(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 105706086 B
(45)授权公告日 2019.03.15

(21)申请号 201480061589.4
(22)申请日 2014.11.11
(65)同一申请的已公布的文献号
申请公布号 CN 105706086 A
(43)申请公布日 2016.06.22
(30)优先权数据
14/077,173 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/US2014/065061 2014.11.11
(87)PCT国际申请的公布数据
WO2015/070239 EN 2015.05.14
(73)专利权人 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 16/2455(2019.01)
H04L 12/24(2006.01)
(56)对比文件
US 2012066337 A1,2012.03.15,
US 8161255 B2,2012.04.17,
US 2010257140 A1,2010.10.07,
US 8255739 B1,2012.08.28,
CN 101110757 A,2008.01.23,
US 2012066337 A1,2012.03.15,
CN 101621781 A,2010.01.06,
US 7716186 B2,2010.05.11,
审查员 曾伟

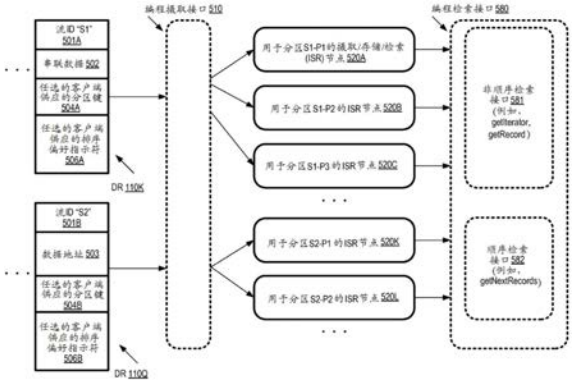
权利要求书3页 说明书46页 附图31页

(54)发明名称

用于获取、存储和消费大规模数据流的管理服务

(57)摘要

一种多租户流管理服务的控制节点接收将数据流初始化的请求，所述数据流包括多个数据记录。基于分区策略，所述控制节点确定将用于配置子系统的参数，以便摄取、存储和检索所述记录。所述控制节点识别将用于检索子系统的节点的资源。所述检索节点被配置成实施编程记录检索接口，包括实施非顺序和顺序访问模式的相应接口。所述控制节点使用选择的资源来配置所述检索节点。



CN 105706086 B

1.一种用于管理数据流的方法,其包括:

由一个或多个计算装置执行:

针对包括多个数据记录序列的特定数据流,确定节点集合,所述节点集合可由一个或多个控制部件配置成基于包括流分区策略的一个或多个策略来执行流管理操作,所述流分区策略用于将包括所述多个数据记录序列的所述特定数据流分区成多个分区;

响应于经由一个或多个编程记录检索接口接收的数据检索请求,提供数据记录,其中所述一个或多个编程记录检索接口包括实现非顺序访问模式的第一检索接口以及实现顺序访问模式的第二检索接口,并且其中与使用所述第一检索接口相关联的计费率与使用所述第二检索接口相关联的计费率不同;以及

至少部分基于所述多个编程记录检索接口的相应使用计数度量,生成与所述特定数据流相关联的客户端计费量。

2.根据权利要求1所述的方法,其还包括由所述一个或多个计算装置执行:

根据所述分区策略,至少部分基于与所述特定数据流的特定数据记录相关联的键,将所述特定数据记录分配到所述特定数据流的第一分区,其中所述键由对应于所述特定数据记录的写入请求指示。

3.根据权利要求1所述的方法,其还包括由所述一个或多个计算装置执行:

至少部分基于特定数据记录被分配到的分区,选择下列项中的一个或多个:(a)负责接受所述特定数据记录的记录摄取子系统的特定节点,(b)负责存储所述特定数据记录的至少一个复本的记录存储子系统的特定节点,以及(c)负责响应于读取请求而从所述记录存储子系统获取所述特定数据记录的检索子系统的特定节点。

4.根据权利要求1所述的方法,其还包括由所述一个或多个计算装置执行:

接收经由一个或多个编程记录提交接口提交的数据记录,其中所述一个或多个编程记录提交接口包括支持数据记录的串联提交的第一提交接口以及使得能够通过引用下列中的一个来提交数据记录的第二提交接口:(a)提供者网络实施的存储服务处的对象地址,(b)通用记录定位符,(c)数据库记录。

5.根据权利要求1所述的方法,其还包括由所述一个或多个计算装置执行:

至少部分基于下列项中的一个,从数据存储子系统的特定节点移除特定数据记录:(a)针对所述特定数据流配置的数据去除重复窗口,(b)与所述特定数据流相关联的数据存档策略,(c)客户端指明的数据保留策略,(d)移除所述特定数据记录的客户端请求,或者(e)所述特定数据记录已被一个或多个数据消费者处理的指示。

6.根据权利要求1所述的方法,其中针对所述特定数据流配置的(a)记录摄取子系统、(b)记录存储子系统或(c)记录检索子系统中的至少一个子系统包括被所述一个或多个控制部件配置成冗余组的成员的多个节点,其中所述冗余组包括:(a)被分配来对所述流的数据记录的集合执行操作的一个或多个主节点,以及(b)被配置成响应于一个或多个触发事件而承担主节点角色的一个或多个非主节点。

7.根据权利要求6所述的方法,其中在特定数据中心处将所述一个或多个主节点中的特定主节点实例化,并且其中在不同的数据中心处将所述一个或多个非主节点中的特定非主节点实例化。

8.根据权利要求6所述的方法,其还包括由所述一个或多个计算装置执行:

检测所述一个或多个触发事件中的触发事件;以及

通知所述一个或多个非主节点中的特定非主节点来承担所述主节点角色。

9.根据权利要求1所述的方法,其中所述一个或多个控制部件包括被配置成冗余组的多个控制节点,包括:主控制节点,其被配置成响应对包括所述特定数据流的一个或多个数据流进行控制平面操作的请求;以及至少一个非主控制节点,其被配置成响应于一个或多个触发事件而承担主节点角色。

10.根据权利要求1所述的方法,其还包括由所述一个或多个计算装置执行:

在提供者网络的网络可访问数据库处,存储所述特定数据流的元数据,所述元数据包括根据所述流分区策略生成的分区映射;以及

从记录检索子系统访问所述元数据,以响应于特定记录检索请求。

11.根据权利要求1所述的方法,其还包括由所述一个或多个计算装置执行:

