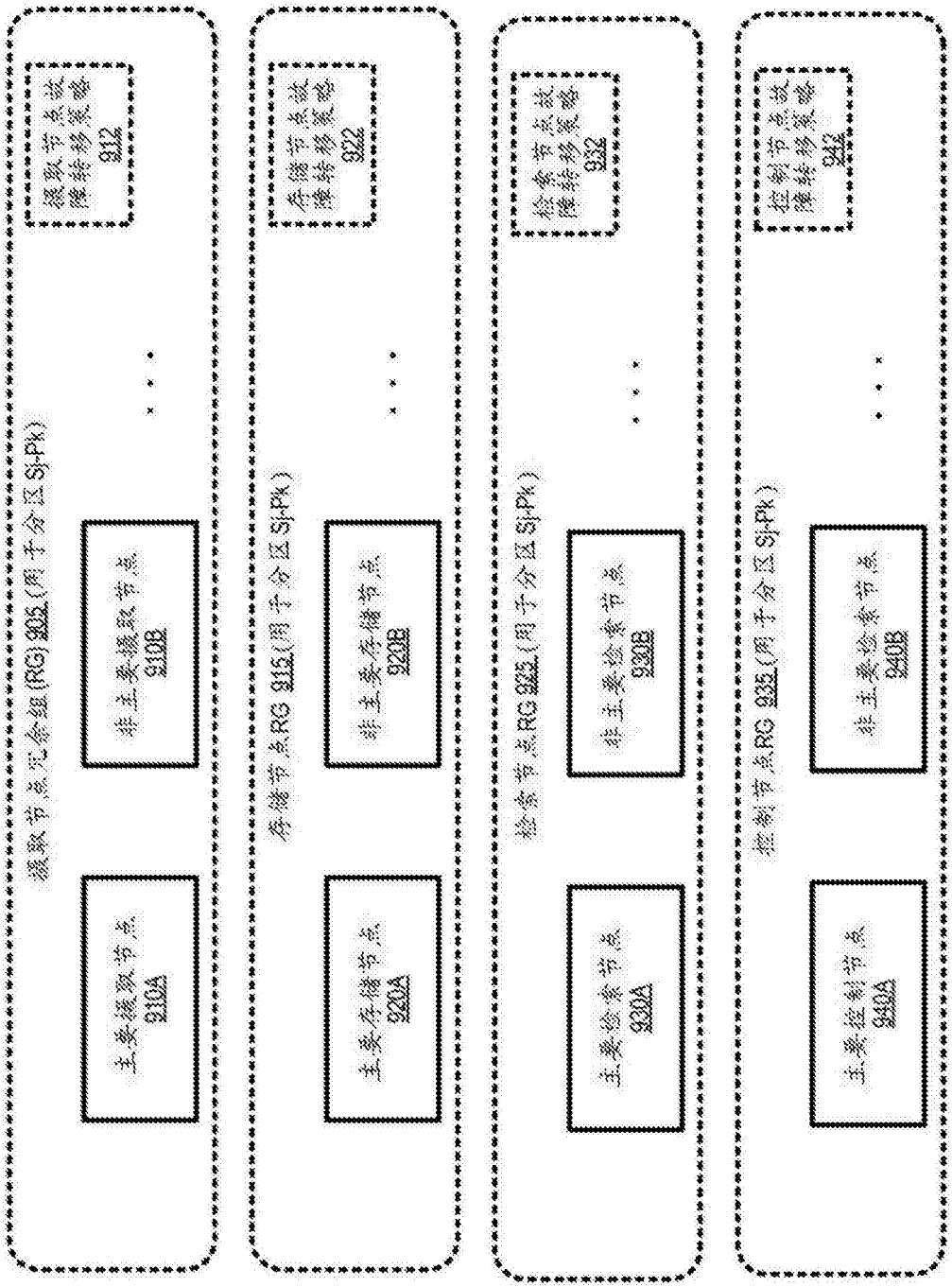

提取节点冗余组(RG) 905(用于分区 Sj-Pk)
主要提取节点 910A
非主要提取节点 910B
提取节点故障转移策略 912
存储节点RG 915(用于分区 Sj-Pk)
主要存储节点 920A
非主要存储节点 920B
存储节点故障转移策略 922
检索节点RG 925(用于分区 Sj-Pk)
主要检索节点 930A
非主要检索节点 930B
检索节点故障转移策略 932
控制节点RG 935(用于分区 Sj-Pk)
主要控制节点 940A
非主要检索节点 940B
控制节点故障转移策略 942

图9

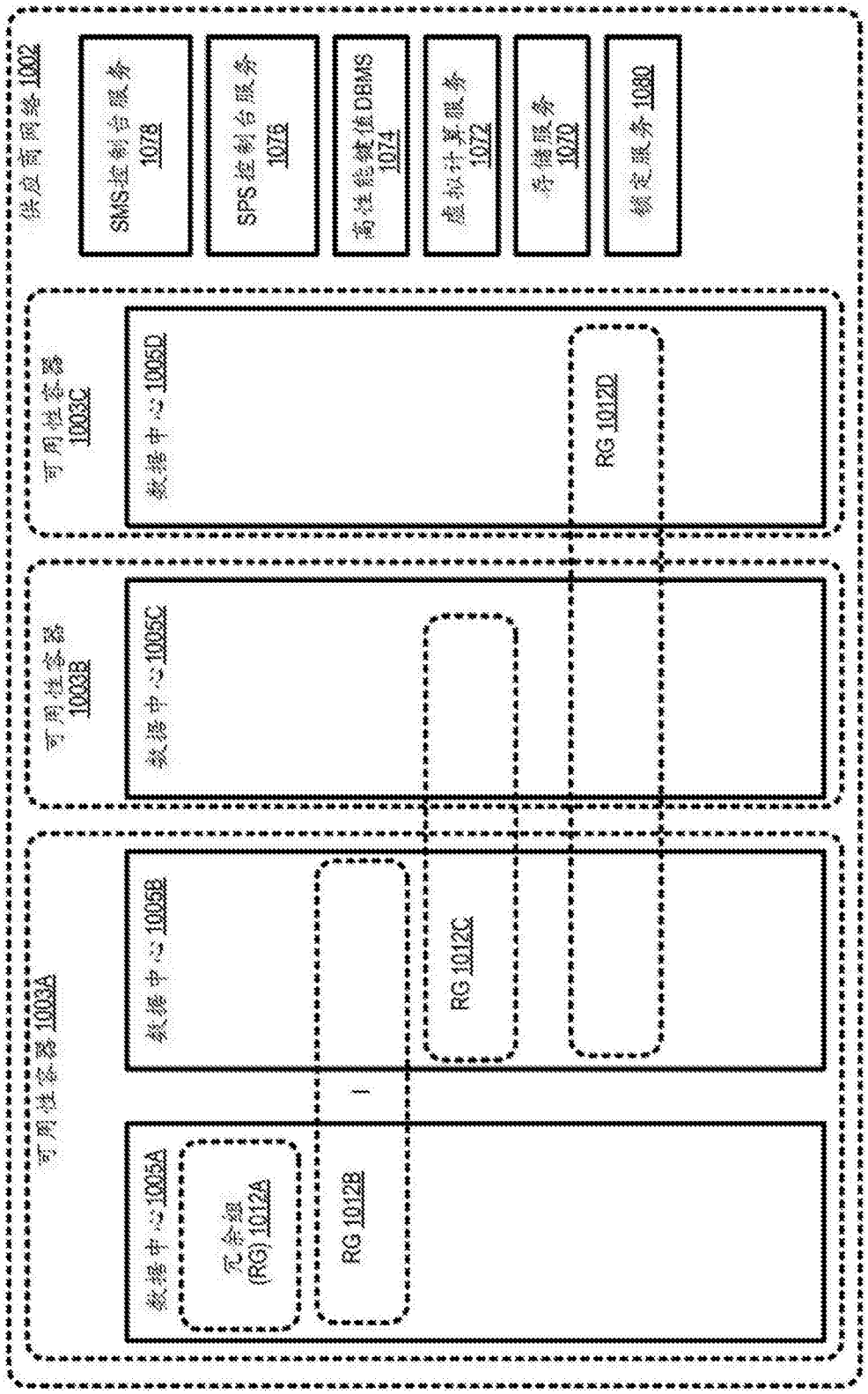

供应商网络 1002
SMS控制台服务 1078
SPS控制台服务 1076
高性能键值DBMS 1074
虚拟计算服务 1072
存储服务 1070
锁定服务 1080
可用性容器 1003C
数据中心 1005D
RG 1012D
可用性容器 1003B
数据中心 1005C
RG 1012C
可用性容器 1003A
数据中心 1005B
数据中心 1005A
冗余组 (RG) 1012A
RG 1012B

图10

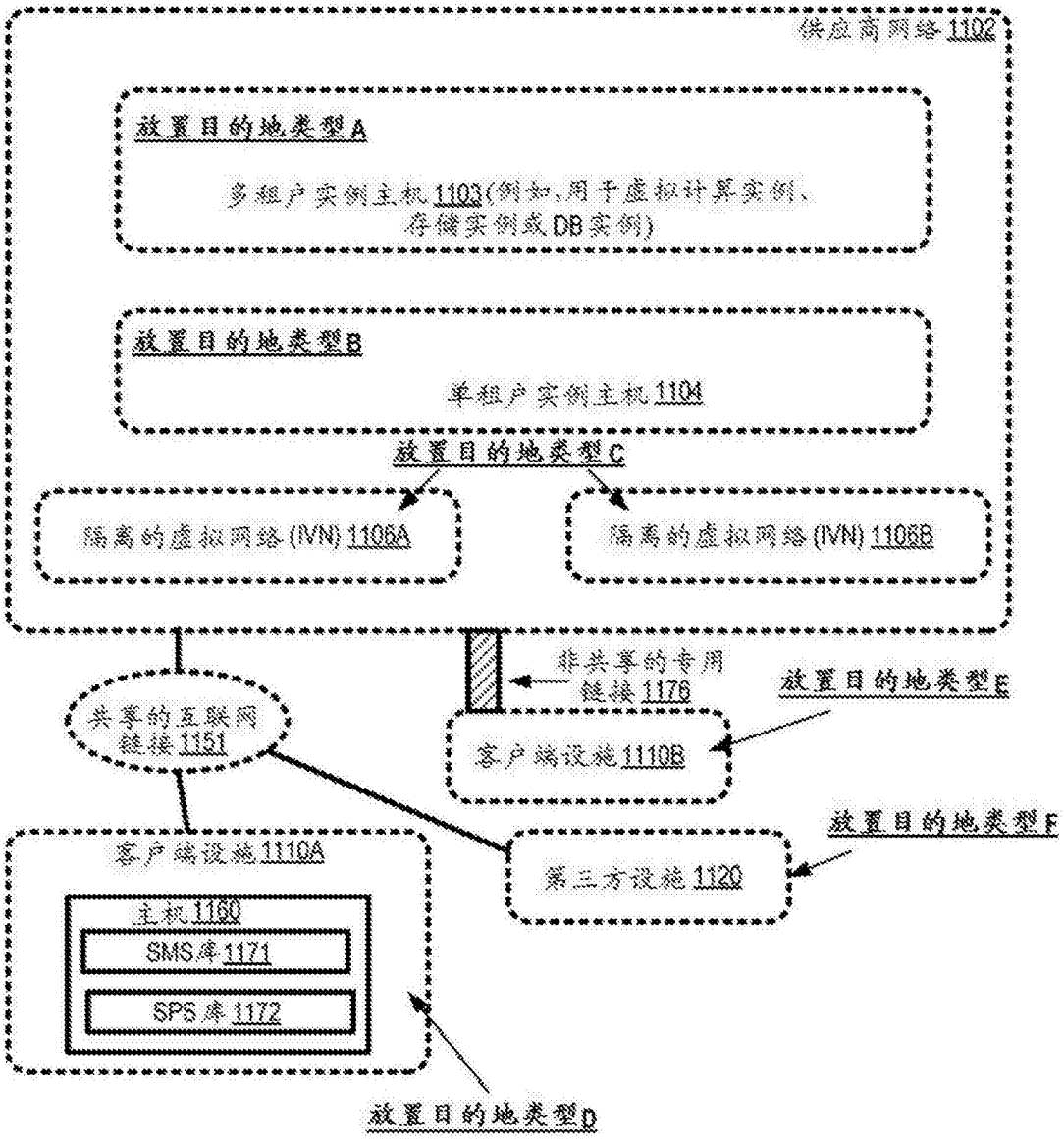

供应商网络 1102
放置目的地类型A
多租户实例主机 1103(例如,用于虚拟计算实例、存储实例或DB实例)
放置目的地类型B
单租户实例主机 1104
放置目的地类型C
隔离的虚拟网络 (IVN) 1106A
隔离的虚拟网络 (IVN) 1106B
非共享的专用链接 1176
放置目的地类型E
共享的互联网链接 1151
客户端设施 1110B
放置目的地类型F
客户端设施 1110A
第三方设施 1120
主机 1160
SMS库 1171
SPS库 1172
放置目的地类型D