生成对应于所述数据流的特定数据记录的特定序列号,所述特定序列号表示相对于所述特定数据记录所属的所述分区的其他数据记录而言,在记录摄取子系统处接收到所述特定数据记录的顺序;

至少部分基于针对所述数据记录生成的相应时间戳,由记录存储子系统按一定顺序存储多个数据记录,所述多个数据记录包括所述特定数据记录;以及

响应于接收到以所述特定序列号作为参数调用所述第一检索接口的读取请求,从所述记录存储子系统中检索所述特定数据记录。

12.一种用于管理数据流的系统,所述系统包括一个或多个处理器以及一个或多个存储器,所述一个或多个存储器包括程序指令,所述程序指令在一个或多个处理器上执行时实施多租户流管理服务的控制节点,其中所述控制节点被配置成:

接收将特定数据流初始化的请求,所述特定数据流包括由一个或多个数据生产者生成的多个数据记录序列;

至少部分基于分区策略,确定将用于针对所述特定数据流来配置一个或多个子系统的一个或多个参数,所述分区策略用于将包括所述多个数据记录序列的所述特定数据流分区成多个分区,所述一个或多个子系统包括记录检索子系统,其中所述一个或多个参数包括将在所述记录检索子系统中实例化的节点的初始数量;

识别将用于所述记录检索子系统的特定节点的一个或多个资源,其中所述特定节点被配置成实施多个编程记录检索接口,包括实现非顺序访问模式的第一检索接口以及实现顺序访问模式的第二检索接口;以及

使用所述一个或多个资源来配置所述记录检索子系统的所述特定节点。

13.根据权利要求12所述的系统,其中所述一个或多个子系统包括:记录摄取子系统,其包括可配置成接收所述特定数据流的数据记录的一个或多个节点;以及记录存储子系统,其可配置成将所述流的数据记录存储在选择的存储位置集合处。

14.根据权利要求13所述的系统,其中所述记录摄取子系统、所述记录存储子系统和所述记录检索子系统中的至少一个子系统包括被配置成冗余组的成员的多个节点,其中所述冗余组包括:(a)被分配来对所述流的数据记录的集合执行操作的一个或多个主节点,以及(b)被配置成响应于一个或多个触发事件而承担主节点角色的一个或多个非主节点。

15.根据权利要求13所述的系统,所述记录摄取子系统、所述记录存储子系统或者所述

记录检索子系统中的一个的至少一个节点包括在提供者网络处实施的虚拟化计算服务的计算实例的部件。

用于获取、存储和消费大规模数据流的管理服务

背景技术

[0001] 随着近年来数据存储的成本下降，并且随着将计算基础设施的各种元件互连的能力提高，有可能收集并分析涉及广泛多种应用的越来越多的数据。例如，移动电话可以生成指示其位置、电话用户使用的应用等等的数据，其中的至少一些数据可以收集和分析，以便将定制化赠券、广告等呈现给用户。对监控摄像机收集的数据进行分析可用于防止和/或处理犯罪案件，并且从嵌入在飞机发动机、汽车或复杂机械内的各个位置的传感器收集的数据可用于各种目的，例如，预防性维护、提高效率以及降低成本。

[0002] 流数据量的增加伴随(并且在一些情况下，可能伴随)商用硬件的使用增加。在为许多类型的应用管理大规模计算资源方面，用于商用硬件的虚拟化技术的出现提供益处，从而允许各种计算资源被多个客户有效且安全地共享。例如，通过为每个用户提供由单个物理计算机托管的一个或多个虚拟机，虚拟化技术可允许在多个用户之间共享单个物理计算机，其中每个这样的虚拟机都是充当不同逻辑计算系统的软件模拟，从而使用户产生他们是给定硬件计算资源的唯一操作员和管理员的错觉，同时还在各个虚拟机之间提供应用隔离和安全保证。此外，一些虚拟化技术能够提供跨两个或更多物理资源的虚拟资源，例如，单个虚拟机具有跨多个不同物理计算系统的多个虚拟处理器。除了计算平台之外，一些大组织还提供使用虚拟化技术建立的各种类型的存储服务。使用这种存储服务，可以在所需的耐用性水平存储大量的数据。

[0003] 尽管可用相对低的成本从各种提供者得到虚拟化计算和/或存储资源，然而，由于许多原因，对大型动态波动数据流的收集、存储和处理进行管理和协调仍然是具有挑战性的课题。随着更多的资源被添加到为处理大型数据流而设立的系统，例如，系统的不同部分之间可能会出现工作负载不平衡。如果不加以解决，除了没有充分利用(且因此浪费)其他资源之外，这种不平衡会导致某些资源出现严重的性能问题。如果此类数据或结果存储在客户端无法控制的设施处，客户端还可能关心流数据的安全性或者分析流数据的结果。随着分布式系统的大小增长，出现故障的频率自然也会增加(例如，偶尔失去连接和/或出现硬件故障)，这些故障也必须被有效解决，以防止流数据收集、存储或分析出现中断而增加成本。

附图说明

[0004] 图1提供根据至少一些实施方案的数据流概念的简单概述。

[0005] 图2提供根据至少一些实施方案的流管理系统(SMS)和流处理系统(SPS)的各个子部件之间的数据流动的概述，所述SPS包括流处理阶段的集合。

[0006] 图3示出根据至少一些实施方案的可以在SMS和SPS处实施的编程接口的相应集合的实例。

[0007] 图4示出根据至少一些实施方案的示例性基于web的接口，所述接口可以被实施以使得SPS客户端能够生成流处理阶段的图形。

[0008] 图5示出根据至少一些实施方案的可以在SMS处实施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，以及检索子系统与SPS交互的实例。

[0012] 图9示出根据至少一些实施方案的冗余组的实例，所述冗余组可以针对SMS或SPS的节点而设立。

[0013] 图10示出根据至少一些实施方案的提供者网络环境，其中给定冗余组的节点可以分布在多个数据中心之间。

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

[0015] 图12a和图12b示出根据至少一些实施方案的分别可由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者与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针对SMS处的摄取数据记录生成的序列号的示例性元素。

[0018] 图14示出根据至少一些实施方案的在SMS处的流数据记录的顺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可以针对SMS和SPS节点作出的流分区映射和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是示出根据至少一些实施方案的可以被执行以支持流记录摄取和流记录检索的编程接口的相应集合的操作方面的流程图。

[0022] 图18a是示出根据至少一些实施方案的可以被执行以配置流处理阶段的操作方面的流程图。

[0023] 图18b是示出根据至少一些实施方案的可以响应于调用客户端库的部件以配置流处理工作者节点而执行的操作方面的流程图。

[0024] 图19是示出根据至少一些实施方案的可以被执行以实施流处理的一个或多个恢复策略的操作方面的流程图。

[0025] 图20是示出根据至少一些实施方案的可以被执行以实施数据流的多个安全选项的操作方面的流程图。

[0026] 图21是示出根据至少一些实施方案的可以被执行以实施数据流的分区策略的操作方面的流程图。

[0027] 图22是示出根据至少一些实施方案的可以被执行以实施数据流的动态重新分区的操作方面的流程图。

[0028] 图23是示出根据至少一些实施方案的可以被执行以实施数据流记录的至少一次(at-least-once)记录摄取策略的操作方面的流程图。

[0029] 图24是示出根据至少一些实施方案的可以被执行以实施数据流的多个持久性策

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者节点使用数据库表来协调它们的工作负载。

[0031] 图26示出根据至少一些实施方案的可以存储在分区分配表中的示例性条目,所述分区分配表用于工作负载协调。

[0032] 图27示出根据至少一些实施方案的可以由流处理阶段的工作者节点执行以便选择分区的操作方面,其中在所述分区上执行处理操作。

[0033] 图28示出根据至少一些实施方案的可以由流处理阶段的工作者节点执行,以便基于从流管理服务控制子系统获取的信息来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以由流处理阶段的工作者节点执行的负载平衡操作的方面。

[0035] 图30是示出可以用于至少一些实施方案的示例性计算装置的框图。

[0036] 虽然本文通过若干实施方案和说明性附图以实例的方式描述实施方案,但所属领域的技术人员将认识到,实施方案并不受限于所述的实施方案或附图。应理解,附图和详细说明并不意图将实施方案限定于所公开的特定形式,相反,本发明将涵盖落在由所附权利要求书界定的精神和范围内的所有修改、等效物和替代。本文中使用的标题只是为了编制的目的,而并不用来限制说明书或权利要求书的范围。本申请全文所用字词“可以”以许可意义(即,表示具有潜在可能)使用,而非强制意义(即,表示必须)。类似地,字词“包括(include)”、“包括(including)”以及“包括(includes)”意指包括但并不限于。

具体实施方式

[0037] 本发明描述用于管理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,所述方法和设备被设计用来处理数百或甚至数千并行的数据生产者和数据消费者。本文中使用的术语“数据流”是指可由一个或多个数据生产者生成并被一个或多个数据消费者访问的数据记录序列,其中假设每个数据记录都是不可变的字节序列。流管理服务(SMS)可以提供编程接口(例如,应用编程接口(API)、web页或web站、图形用户接口或者命令行工具),以使得能够创建、配置和删除流,以及在一些实施方案中,提交、存储和检索流数据记录。涉及与SMS控制部件交互的一些类型的流操作(例如,流创建或删除,或者下文描述的那种动态分区操作)在本文中可称为“控制平面”操作,而一般(例如,在正常操作条件下)不需要与控制部件交互的操作(例如,数据记录提交、存储和检索)在本文中可称为“数据平面”操作。在一些此类实施方案中,计算、存储和联网资源的动态供应集合可以用来实施服务,例如,基于允许流管理工作负载以可扩展的方式分布在很多服务部件之间的各种分区策略,如下文进一步详细地描述。本文中使用的缩略词SMS可以指流管理服务,并且也指流管理系统,所述流管理系统包括用来实施流管理服务的虚拟和/或物理资源的集合。

[0038] 在各实施方案中,SMS的一些客户可以开发直接调用SMS编程接口的应用。然而,在至少一些实施方案中,除了SMS接口之外,还可为客户提供较高层次的抽象或应用层处理框架,从而可为不希望使用SMS直接支持的较低层次流管理功能来开发应用的那些客户端简化流处理的各个方面。这样的框架可以提供自己的编程接口(例如,建立在SMS接口之上),从而与较低层次的流管理操作相比,使客户能够更集中于将要使用流记录实施的业务逻

辑。在一些实施方案中，较高层次框架可以作为本身具有控制平面和数据平面部件的流处理服务(SPS)来实施，从而可提供高级功能，例如，用于流处理的自动化资源供应、处理节点的自动化故障转移、构建任意流处理工作流程图的能力、支持短暂流、基于工作负载改变或其他触发条件的动态重新分区等等。在至少一些实施方案中，流管理服务、流处理服务或这两个服务都可以作为虚拟化环境中的多租户管理式网络可访问的服务来实施。也就是说，在此类实施方案中，至少在一些情况下，各种物理资源(例如，计算机服务器或主机、存储装置、网络装置等)可以在不同客户的流之间共享，而根本不必让客户确切地意识到资源是如何共享的，或者甚至让客户意识到正在共享给定的资源。基于各种适用的策略(其中一些策略可是客户端可选择的)，管理的多租户流管理和/或处理管理服务的控制部件可以动态添加、移除或者重新配置用于特定流的节点或资源。此外，控制部件还可以负责：透明地实施各种类型的安全协议(例如，以确保一个客户端的流应用无法访问另一客户端的数据，即使这两个客户端可以共享至少一些硬件或软件)、监控资源使用以计费、生成可用于审计或除错的日志信息，等等。从管理的多租户服务的客户端的角度，所述服务实施的控制/管理功能可以消除支持大规模流应用中涉及的大部分复杂性。在一些情形下，此类多租户服务的客户能够表明他们不希望针对至少一些类型的流相关应用共享资源，在所述情况下，对于这些类型的操作而言，可以将一些物理资源至少临时指定为单租户的(即，限于代表单个客户或客户端执行的操作)。