图11

图12a

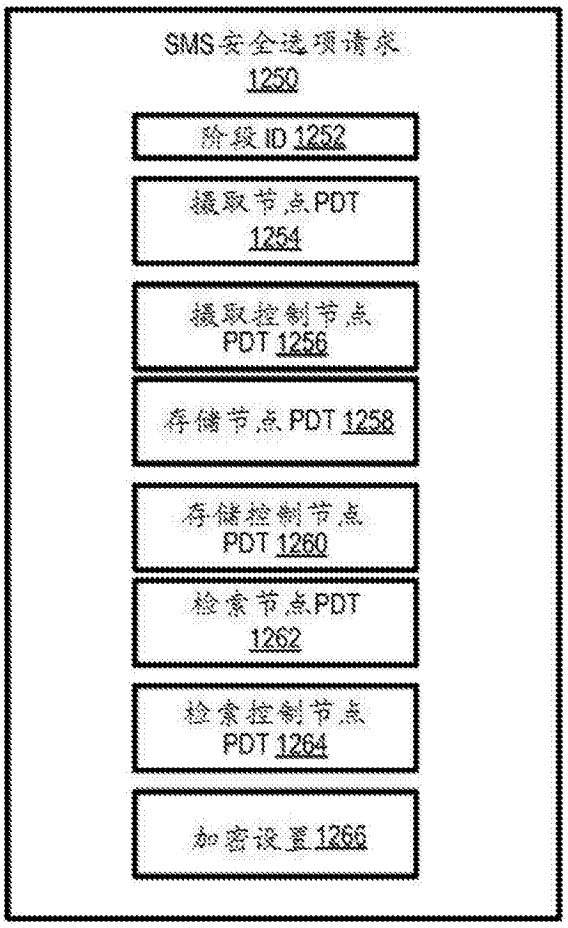


图12b

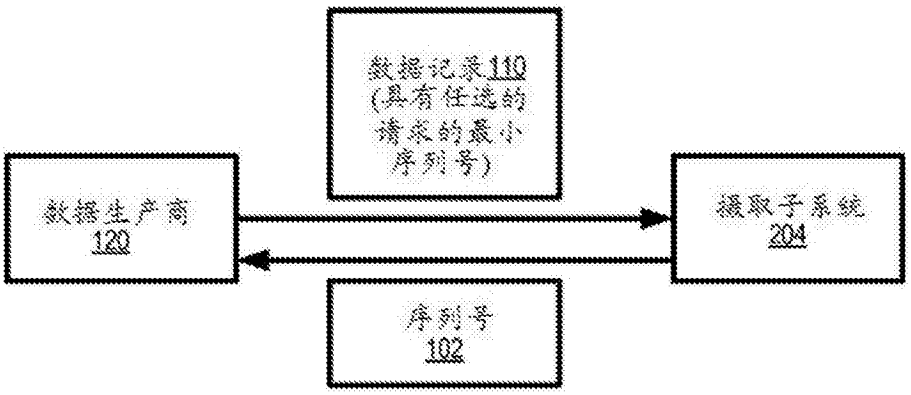


图13a

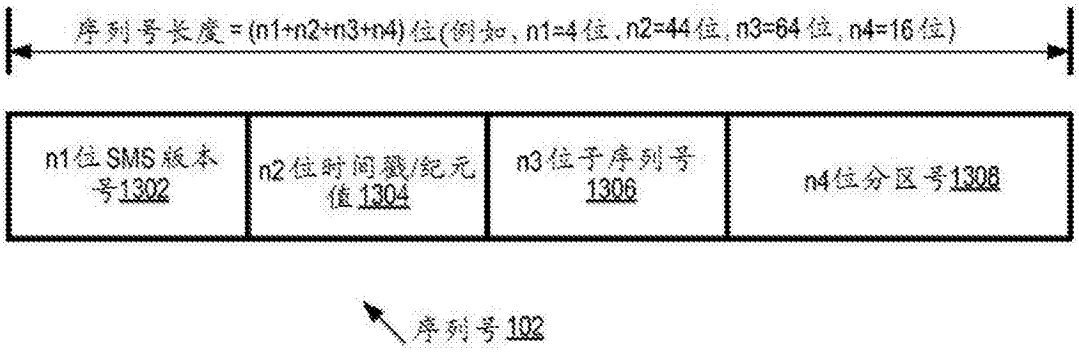

序列号长度=(n1+n2+n3+n4)位(例如，n1=4位，n2=44位，n3=64位，n4=16位)
n1位SMS版本号1302
n2位时间戳/纪元值1304
n3位序列号1306
n4位分区号1308
序列号102

图13b

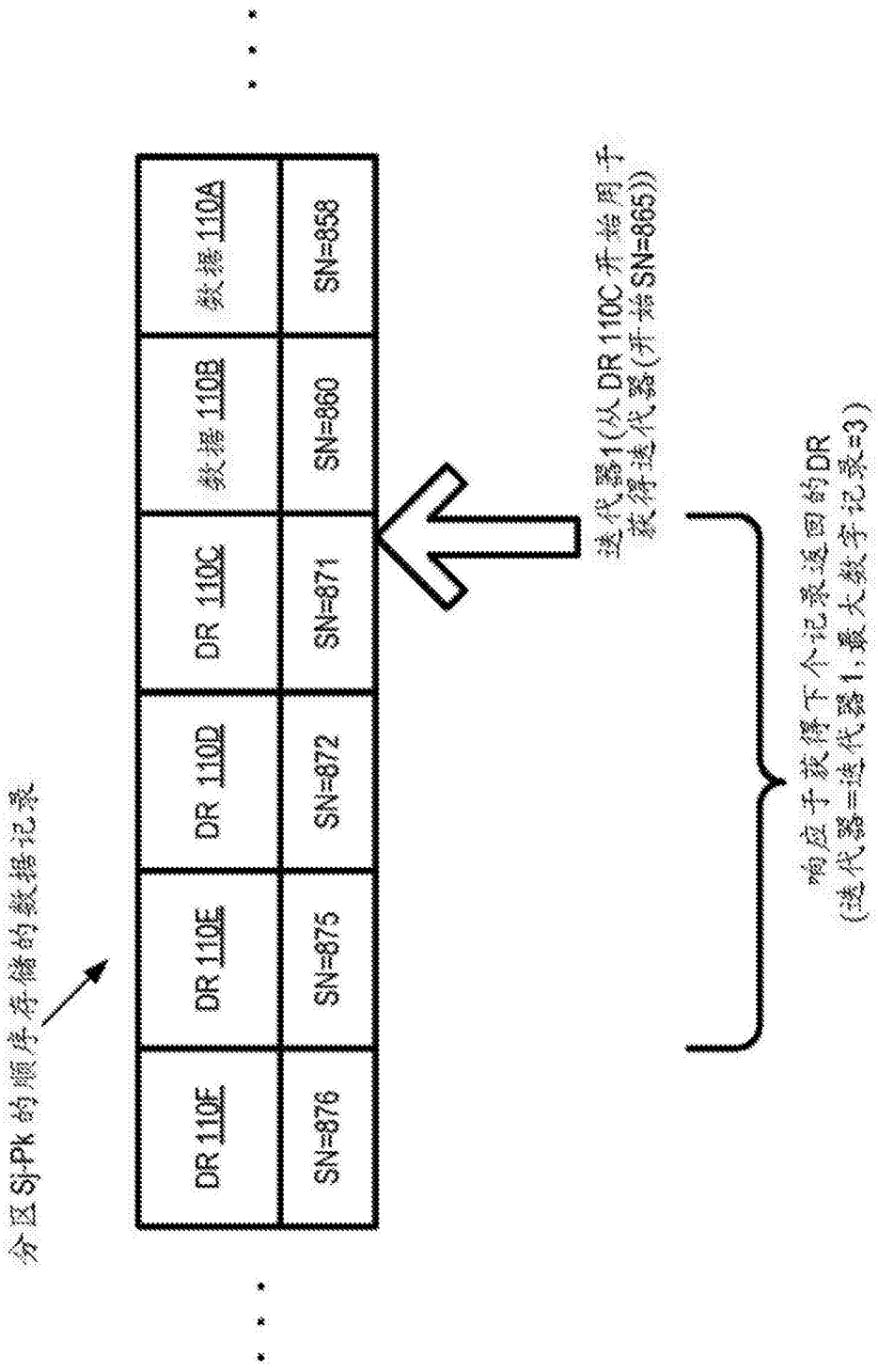

分区 SI-PK 的顺序存储的数据记录
DR 110E
SN=876
DR 110E
SN=875
DR 110D
SN=872
DR 110C
SN=871
数据 110B
SN=860
数据 110A
SN=858
迭代器1(从 DR 110C 开始用于获得迭代器(开始SN=865))
响应于获得下个记录返回的DR
(迭代器=迭代器1,最大数字记录=3)

图14

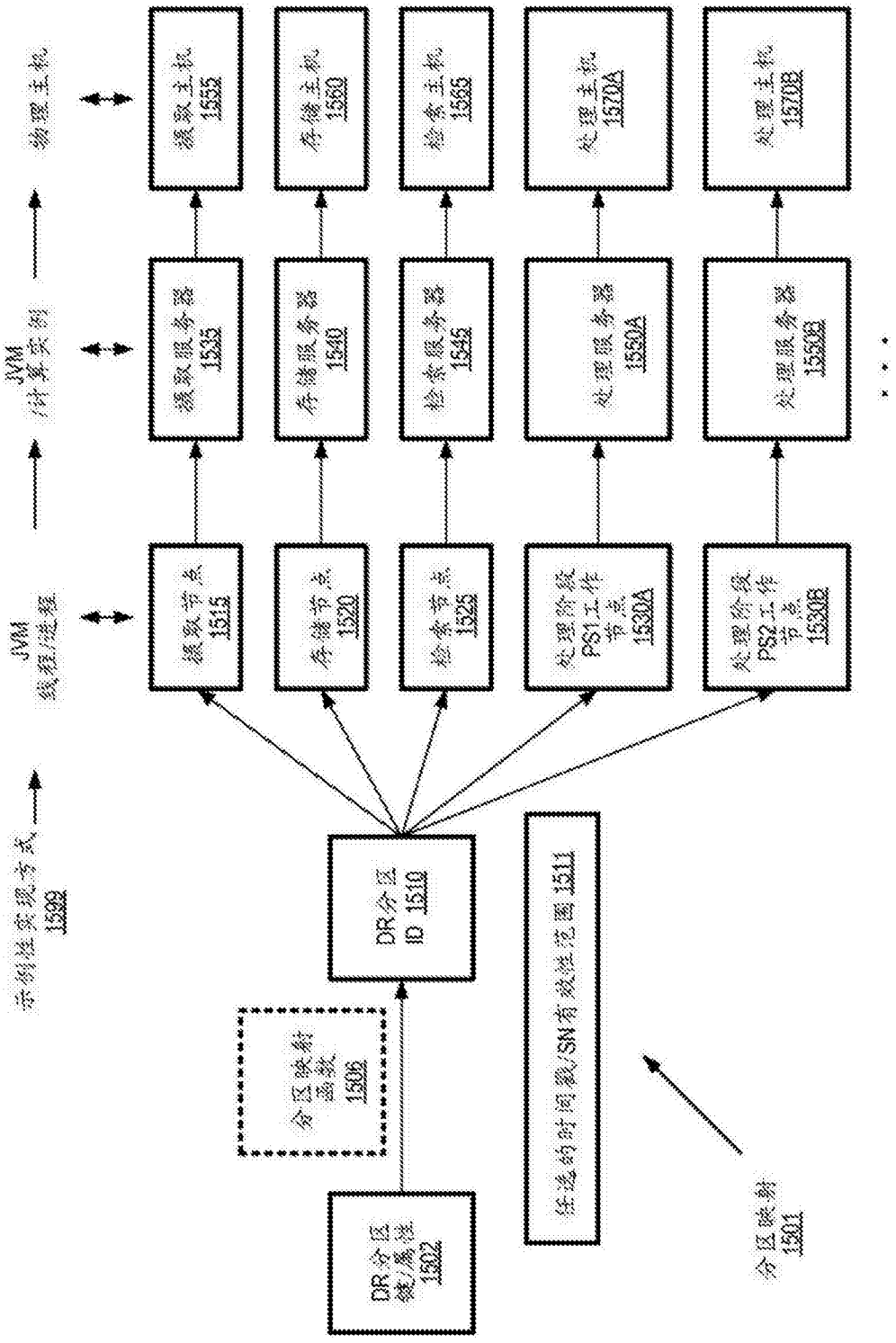

物理主机
JVM
计算实例
JVM
线程/进程
示例性实现方式
1599
获取主机
1555
存储主机
1560
检索主机
1565
处理主机
1570A
处理主机
1570B
获取服务器
1535
存储服务器
1540
检索服务器
1545
处理服务器
1550A
处理服务器
1550B
获取节点
1515
存储节点
1520
检索节点
1525
处理阶段
PS1工作
节点
1530A
处理阶段
PS2工作
节点
1530B
DR分区
ID 1510
分区映射
函数
1506
DR分区
键/属性
1502
传送的时间戳/SN有效性范围 1511
分区映射
1501

图15

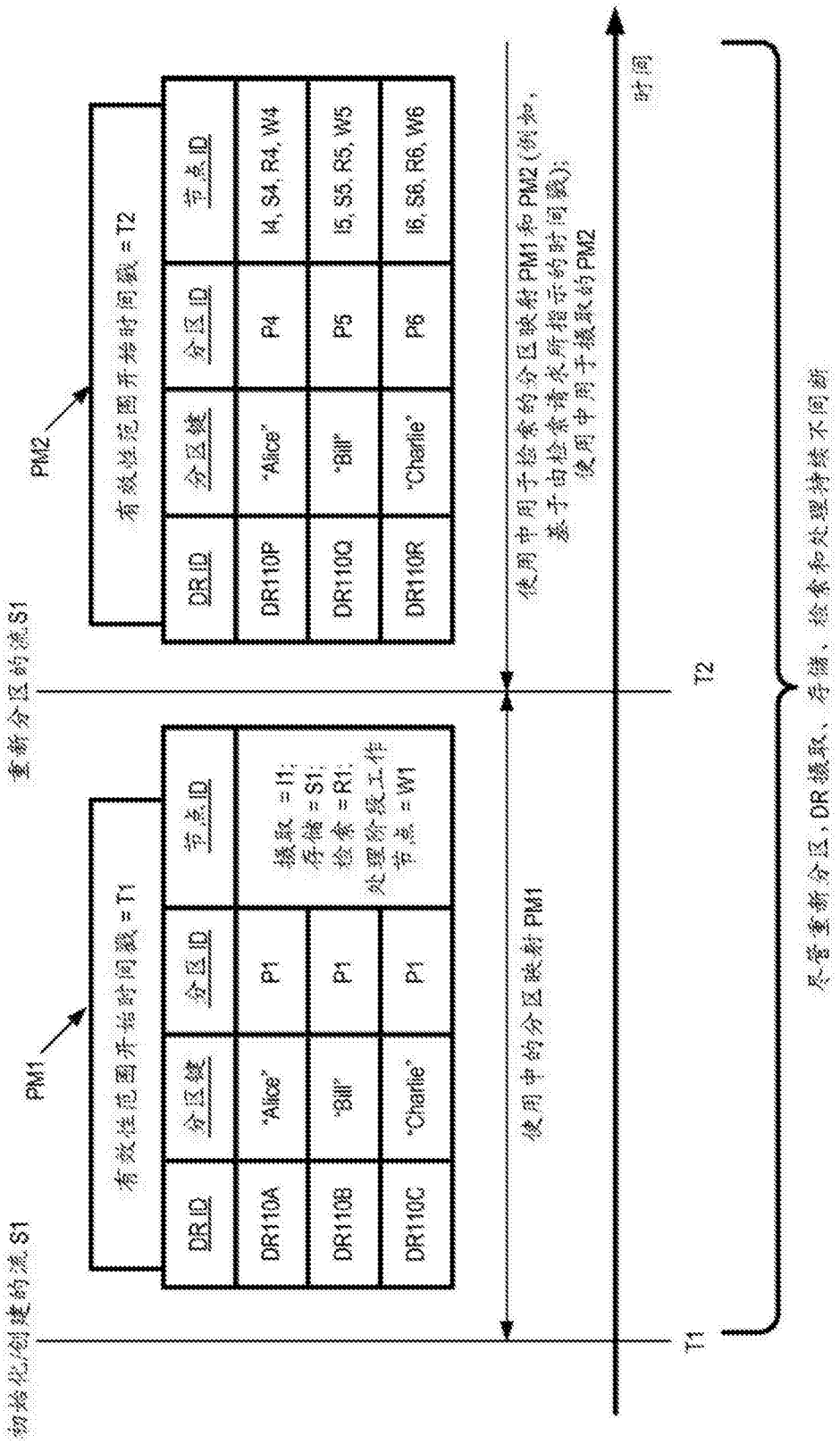

初始化创建的流 S1
PM1
有效性范围开始时间戳 = T1
DR ID
分区键
分区 ID
节点 ID
DR110A
"Alice"
P1
DR110B
"Bill"
P1
DR110C
"Charlie"
P1
提取 = I1;
存储 = S1;
检索 = R1;
处理阶级工作
节点 = W1
使用中的分区映射 PM1
T1
重新分区的流 S1
PM2
有效性范围开始时间戳 = T2
DR110P
"Alice"
P4
I4, S4, R4, W4
DR110Q
"Bill"
P5
I5, S5, R5, W5
DR110R
"Charlie"
P6
I6, S6, R6, W6
使用中用于检索的分区映射 PM1 和 PM2 (例如，
基于由检索请求所指示的时间戳);
使用中用于提取的 PM2
T2
时间
尽管重新分区，DR 提取、存储、检索和处理持续不间断

图16

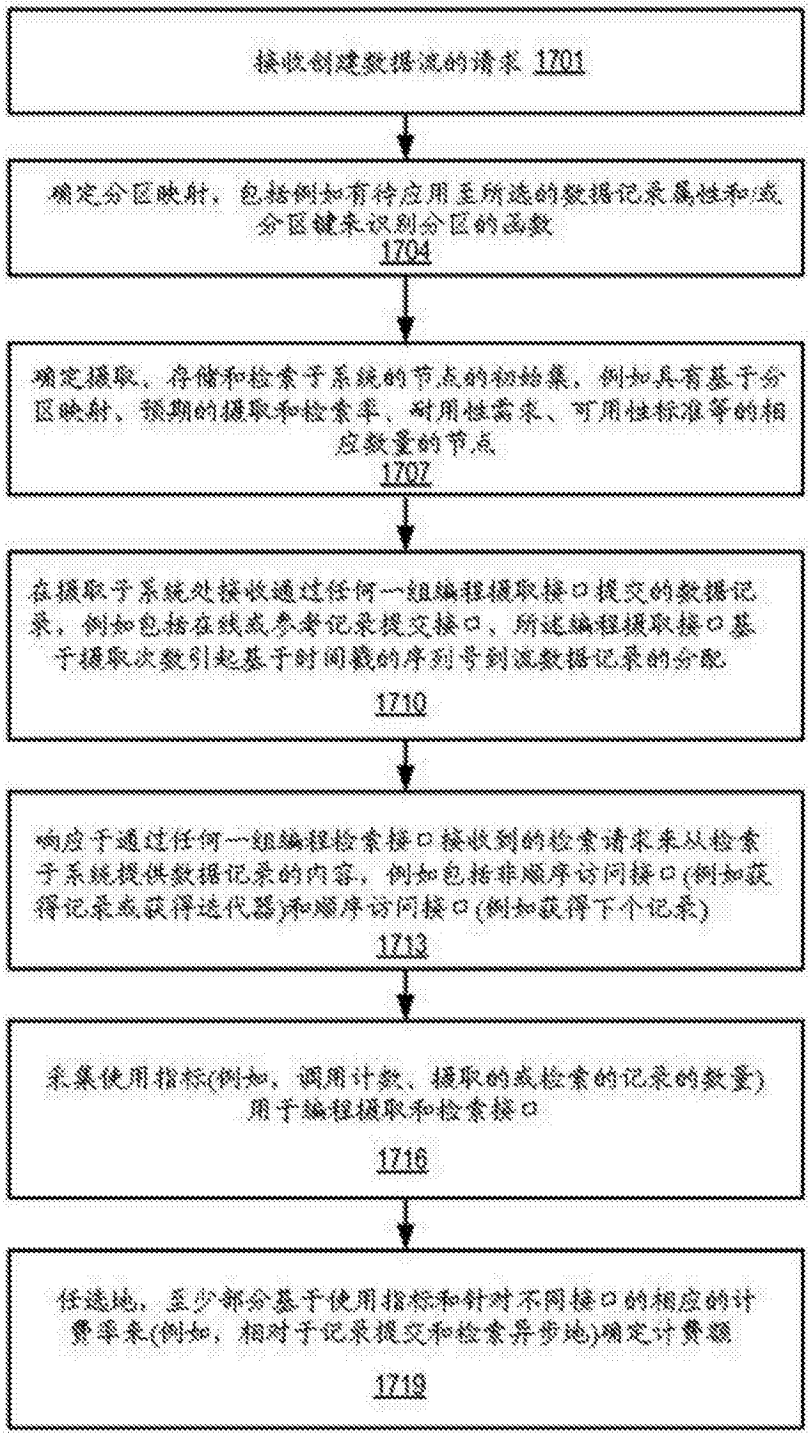

接收创建数据流的请求 1701
确定分区映射，包括例如有待应用至所选的数据记录属性和/或分区键来识别分区的函数 1704
确定提取、存储和检索于系统的节点的初始集，例如具有基于分区映射、预期的提取和检索率、耐用性需求、可用性标准等的相应数量的节点 1707
在提取于系统处接收通过任何一组编程提取接口提交的数据记录，例如包括在线或参考记录提交接口、所述编程提取接口基于提取次数引起基于时间戳的序列号到达数据记录的分配 1710
响应于通过任何一组编程检索接口接收到的检索请求来从检索于系统提供数据记录的内容，例如包括非顺序访问接口(例如获得记录或获得迭代器)和顺序访问接口(例如获得下个记录) 1713
采集使用指标(例如，调用计数、提取的或检索的记录的数量)用于编程提取和检索接口 1716
任选地，至少部分基于使用指标和针对不同接口的相应的计费率来(例如，相对于记录提交和检索异步地)确定计费额 1719