[0039] 在各实施方案中，可以采用许多不同的方法来实施SMS和/或SPS控制平面和数据平面操作。例如，针对控制平面操作，在一些实施中，可以设置控制服务器或节点的冗余组。冗余组可以包括多个控制服务器，其中一个服务器被指定为负责响应于有关各种流的管理请求的主服务器，而在诸如当前的主服务器出现故障(或失去连接)的触发条件下，另一服务器可以被指定接管作为主服务器。在另一实施中，在网络可访问的数据库服务处创建的一个或多个表可以用来存储各种流的控制平面元数据(例如，分区映射)，并且各个摄取、存储或检索节点可以根据需要访问所述表以获取数据平面操作所需的元数据的子集。下文提供有关不同实施方案中的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，在实施流管理服务的一些实施方案中，可能不必实施提供较高层次原语的流处理服务。在其他实施方案中，只有流处理服务的高层次编程接口可以暴露于客户，并且客户端可能无法使用较低层次的流管理接口。

[0040] 根据一些实施方案，流管理系统可以包括多个可独立配置的子系统，所述子系统包括：记录摄取子系统，其主要负责获取或收集数据记录；记录存储子系统，其主要负责根据适用的持久性或耐久性策略来保存数据记录内容；以及记录检索子系统，其主要负责响应于指向存储记录的读取请求。在一些实施方案中，还可以实施控制子系统，其包括负责配置其余子系统的一个或多个管理或控制部件，例如，通过针对诸如虚拟或物理服务器等所选资源处的摄取、存储和检索子系统中的每个，动态确定和/或初始化节点的所需数量。摄取、存储、检索和控制子系统中的每个均可以使用相应多个硬件和/或软件部件进行实施，所述部件可以统称为子系统的“节点”或“服务器”。因此，SMS的各种资源在逻辑上可以说成属于四个功能类别中的一个：摄取、存储、检索或者控制。在一些实施中，可以针对其他子系统中的每个来建立控制部件的相应集合，例如，可以实施独立的摄取控制子系统、存储控制子系统和/或检索控制子系统。每个这样的控制子系统可以负责识别将用于对应子系统的

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。在一些实施中，可以提前设置能够执行各种类型的SMS和/或SPS功能的节点池，并且根据需要，可以将从这些池中选择的成员分配给新的流或新的处理阶段。

[0041] 在至少一些实施方案中，可以实施流分区策略和相关联的映射，例如，以便在摄取、存储、检索和/或控制节点的不同集合之间分发数据记录的子集。例如，基于为特定数据流选择的分区策略并且基于其他因素，例如，预期的记录摄取率和/或检索率，控制部件可以确定最初(即，在流创建时)应该为摄取、存储和检索建立多少节点(例如，进程或线程)，以及这些节点应该如何映射到虚拟机和/或物理机。随着时间的推移，与给定流相关联的工作负载可以增加或减少，从而(在其他触发条件下)可导致流的重新分区。这种重新分区可以涉及改变各种参数，例如，将用来确定记录的分区的功能；所用的分区键；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，可以使用下文进一步详细描述的技术来动态实施重新分区，而不必中断数据记录的流动。在一些实施方案中，针对不同的数据流可以使用不同的分区方案和重新分区触发准则，例如，基于客户端提供的参数或者SMS控制节点的探试法。在一些实施方案中，例如，基于客户端偏好、流的预期使用期限或者其他因素，有可能限制重新分区的数量和/或频率。

[0042] 在不同的实施方案中，可以实施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，被配置成代表SMS的客户来调用SMS的编程接口的可执行部件或模块)可以使用串联提交接口或按引用提交接口。对于串联提交而言，在此类实施方案中，数据记录的内容或主体可以被包括作为提交请求的一部分。相反，在按引用提交请求中，可以提供从中可获取数据记录的内容或主体的地址(例如，存储装置地址、数据库记录地址或者URL(通用记录定位符))。在一些实施中，还可以支持或者改为支持混合提交接口，其中数据记录的前N个字节可以串联的方式包括，而剩余的字节(如果有的话)以引用的方式提供。在这种情形下，短记录(主体长度少于N个字节)可以由提交请求充分说明，而较长记录的部分可能需要从对应的地址获取。

[0043] 除了在摄取期间用于指定记录内容的不同替代物之外，在一些实施方案中，还可实施多种确认或去除重复相关的摄取策略。例如，对于一些流应用而言，客户端可能希望确保每一个数据记录都被SMS可靠地摄取。在大型分布式流管理环境中，包可能会丢失，或者沿着数据生产者与摄取节点之间的路径，可能不时地发生各种故障，从而可能导致丢失一些提交的数据。因此，在一些实施方案中，SMS可以实施至少一次摄取策略，根据所述摄取策略，记录提交者可以一次或多次提交相同的记录直至接收到来自摄取子系统的肯定确认为止。在正常操作条件下，记录可以提交一次，并且在接收摄取节点已经获取并存储记录之后，提交者可以接收到确认。如果确认丢失或延迟，或者如果记录提交请求本身丢失，那么提交者可以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为止。例如，基于如果提交者已接收到确认则记录将不会被重新提交的预期，无论是否重复，摄取节点可以针对每个提交都生成确认。然而，在至少一些实施方案中，摄取节点可以负责识别相同的数据记录已被多次提交，以及负责避免不必要地存储重复数据的新复本。在一个实施方案中，可以支持至少一次摄取策略的至少两个版本，在一个版本(可被称为“至少一次摄取、无重复”)中，SMS负责去除重复数据记录(即，确保，响应于两次或更多提交的集合中的仅一个，

数据存储在SMS存储子系统)，并且在一个版本中，准许SMS重复数据记录存储(可被称为“至少一次、准许重复”)。所述至少一次、准许重复的方法可以用于数据记录重复很少或没有造成消极后果的流应用，和/或用于本身执行重复消除的流应用。也可以支持其他摄取策略，例如，最大努力(best-effort)摄取策略，其中提交的每个数据记录无需确认。在至少一些实施方案中，如果实行最大努力摄取策略，那么可以接受丢失少许数据记录。在各实施方案中，客户端可以选择他们希望使用哪些摄取策略用于各种流。