图17

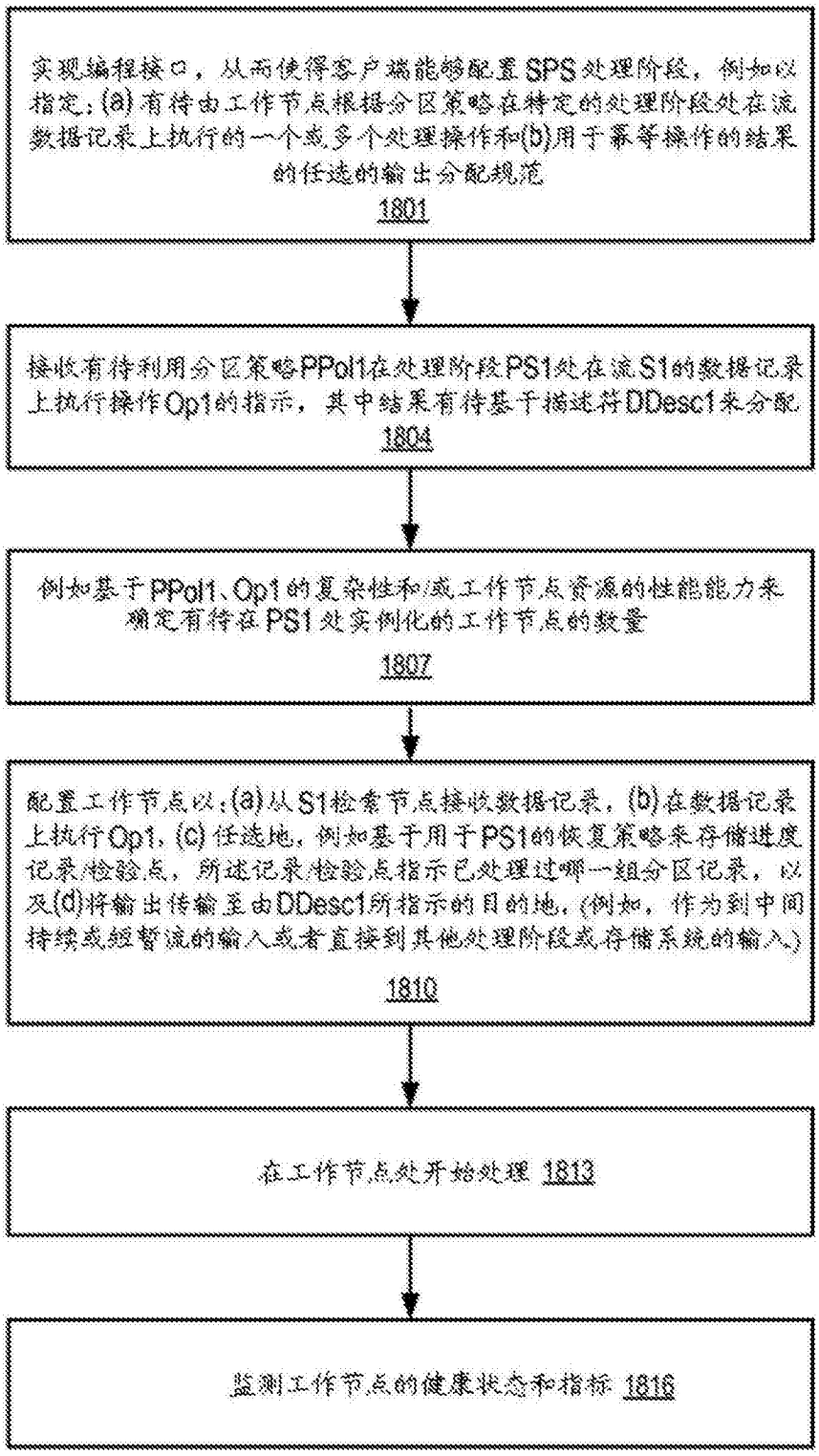

实现编程接口，从而使得客户端能够配置SPS处理阶段，例如以指定：(a)有待由工作节点根据分区策略在特定的处理阶段处在流数据记录上执行的一个或多个处理操作和(b)用于幂等操作的结果的任选的输出分配规范
1801
接收有待利用分区策略PPol1在处理阶段PS1处在流S1的数据记录上执行操作Op1的指示，其中结果有待基于描述符DDesc1来分配
1804
例如基于PPol1、Op1的复杂性和/或工作节点资源的性能能力来确定有待在PS1处实例化的工作节点的数量
1807
配置工作节点以：(a)从S1检索节点接收数据记录，(b)在数据记录上执行Op1，(c)任选地，例如基于用于PS1的恢复策略来存储进度记录/检验点，所述记录/检验点指示已处理过哪一组分区记录，以及(d)将输出传输至由DDesc1所指示的目的地，(例如，作为到中间持续或短暂流的输入或者直接到其他处理阶段或存储系统的输入)
1810
在工作节点处开始处理 1813
监测工作节点的健康状态和指标 1816

图18a

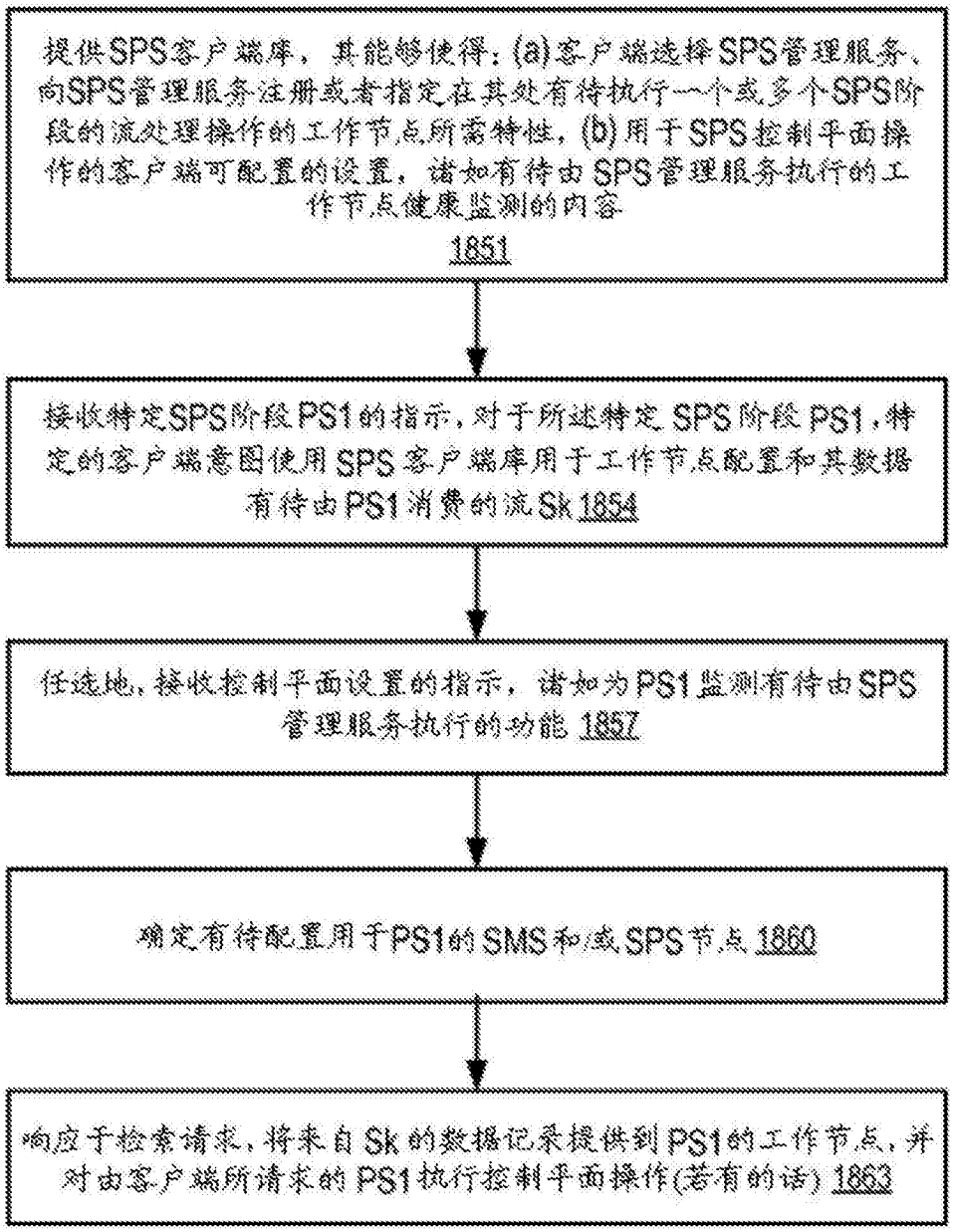

提供SPS客户端库，其能够使得：(a)客户端选择SPS管理服务、向SPS管理服务注册或者指定在其处有待执行一个或多个SPS阶段的流处理操作的工作节点所需特性，(b)用于SPS控制平面操作的客户端可配置的设置，诸如有待由SPS管理服务执行的工作节点健康监测的内容
1851
接收特定SPS阶段PS1的指示，对于所述特定SPS阶段PS1，特定的客户端意图使用SPS客户端库用于工作节点配置和其数据有待由PS1消费的流Sk 1854
任选地，接收控制平面设置的指示，诸如为PS1监测有待由SPS管理服务执行的功能 1857
确定有待配置用于PS1的SMS和/或SPS节点 1860
响应于检索请求，将来自Sk的数据记录提供到PS1的工作节点，并对由客户端所请求的PS1执行控制平面操作(若有的话) 1863

图18b

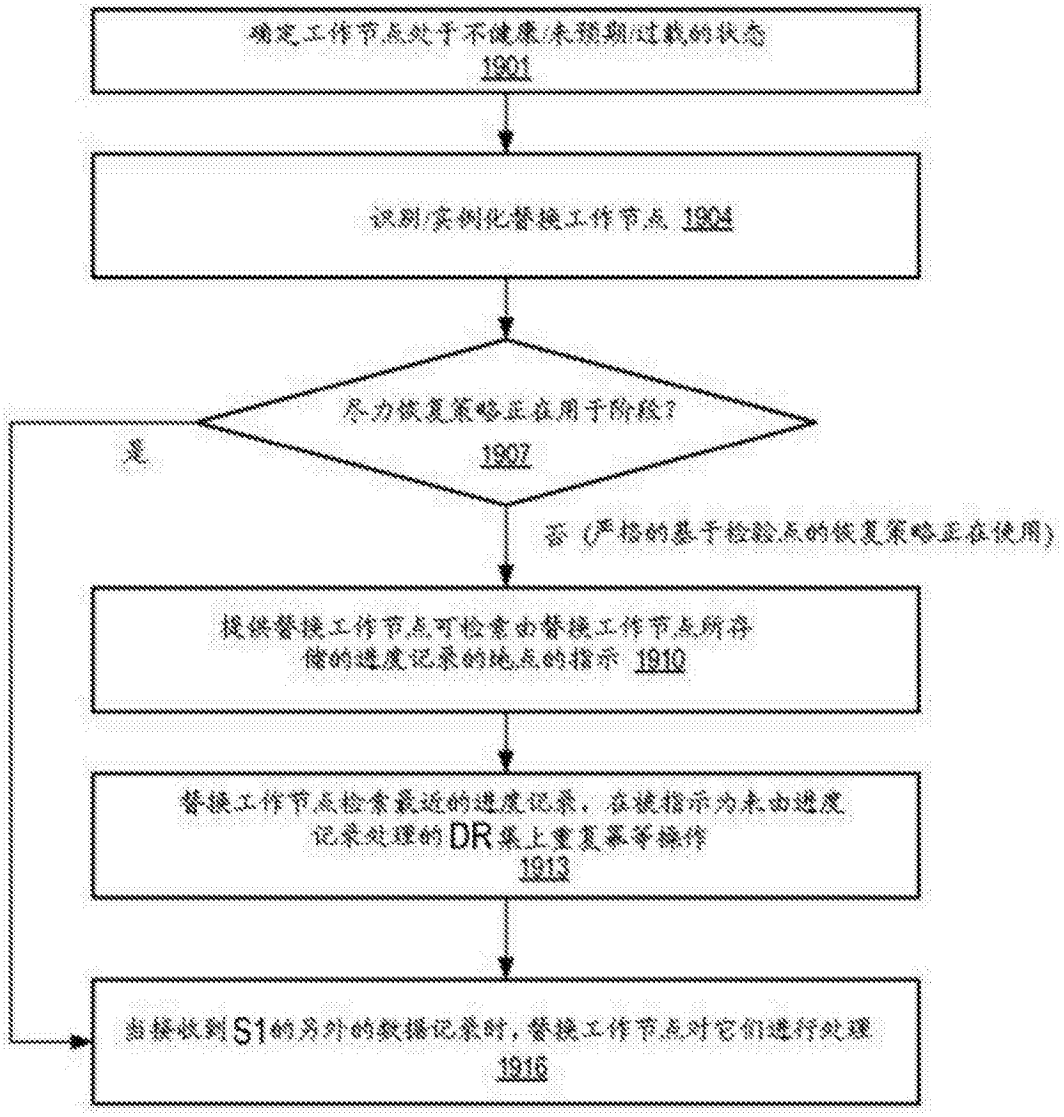

确定工作节点处于不健康/未预期/过载的状态
1901
识别/实例化替换工作节点 1904
尽力恢复策略正在用于阶段?
1907
是
否 (严格的基于检验点的恢复策略正在使用)
提供替换工作节点可检查由替换工作节点所存储的速度记录的地点的指示 1910
替换工作节点检查最近的速度记录，直到指示为未由速度记录处理的DR集上重复等操作
1913
当接收到S1的另外的数据记录时，替换工作节点对它们进行处理
1916

图19

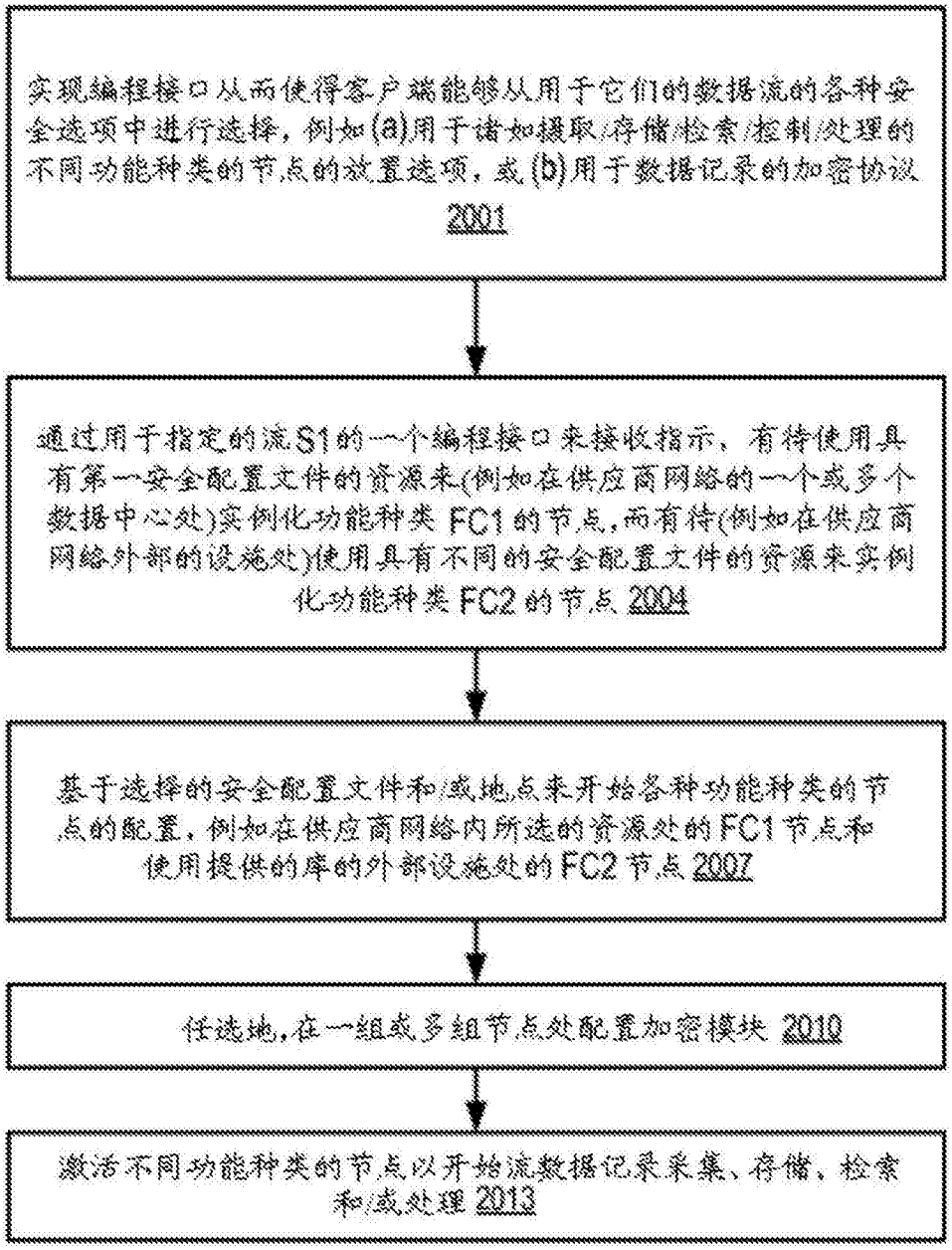

实现编程接口从而使得客户端能够从用于它们的数据流的各种安全选项中进行选择，例如(a)用于诸如提取/存储/检索/控制/处理的不同功能种类的节点的放置选项，或(b)用于数据记录的加密协议 2001
通过用于指定的流S1的一个编程接口来接收指示，有待使用具有第一安全配置文件的资源来(例如在供应商网络的一个或多个数据中心处)实例化功能种类 FC1 的节点，而有待(例如在供应商网络外部的设施处)使用具有不同的安全配置文件的资源来实例化功能种类 FC2 的节点 2004
基于选择的安全配置文件和/或地点来开始各种功能种类的节点的配置，例如在供应商网络内所选的资源处的 FC1 节点和使用提供的库的外部设施处的 FC2 节点 2007
任选地，在一组或多组节点处配置加密模块 2010
激活不同功能种类的节点以开始流数据记录采集、存储、检索和/或处理 2013

图20

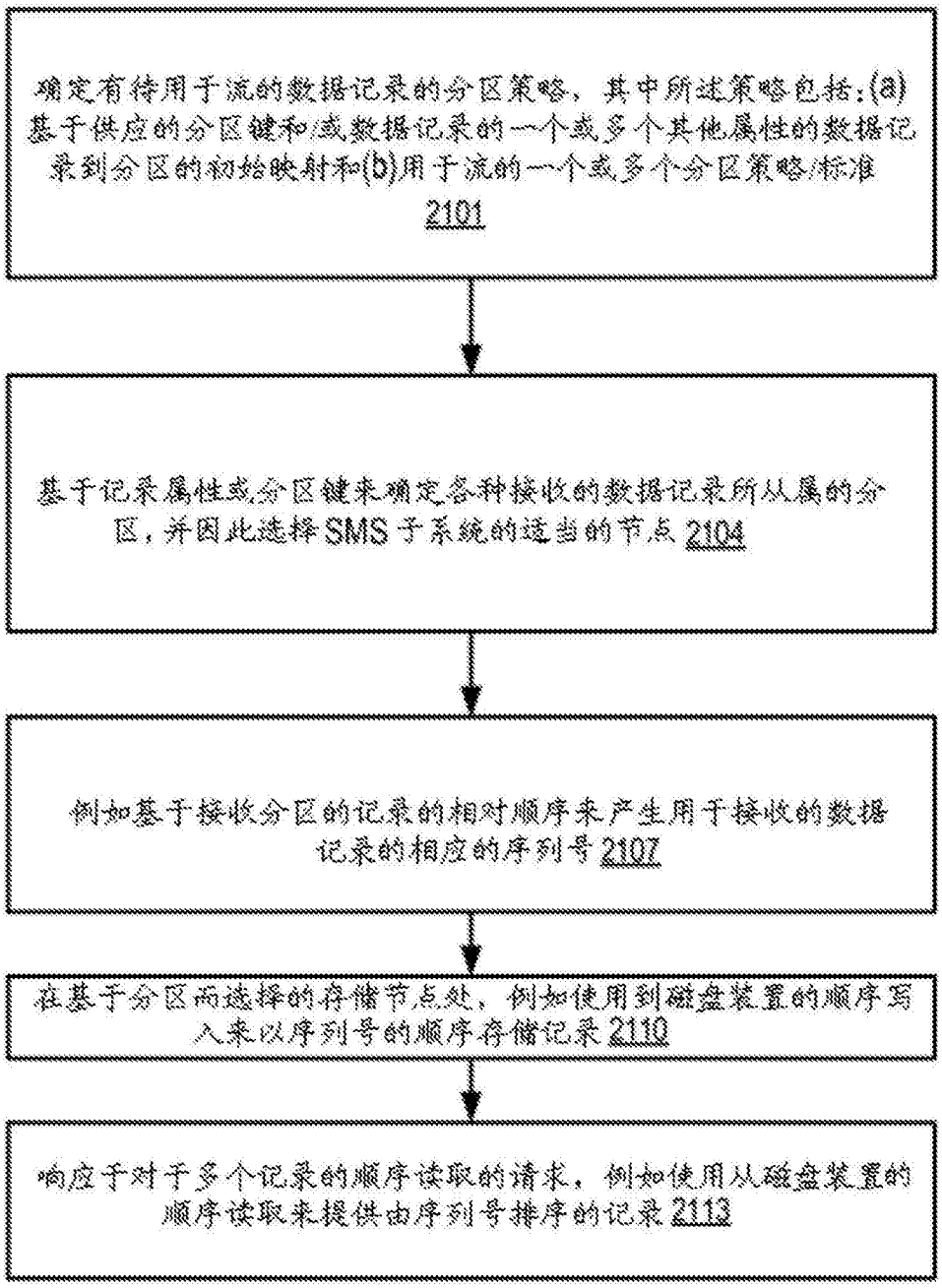

确定有待用于流的数据记录的分区策略，其中所述策略包括:(a)基于供应的分区键和/或数据记录的一个或多个其他属性的数据记录到分区的初始映射和(b)用于流的一个或多个分区策略/标准 2101
基于记录属性或分区键来确定各种接收的数据记录所从属的分区，并因此选择SMS子系统的适当的节点 2104
例如基于接收分区的记录的相对顺序来产生用于接收的数据记录的相应的序列号 2107
在基于分区而选择的存储节点处，例如使用到磁盘装置的顺序写入来以序列号的顺序存储记录 2110
响应于对于多个记录的顺序读取的请求，例如使用从磁盘装置的顺序读取来提供由序列号排序的记录 2113

图21

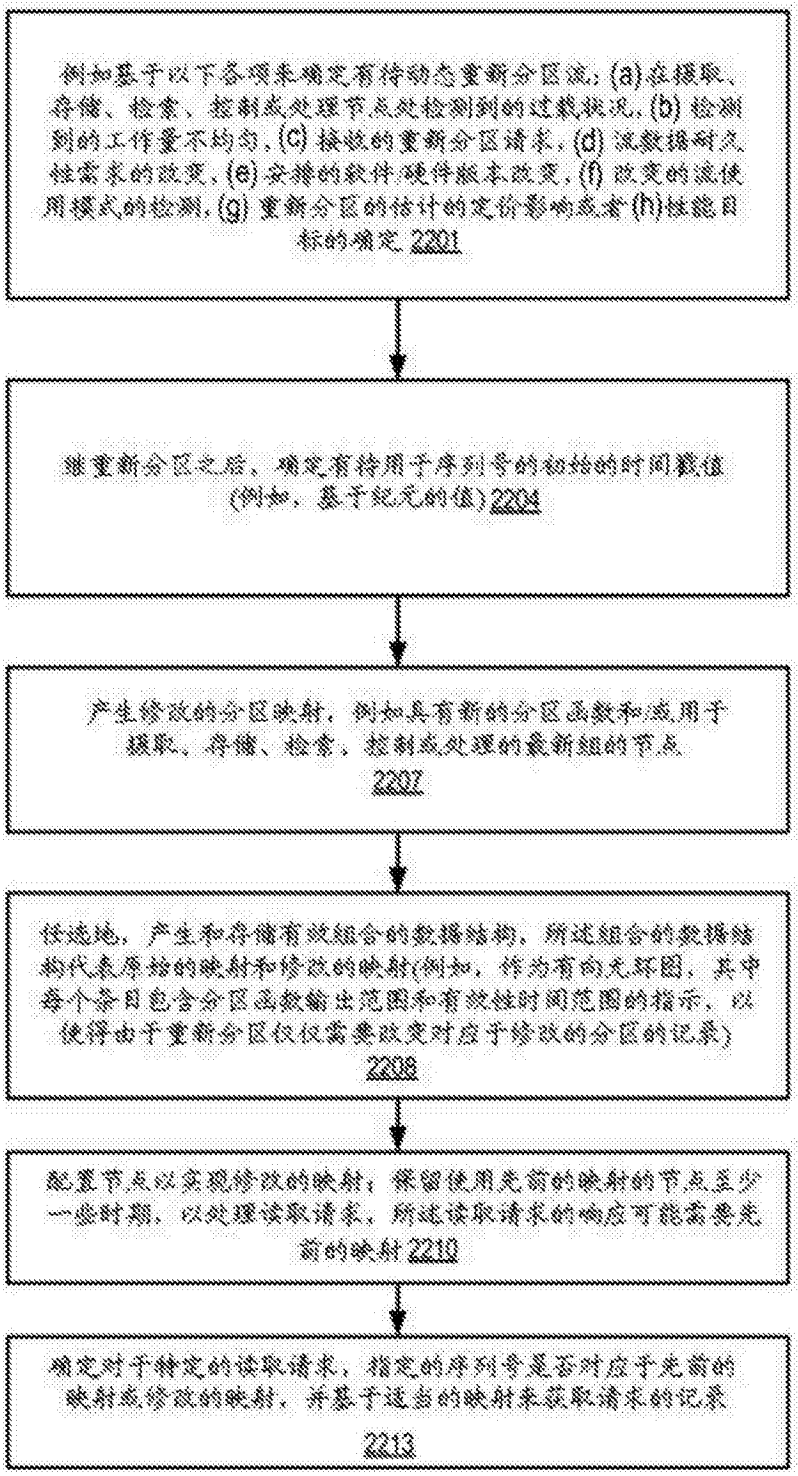

例如基于以下各项来确定有待动态重新分区或：(a)在提取、存储、检索、控制或处理节点处检测到的过载状况，(b)检测到的工作量不均匀，(c)接收的重新分区请求，(d)流数据耐久性需求的改变，(e)安排的软件/硬件版本改变，(f)改变的流使用模式的检测，(g)重新分区的估计的定价影响或者(h)性能目标的确定 2201
继重新分区之后，确定有待用于序列号的初始的时间戳值(例如，基于纪元的值) 2204
产生修改的分区映射，例如具有新的分区函数和应用于提取、存储、检索，控制或处理的最新组的节点 2207
任选地，产生和存储有效组合的数据结构，所述组合的数据结构代表原始的映射和修改的映射(例如，作为有向无环图，其中每个条目包含分区函数输出范围和有效性时间范围的指示，以使得由于重新分区仅仅需要改变对应于修改的分区的记录) 2208
配置节点以实现修改的映射；保留使用先前的映射的节点至少一些时期，以处理读取请求，所述读取请求的响应可能需要先前的映射 2210
确定对于特定的读取请求，指定的序列号是否对应于先前的映射或修改的映射，并基于适当的映射来获取请求的记录 2213