[0044] 至于流记录的存储，在至少一些实施方案中，还可支持许多替代策略。例如，客户端能够从SMS支持的若干策略中选择一个持久性策略，所述策略管理记录存储的一些方面，例如，将要存储的给定数据记录的复本数量、将用于复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设备、固态装置(SSD)、网络附接的存储装置等)等等。例如，如果客户端为基于磁盘的存储选择N个复本持久性策略，那么直到将记录的N个复本安全写入N个相应的磁盘装置，数据记录提交才被视作完成。在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将给定分区的输入数据记录相继写入磁盘，例如，以避免磁盘寻道的性能影响。使用下文所述的各种技术，例如，包括使得能够基于摄取时间进行按序记录检索的基于时间戳的技术，可以为数据记录生成(并随其一起存储)序列号。在至少一些实施方案中，给定分区的数据记录可以存储在一起，例如，在磁盘上连续存储，并且与其他分区的数据记录分开。在一些实施中，根据(客户端或SMS选择的)保留策略或者去除重复时间窗策略(表明时间段、在提交任何给定数据记录之后、SMS可被要求确保SMS存储子系统没有存储该给定数据记录的复本的时间段，即使提交了一些复本)，至少一些数据记录可以存档到不同类型的存储服务和/或在离开SMS的一时间段之后被删除。这种移除操作在本文中可称为流“修整”。在一些实施方案中，客户端可以提交流修整请求，例如，通知SMS指定的数据记录不再被需要，且因此从提交修整请求或明确请求删除指定数据记录的客户端的角度，可以删除所述数据记录。在可能有多个客户端消费给定流的数据记录的情形下，SMS可以负责确保在所有感兴趣消费者访问给定记录之前，所述给定记录不会过早删除或修整。在一些实施中，如果给定流有N个数据消费者，那么在删除所述流的给定记录R之前，SMS可以等到确定所有N个数据消费者都已读取或处理R为止。例如，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中的进度的相应指示，SMS可确定R已被所有消费者读取。在一些实施方案中，一些类型的数据消费者(例如，测试相关应用)可以接受在数据记录被访问之前删除数据记录的至少小子集。因此，在至少一些实施方案中，应用能够通知SMS有关可接受在检索之前删除数据，并且SMS可以根据通知来安排删除。在一些实施方案中，可以实施存档策略，例如，作为数据保留策略的一部分，从而表明(例如)应复制流数据记录至的存储装置的类型，以及将用于此类复本的调度策略。

[0045] 在至少一些实施方案中，多个编程接口还可被支持用于记录检索。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，getIterator)可以用来将迭代器或游标实例化并且定位在流的分区内的指定逻辑偏移量处(例如，基于序列号或时间戳)。随后，不同的编程接口(例如，getNextRecords)可以用来从迭代器的当前位置开始按顺序读取指定数量的数据记录。迭代器的实例化实际上可以允许客户端在流分区内指定记录检索的任意或随机开始位置。在此类实施方案中，如果客户端希望以随机访问的模式读取数据记录，那么客户端可能需要重复创建新的迭代器。在基于旋转磁盘的存储系统中，频

繁随机访问所需的磁盘寻道可显著影响I/O响应时间。因此，在至少一些实施方案中，作为激励客户端顺序读取而不是随机读取流数据记录，可将与应用于顺序读取访问不同(例如，更高)的计费率应用于随机读取访问。因此，例如，在一些实施中，可以对客户端为每个getIterator调用支付X个货币单元，而每次经由getNextRecords检索的记录要支付Y个货币单元，其中X>Y。当针对其他操作种类(例如，摄取)支持替代客户端接口时，在至少一些实施方案中，这些替代接口的计费率或价格也可不同，例如，与在线提交请求相比，客户端可能要为按引用提交请求支付更多费用，就好像与顺序读取相比，客户端可能要为随机读取支付更多费用。在各实施方案中，其他因素也会影响计费，例如，数据记录的大小、写入与读取请求随时间的分布、所选的持久性策略等等。

[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处理工作流程，其中在给定阶段执行的处理的输出可以用作零阶段或更多其他阶段的输入。在一些实施方案中，分区策略(类似于针对SMS描述的用于摄取、存储和检索数据记录的那些)可用来将处理工作负载分在各个阶段处的多个工作者节点之间。在一个这样的实施方案中，可以实施编程SPS接口，从而使得客户端能够指定用于任何给定阶段的各种配置设置，例如，包括用于所述阶段的输入数据源(例如，从中检索数据记录的一个或多个流，以及用于所述流的分区策略)、将在所述阶段执行的处理操作以及用于来自所述阶段的输出或结果分布的描述符或说明(例如，输出将被保存到存储位置、发送到网络端点，还是以不同的流的形式馈送到一个或多个其他处理阶段)。在至少一些实施方案中，为SPS阶段指定的处理操作可以是幂等的：也就是说，如果多次对同一输入数据执行给定的处理操作，那么操作的结果不会与只执行一次操作就获取的结果不同。如果处理操作是幂等的，则可以简化故障(例如，SPS阶段的工作者节点故障)的恢复，如下文进一步详细地描述。根据一些实施方案，一些或所有SPS阶段可以准许非幂等的处理操作。

[0047] 至少部分基于配置信息例如输入流分区策略以及接着经由SPS编程接口接收的处理操作的性质，在各实施方案中，SPS控制服务器可以确定针对处理工作流程的各个阶段，最初应设置多少个工作者节点。当确定工作者节点的初始数量和放置时，也可考虑将用于工作者节点的资源的性能能力(例如，所用的虚拟机或物理机)。所选数量的工作者节点(在一些实施中，每一个都可以包括可执行线程或可执行进程)可以实例化。例如，每个工作者节点可被配置成获取来自适当输入源(例如，来自一个或多个流分区的检索节点)的数据记录、对数据记录执行指定的处理操作以及将处理结果传输到指定的输出目的地。此外，在至少一些实施方案中，可以实施检查点方案，根据所述方案，给定的工作者节点可被配置成存储进程记录或检查点，其表示已经在所述工作者节点处处理的分区的那部分，其中假设按顺序处理分区记录。例如，在一些实施中，工作者节点可以定期(例如，每N秒一次或者一旦处理了每R个数据记录后)和/或响应于来自SPS控制服务器的检查点请求，将进程记录写入到永久存储装置中。