图22

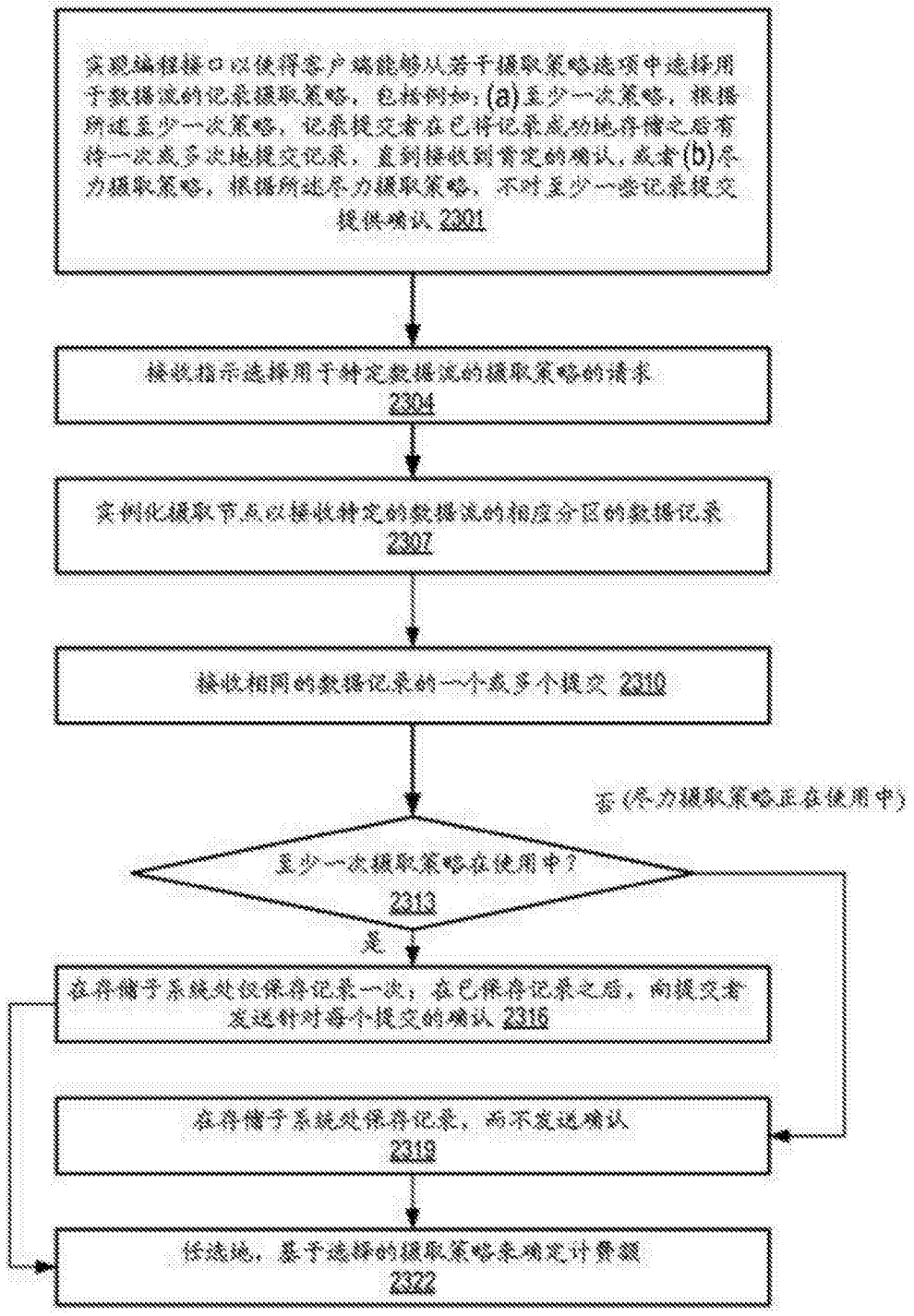

实现编程接口以使得客户端能够从若干提取策略选项中选择用于数据流的记录提取策略，包括例如：(a)至少一次策略，根据所述至少一次策略，记录提交者在已将记录成功地存储之后有待一次或多次地提交记录，直到接收到肯定的确认，或者(b)尽力提取策略，根据所述尽力提取策略，不对至少一些记录提交提供确认 2301
接收指示选择用于特定数据流的提取策略的请求 2304
实例化提取节点以接收特定的数据流的相应分区的数据记录 2307
接收相同的数据记录的一个或多个提交 2310
否(尽力提取策略正在使用中)
至少一次提取策略在使用中？ 2313
是
在存储子系统处保存记录一次；在已保存记录之后，向提交者发送针对每个提交的确认 2316
在存储子系统处保存记录，而不发送确认 2319
任选地，基于选择的提取策略来确定计费额 2322

图23

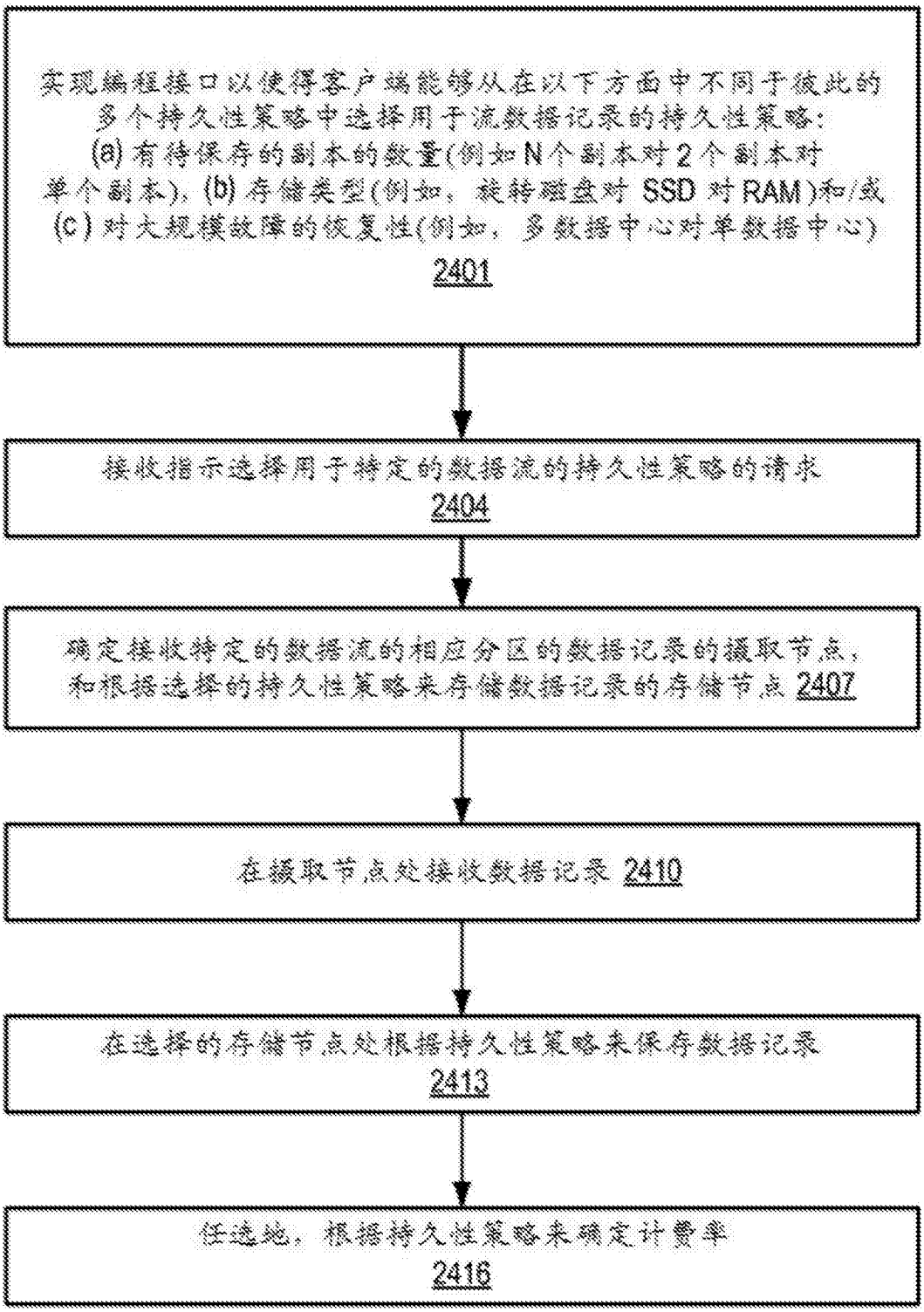

实现编程接口以使得客户端能够从在以下方面中不同于彼此的多个持久性策略中选择用于流数据记录的持久性策略：(a)有待保存的副本的数量(例如N个副本对2个副本对单个副本)，(b)存储类型(例如，旋转磁盘对SSD对RAM)和/或(c)对大规模故障的恢复性(例如，多数据中心对单数据中心)
2401
接收指示选择用于特定的数据流的持久性策略的请求
2404
确定接收特定的数据流的相应分区的数据记录的摄取节点，和根据选择的持久性策略来存储数据记录的存储节点 2407
在摄取节点处接收数据记录 2410
在选择的存储节点处根据持久性策略来保存数据记录
2413
任选地，根据持久性策略来确定计费率
2416