[0048] 在一些实施方案中，进程记录可用于从工作者节点故障中快速恢复。例如，SPS控制服务器可以监控各个工作者节点随着时间变化的健康状况，例如，使用心跳机制和/或通过监控资源利用水平(例如，CPU利用水平、I/O装置利用水平或者网络利用水平)。响应于SPS控制服务器确定特定工作者节点处于不希望或不健康的状态(例如，如果它无响应或超负载的话)，置换工作者节点可被实例化，以接管特定工作者节点的职责。置换工作者节点

可以访问被置换的工作者节点存储的最近进程记录，以识别置换工作者节点应处理的数据记录的集合。在处理操作是幂等的实施方案中，即使重复一些操作(例如，因为在置换工作者的实例化之前的一段时间写入最近的进程记录)，处理的整体结果也不会受故障和置换影响。在一些实施中，除了存储表示已经被它处理的给定流或分区的子集的进程记录之外，工作者节点也可被配置成存储累积的应用状态信息。例如，如果流处理工作流程负责基于分析表示服务使用度量的流数据记录来确定特定服务的客户端计费量，那么工作者节点可定期存储为各个客户端确定的累积计费量。

[0049] 在至少一些实施方案中，SPS控制服务器还可被配置成，通过开始其他动作，例如，针对各个阶段来请求输入流的动态重新分区、改变在给定阶段分配到给定分区的工作者节点的数量、将更高性能的工作者节点分配给一些阶段或者将工作者节点从一个物理资源转移到具有不同性能能力的另一物理资源，来响应于各种其他触发条件，例如，改变工作负载水平或检测到工作负载不平衡(例如，如果一个分区的摄取率不成比例地高于其他分区的话)。在一些实施方案中，例如，响应于SPS控制服务器确定将要针对给定阶段实施最大努力恢复策略而非基于检查点的恢复策略，至少一些SPS阶段的工作者节点可以不存储上述类型的进程记录。在此类最大努力恢复策略的一些实施中，置换工作者节点可以在接收到新的数据时就简单进行处理，而无需访问进程记录。在一些实施方案中，如果客户端希望在SPS阶段实施最大努力恢复策略，那么在所述阶段执行的流处理操作不必是幂等的。在将在SPS阶段对流记录执行非幂等处理操作的一些实施方案中，可能不支持基于检查点的恢复，而是可使用不同的恢复方案，例如，最大努力恢复。在至少一个实施方案中，SPS阶段可以只允许幂等流处理操作。

[0050] 一些流的数据记录可能含有敏感或机密信息，或者在SPS阶段执行的处理操作可以包括使用专有算法，而竞争者对所述专有算法的恢复可能遇到问题。因此，客户端可以关心各种流管理和处理操作的安全性，尤其是在使用位于客户端本身没有完全控制的提供者网络数据中心处的资源执行操作的情况下。由诸如公司或公共部门组织等实体设置的网络将可经由互联网和/或其他网络访问的一个或多个网络可访问服务(例如，各种类型的基于云的数据库、计算或存储服务)提供到分布式客户端集合，所述网络在本文中可称为提供者网络。在一些实施方案中，客户端能够针对它们的数据流从多个安全相关选项中进行选择。如上文所述，组合的SPS和SMS配置可包括属于许多不同功能类别的节点，例如，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处理或工作者节点。在一些实施方案中，客户端可用的安全相关选择可以包括用于各种类型的节点的放置或位置的选项。例如，在一个实施方案中，客户端能够请求用于流工作流程的一个或多个处理阶段的SPS工作者节点在位于客户端拥有的设施上的计算装置处实施，即使最初使用位于提供者网络处的资源收集和/或存储流记录。响应于这种放置请求，用于给定流的不同功能类别的节点可以在具有不同安全特性或配置文件的相应资源集处实例化。

[0051] 在不同实施方案中，资源集的各安全相关特性可以彼此不同，例如，包括物理位置、所用的物理安全协议(例如，所述协议具有资源的物理访问权)、网络隔离等级(例如，各个实体可见的资源的网络地址的程度)、多租户对单租户等等。在一些实施方案中，客户端能够在提供者网络内建立隔离虚拟网络(IVN)，其中给定客户端可以实质上控制该客户端的IVN内所包括的各种装置的网络配置。具体而言，客户端能够限制对分配给其IVN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。在此类实施方案中,客户端可以请求其SMS或SPS节点的某些子集在指定的IVN内被实例化。在将诸如虚拟化实例主机(它通常可被配置成多租户主机)等提供者网络资源用于各种类别的SMS或SPS节点的实施方案中,客户端可请求一些节点集合在实例主机上进行实例化,所述实例主机被限制于实施只属于所述客户端的实例(即,一些SMS或SPS节点可以在被配置成单租户主机的实例主机处实施)。

[0052] 在一些实施方案中,作为另一安全相关选项,客户端可以请求特定流的数据记录,所述数据记录在通过网络链路传输之前进行加密---例如,在SMS处被摄取之前、在摄取与存储子系统之间、在存储与检索子系统之间、在检索子系统与SPS工作者节点之间和/或在工作者节点与SPS输出目的地之间。在一些实施方案中,客户端可以指定将要使用的加密算法。在一个实施方案中,安全联网协议,例如TLS(传输层安全)或SSL(安全套接字层)协议,可以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提供根据至少一些实施方案的数据流概念的简单概述。如图所示,流100可包括多个数据记录(DR)110,例如,DR 110A、110B、110C、110D和110E。一个或多个数据生产者120(也可被称为数据源),例如数据生产者120A和120B,可执行写入操作151,以生成流100的数据记录的内容。在不同的实施方案中,例如移动电话或平板应用、传感器阵列、社交媒体平台、日志应用或系统日志部件、各种监控代理等等,许多不同类型的数据生产者可以生成数据流。一个或多个数据消费者130(例如,数据消费者130A和130B)可执行读取操作152,以访问由数据生产者120生成的数据记录的内容。在一些实施方案中,例如,数据消费者130可包括流处理阶段的工作者节点。