图24

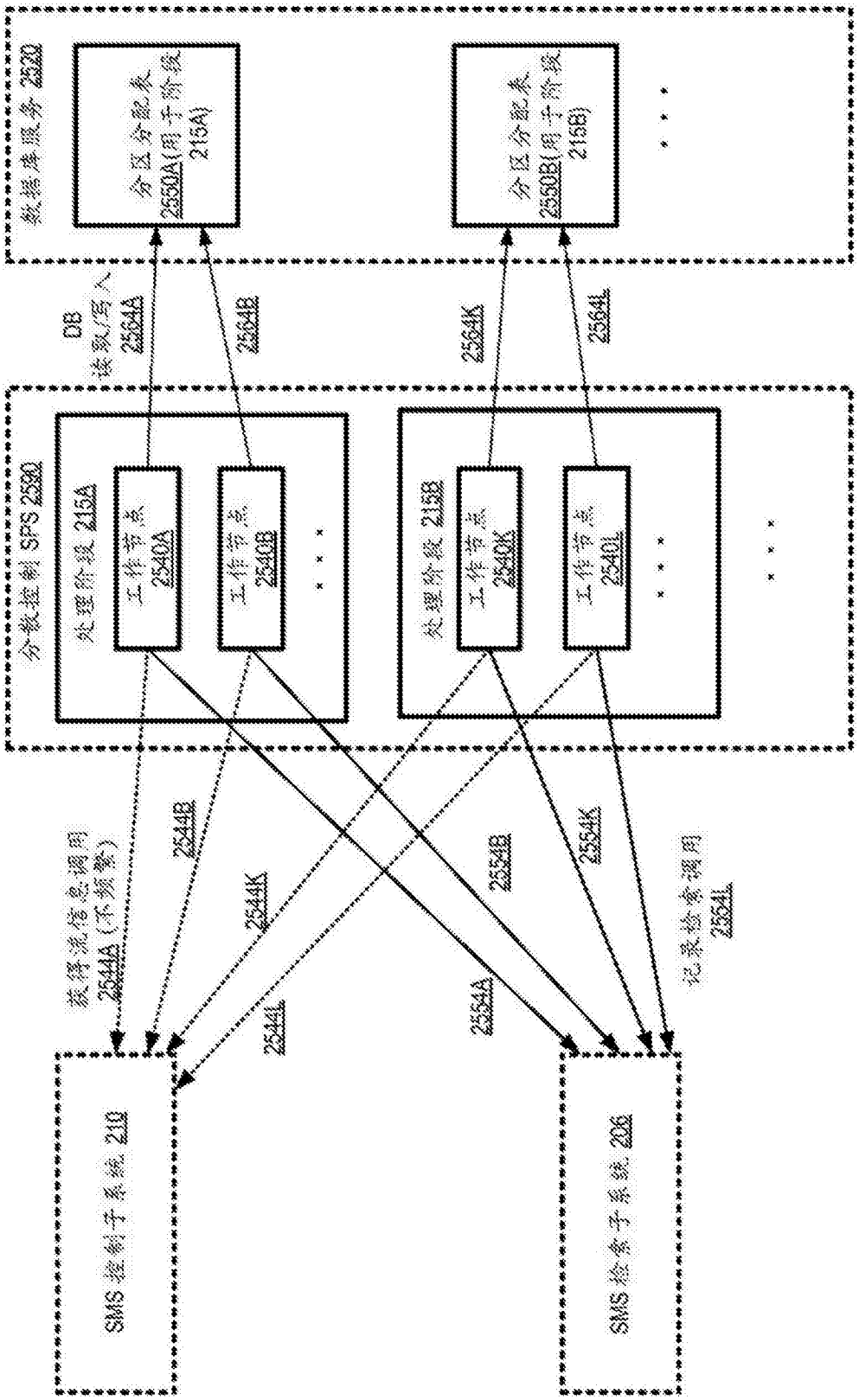

数据库服务 2520
分区分配表 2550A(用于阶段 215A)
分区分配表 2550B(用于阶段 215B)
DB 读取/写入
2564A
2564B
2564K
2564L
分散控制 SPS 2530
处理阶段 215A
工作节点 2540A
工作节点 2540B
处理阶段 215B
工作节点 2540K
工作节点 2540L
获得流信息调用 2544A (不频繁)
2544B
2544K
2544L
2554A
2554B
2554K
记录检索调用 2554L
SMS 控制子系统 210
SMS 检索子系统 206

图25

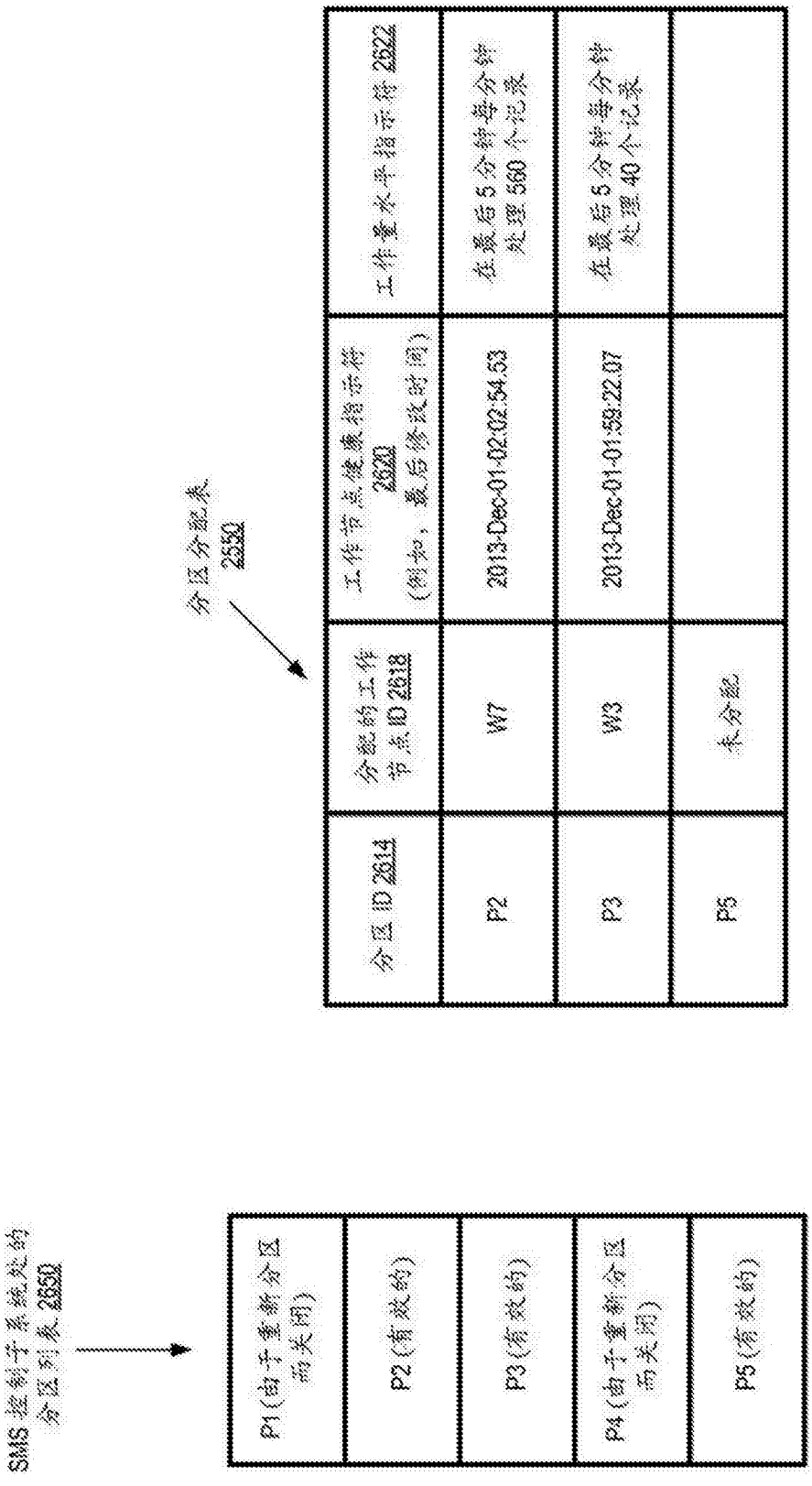

分区分配表
2550
分区 ID 2614
分配的工作节点 ID 2618
工作节点健康指示符 2620 (例如，最后修改时间)
工作量水平指示符 2622
P2
W7
2013-Dec-01-02:02:54.53
在最后 5 分钟每分钟处理 560 个记录
P3
W3
2013-Dec-01-01:58:22.07
在最后 5 分钟每分钟处理 40 个记录
P5
未分配
SMS 控制子系统处的分区列表 2650
P1 (由于重新分区而关闭)
P2 (有效的)
P3 (有效的)
P4 (由于重新分区而关闭)
P5 (有效的)

图26

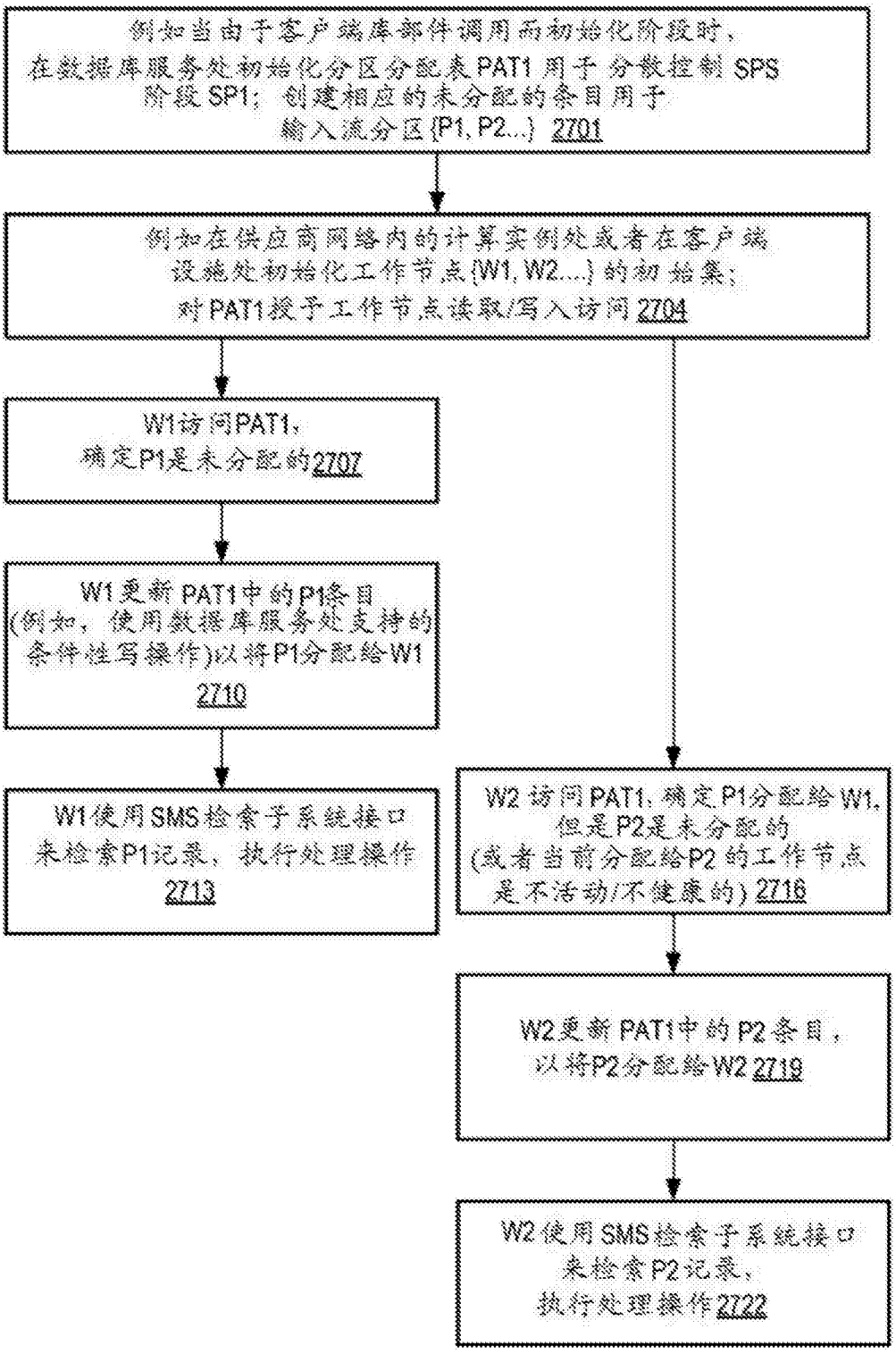

例如当由于客户端库部件调用而初始化阶段时，
在数据库服务处初始化分区分配表PAT1用于分散控制SPS
阶段SP1；创建相应的未分配的条目用于
输入流分区{P1, P2...} 2701
例如在供应商网络内的计算实例处或者在客户端
设施处初始化工作节点{W1, W2...}的初始集；
对PAT1授予工作节点读取/写入访问2704
W1访问PAT1，
确定P1是未分配的2707
W1更新PAT1中的P1条目
(例如，使用数据库服务处支持的
条件性写操作)以将P1分配给W1
2710
W1使用SMS检索子系统接口
来检索P1记录，执行处理操作
2713
W2访问PAT1，确定P1分配给W1，
但是P2是未分配的
(或者当前分配给P2的工作节点
是不活动/不健康的) 2716
W2更新PAT1中的P2条目，
以将P2分配给W2 2719
W2使用SMS检索子系统接口
来检索P2记录，
执行处理操作2722