[0055] 在至少一些实施方案中,存储在SMS中的给定数据记录110可包括数据部分101(例如,DR 110A、110B、110C、110D和110E的相应数据部分101A、101B、101C、101D和101E)以及序列号SN 102(例如,DR 110A、110B、110C、110D和110E的相应SN 102A、102B、102C、102D和102E)。在所述实施方案中,序列号102可以表示在流管理系统处(或者在流管理系统的特定节点处)接收到DR的顺序。在一些实施中,数据部分101可以包括不可变的非解译字节序列:也就是说,一旦写入操作152完成,作为写入的结果生成的DR内容不可以被SMS改变,并且一般来说,SMS可能没有意识到数据的语义。在一些实施中,给定流100的不同数据记录可以包括不同的数据量,而在其他实施中,给定流的所有数据记录都可以是相同的大小。在至少一些实施中,SMS的节点(例如,摄取子系统节点和/或存储子系统节点)可负责生成SN 102。如下文进一步详细地描述,数据记录的序列号无需始终连续。在一个实施中,作为写入请求的一部分,客户端或数据生产者120可以提供将用于对应数据记录的最小序列号的指示。在一些实施方案中,例如,通过提供从中可获取数据部分的存储装置地址(例如,装置名和装置内的偏移)或网络地址(例如,URL),数据生产者120可以提交含有数据记录的数据部分的指针(或地址)的写入请求。

[0056] 在各实施方案中,流管理服务可以负责接收来自数据生产者120的数据、存储所述数据以及使得数据消费者130能够以一种或多种访问模式来访问数据。在至少一些实施方案中,流100可以被分区或“分片(sharded)”,以对接收、存储和检索数据记录的工作负载进行分配。在此类实施方案中,基于数据记录的一个或多个属性可以为输入数据记录110选择

分区或分片，并且基于分区可以识别用来摄取、存储或检索数据记录的具体节点。在一些实施中，数据生产者120可以为每个写入操作提供明确的分区键，所述分区键可用作分区属性，并且此类键可以映射到分区标识符。在其他实施中，基于数据生产者120的标识、数据生产者的IP地址等因素或者甚至基于提交的数据内容，SMS可以推断出分区ID。在数据流被分区的一些实施中，可以在每个分区的基础上分配序列号，例如，尽管序列号可以表示接收到特定分区的数据记录的顺序，但两个不同分区中的数据记录DR1和DR2的序列号可以不必表示接收到DR1和DR2的相对顺序。在其他实施中，可以在流宽度而不是每个分区的基础上分配序列号，从而如果分配给数据记录DR1的序列号SN1低于分配给数据记录DR2的序列号SN2，那么无论DR1和DR2属于哪个分区，这都将意味着DR1比DR2早被SMS接收到。

[0057] 在各实施方案中，SMS支持的检索或读取接口可允许数据消费者130按顺序和/或以随机顺序访问数据记录。在一个实施方案中，可以支持读取应用编程接口(API)的基于迭代器的集合。数据消费者130可以针对数据流提交获取迭代器的请求，其中迭代器的初始位置由指定的序列号和/或分区标识符指示。在启动程序被实例化之后，数据消费者可以提交从所述流或分区内的初始位置开始按顺序读取数据记录的请求。如果数据消费者希望以某一随机顺序读取数据记录，那么在此类实施方案中，每次读取都可能要实例化新的迭代器。在至少一些实施中，给定分区或流的数据记录可以按序列号顺序写入到基于磁盘的存储设备，一般使用避免磁盘寻道的顺序写入操作。顺序读取操作也可避免磁盘寻道的开销。因此，在一些实施方案中，可以使用价格激励来鼓励数据消费者更多地执行顺序读取而不是随机读取，例如，随机访问读取操作(例如，迭代器实例化)可比顺序访问读取操作具有更高的相关计费率。

[0058] 示例性系统环境

[0059] 图2提供根据至少一些实施方案的流管理系统(SMS)和流处理系统(SPS)的各个子部件之间的数据流动的概述，所述SPS包括流处理阶段的集合。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以及SMS控制子系统210。每个SMS子系统都可包括一个或多个节点或部件，例如，所述节点或部件使用在提供者网络的各个资源(或者客户端拥有的设施或第三方设施)处实例化的相应可执行线程或进程来实施，如下文所述。摄取子系统204的节点可以(例如，通过SMS控制子系统210的节点)被配置成基于针对流使用的分区策略而获取来自数据生产者120(例如，120A、120B和120C)的特定数据流的数据记录，并且每个摄取节点可以将接收的数据记录传递给存储子系统206的对应节点。根据针对流所选的持久性策略，存储子系统节点可将数据记录保存在各种类型的存储装置中的任一个上。检索子系统208的节点可响应于来自数据消费者(例如，SPS 290的工作者节点)的读取请求。SPS 290的流处理阶段215(例如，阶段215A、215B、1215C和215D)可以在SPS控制子系统220的帮助下建立起来。每个阶段215都可包括一个或多个工作者节点，所述工作者节点被SPS控制子系统220配置成对接收的数据记录执行处理操作的集合。如图所示，一些阶段215(例如，215A和215B)可以直接从SMS 280获取数据记录，而其他阶段(例如，215C和215D)可以从其他阶段接收到它们的输入。在一些实施方案中，多个SPS阶段215可以并行操作，例如，在阶段215A和215B处，可以同时对从相同流中检索的数据记录执行不同的处理操作。应注意，用于特定流的类似于图2所示那些的相应子系统和处理阶段也可针对其他流进行实例化。

[0060] 在至少一些实施方案中，图2所示的子系统和处理阶段的至少一些节点可以使用提供者网络资源来实施。如之前所述，由诸如公司或公共部门组织等实体设置的网络将可经由互联网和/或其他网络访问的一个或多个网络可访问服务(例如，各种类型的基于云的数据库、计算或存储服务)提供到分布式客户端集合，所述网络在本文中可称为提供者网络。服务中的一些可用来建立更高级别的服务，例如，计算、存储或数据库服务可用作流管理服务或流处理服务的构建块。提供者网络的至少一些核心服务可以打包以供客户端用在被称为“实例”的服务单元中：例如，被虚拟化计算服务实例化的虚拟机可代表“计算实例”，并且被存储服务实例化的存储装置(例如，块级容量)可被称为“存储实例”，或者数据库管理服务器可被称为“数据库实例”。实施提供者网络的各种网络可访问服务的此类单元的计算装置，例如，服务器，在本文中可被称为“实例主机”或者更简单称为“主机”。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，并且特定实例主机处的计算实例的集合可以用来实施节点用于一个或多个客户端的各种不同的流。在一些实施方案中，存储实例可用于存储各个流的数据记录，或者作为流处理阶段的结果的目的地。随着时间的推移，控制子系统节点可以响应于各种触发条件而动态更改其他子系统的群体，例如，通过添加或移除节点、改变节点到进程或计算实例或者实例主机的映射或者对给定的流进行重新分区而同时仍继续接收、存储和处理数据记录，如下文参考图15和图16所述。