图27

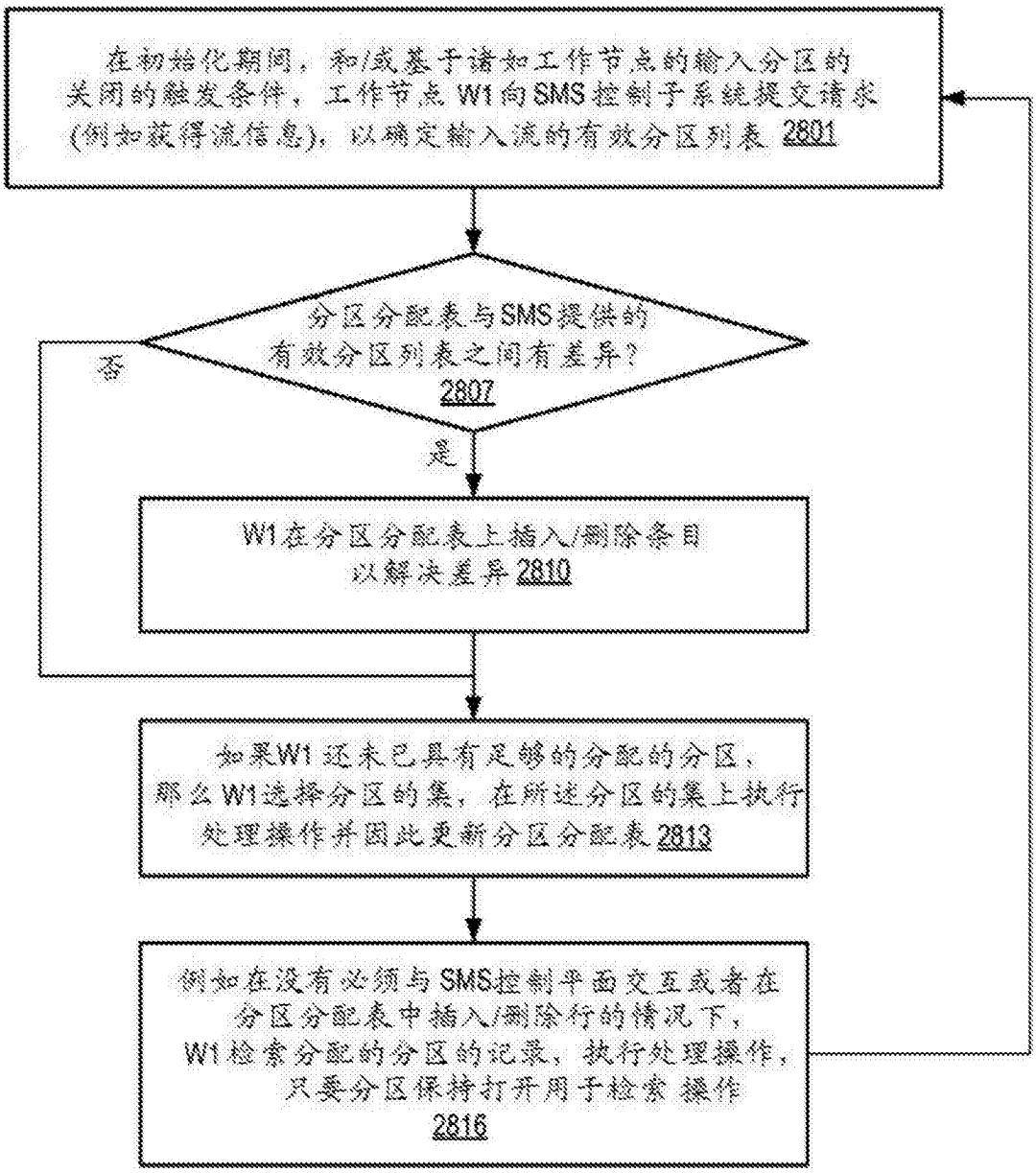

在初始化期间，和/或基于诸如工作节点的输入分区的关闭的触发条件，工作节点W1向SMS控制子系统提交请求(例如获得流信息)，以确定输入流的有效分区列表 2801
分区分配表与SMS提供的有效分区列表之间有差异？ 2807
否
是
W1在分区分配表上插入/删除条目以解决差异 2810
如果W1还未已具有足够的分配的分区，那么W1选择分区的集，在所述分区的集上执行处理操作并因此更新分区分配表 2813
例如在没有必须与SMS控制平面交互或者在分区分配表中插入/删除行的情况下，W1检索分配的分区的记录，执行处理操作，只要分区保持打开用于检索操作 2816

图28

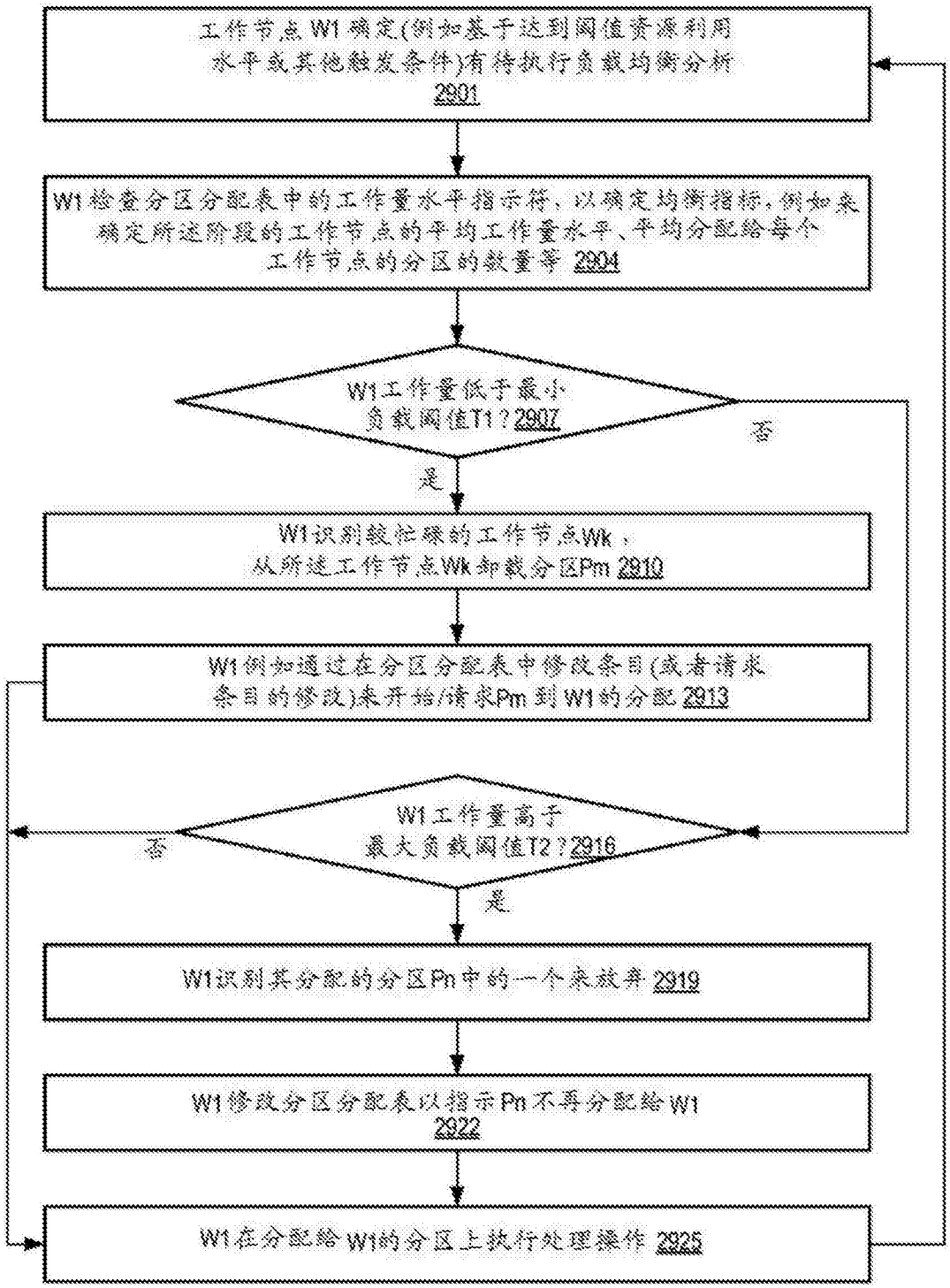

工作节点W1确定(例如基于达到阈值资源利用水平或其他触发条件)有待执行负载均衡分析 2901
W1检查分区分配表中的工作量水平指示符，以确定均衡指标，例如来确定所述阶段的工作节点的平均工作量水平、平均分配给每个工作节点的分区的数量等 2904
W1工作量低于最小负载阈值T1？2907
否
是
W1识别被忙碌的工作节点Wk，从所述工作节点Wk卸载分区Pm 2910
W1例如通过在分区分配表中修改条目(或者请求条目的修改)来开始/请求Pm到W1的分配 2913
W1工作量高于最大负载阈值T2？2916
否
是
W1识别其分配的分区Pn中的一个来放弃 2919
W1修改分区分配表以指示Pn不再分配给W1 2922
W1在分配给W1的分区上执行处理操作 2925

图29

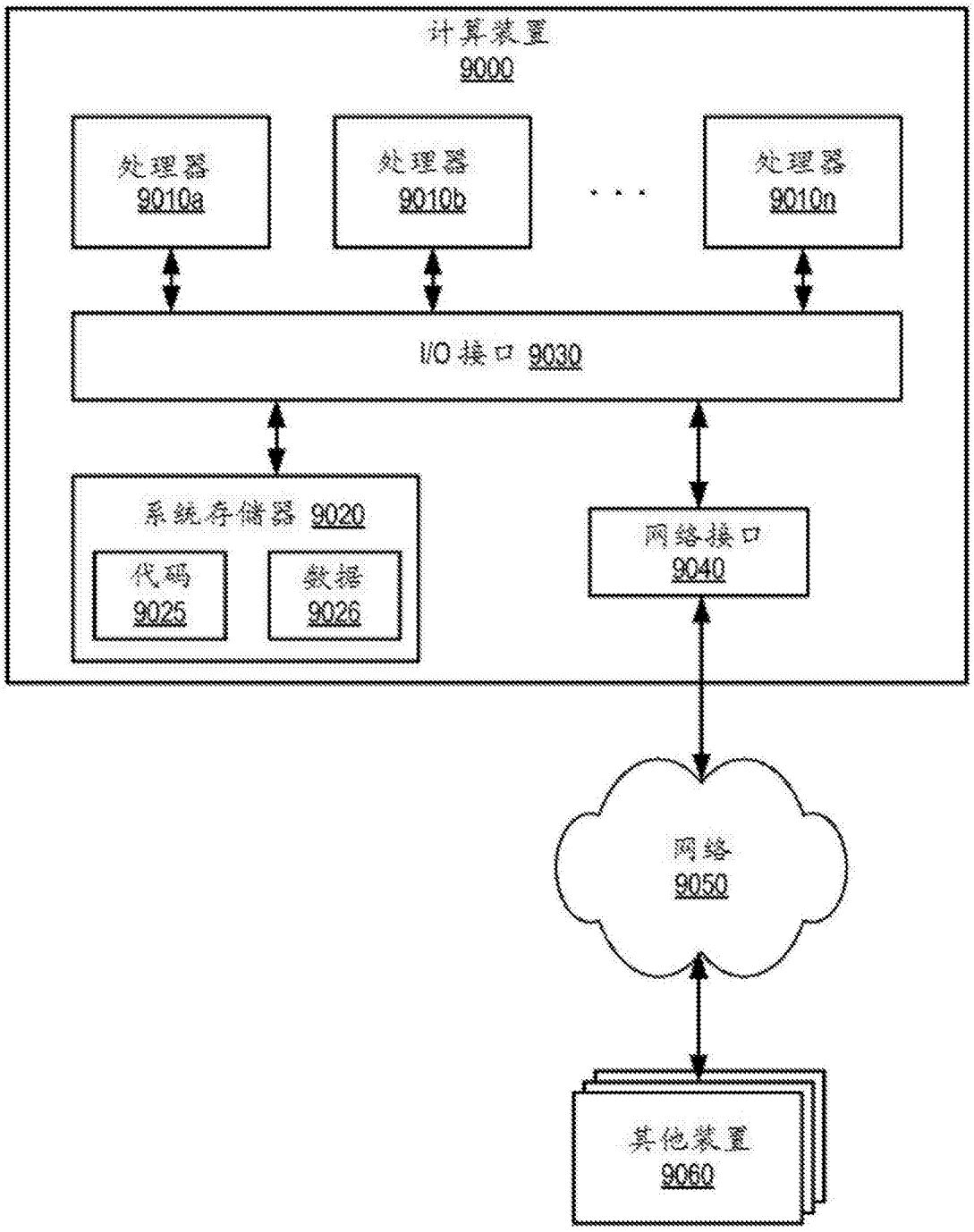

计算装置
9000
处理器
9010a
处理器
9010b
处理器
9010n
I/O 接口 9030
系统存储器 9020
代码
9025
数据
9026
网络接口
9040
网络
9050
其他装置
9060

图30