[0061] 在提供者网络资源用于流相关操作的实施方案的上下文中，术语“客户端”在用作给定通信的源或目的地时，可以指由能够访问和利用提供者网络的至少一个网络可访问服务的实体(例如，组织、有多个用户的群组或者单个用户)拥有、管理或者分配给所述实体的计算装置、进程、硬件模块或软件模块中的任一个。一个服务的客户端本身可使用另一服务的资源来实施，例如，流数据消费者(流管理服务的客户端)可包括计算实例(虚拟化计算服务提供的资源)。

[0062] 给定的提供者网络可包括很多数据中心(它们可以分布在不同的地理区域)，所述数据中心托管了实施、配置和分布由提供者供应的基础设施和服务所需的各种资源池，例如，物理和/或虚拟化计算机服务器的集合、各自带有一个或多个存储装置的存储服务器、联网设备等等。在各实施方案中，许多不同的硬件和/或软件部件可以共同用来实施每一个服务，其中一些所述部件可在不同的数据中心或不同的地理区域被实例化或执行。使用位于客户端拥有或客户端管理的驻地处、或者提供者网络外部的数据中心的装置，和/或提供者网络内的装置，客户端可与提供者网络处的资源和服务交互。应注意，尽管提供者网络用作可以实施本文所述的很多流管理和处理技术的一个示例性上下文，但除了提供者网络之外，这些技术也可应用于其他类型的分布式系统，例如，应用于单个企业实体为自己的应用而操作的大规模分布式环境。

[0063] 编程接口实例

[0064] 如上文所述，在至少一些实施方案中，SPS可以利用SMS编程接口来建立更高级别的功能，所述功能可更容易被SPS客户端用来实施各种基于流的应用所需的业务逻辑。在考虑SPS与SMS功能之间的差异时，类比可能有帮助：SPS功能通常可以与使用更高级语言(例如，C++)的编程语言结构比较，而SMS功能通常可以与所述编程语言结构被编译程序转译成

的汇编语言指令比较。直接使用汇编语言指令来实施相同的操作是有可能的，但对于许多类的顾客或用户来说，采用更高级语言的编程一般可能更容易。类似地，有可能使用SMS提供的原语来实施各种应用，但使用SPS特征会更容易。SPS处理操作(例如，对数据记录执行的幂等处理操作)可对流记录的数据内容实施，而SMS操作被执行以获取、存储和检索记录本身，通常不需要考虑记录的内容。图3示出根据至少一些实施方案的可以在SMS和SPS处实施的编程接口的相应集合的实例。通过实例指出用于SMS和SPS两者的许多不同的应用编程接口(API)。所示API并不意图详尽列出任何给定的实施中支持的那些，并且给定的实施中可能不支持所示API中的一些。

[0065] 如箭头350所示，SPS客户端375可以调用SPS编程接口305，以配置处理阶段。在不同的实施方案中，可以实施各种类型的SPS编程接口305。例如，createStreamProcessingStage API可以使得客户端能够针对指定的输入流请求配置新的处理阶段215，从而所述阶段的工作者节点各自被配置成执行接口调用中指定的幂等操作集合，以及将结果分布到输出分布描述符或策略所指示的目的地。在createStreamProcessingStage API或其等效物的一些版本中，客户端也可请求创建输入流，而在其他版本中，可能需要在创建处理阶段之前就创建输入流。可以为工作者节点指定恢复策略，例如，表明要使用基于检查点的恢复技术还是优选最大努力恢复技术。在一些实施方案中，可以支持initializeWorkerNode API，以请求在指定阶段明确将工作者节点实例化。在实施基于检查点的恢复的实施方案中，可以支持saveCheckpoint API，以允许客户端请求由工作者节点生成进程记录。

[0066] 在不同的实施方案中，可以支持各种类型的SPS输出管理API，例如，setOutputDistribution API，借此客户端可以表明将使用在指定阶段执行的处理操作的结果来创建的一个或多个流，以及将用于新创建的流的特定分区策略。一些处理阶段可被配置成主要用于重新分区，例如，基于记录属性集A1将数据记录映射到N1分区的一个分区功能PF1可用于输入流S1，并且处理阶段可用来实施不同的分区功能PF2，以便将这些相同的数据记录映射到N2分区(使用不同的属性集A2或相同的属性集A1)。一些SPS API，例如linkStages，可用来配置包括多个阶段的任意图(例如，有向非循环图)。在一些实施方案中，可以支持到第三方或开源流处理框架或服务的连接器。在一个这样的实施方案中，SPS阶段可用来准备数据记录(例如，通过将在所述阶段处执行的处理操作的结果适当格式化)，以便由现存的第三方或开源系统消费。在所述实施方案中，诸如createThirdPartyConnector等API可用来设立此类连接器，并且将SPS阶段的结果适当转换成与第三方系统兼容的格式可由作为createThirdPartyConnector调用的结果而实例化的一个或多个连接器模块执行。

[0067] SPS可以调用SMS API 307，以执行它的至少一些功能，如箭头352所示。例如，在所述实施方案中，SMS API 307可包括createStream和deleteStream(分别用以创建和删除流)以及getStreamInfo(用以获取流的元数据，例如，负责给定分区的各种类型的节点的网络地址)。putRecord接口可用来写入数据记录，而getIterator和getNextRecords接口可分别用于非顺序读取和顺序读取。在一些实施方案中，repartitionStream接口可用来请求指定流的动态重新分区。希望如此的客户端370可以直接调用SMS SPI 307，如箭头354所示。如之前所述，在其他实施方案中，也可实施各种其他SMS和/或SPS API，并且在一些实施方

案中，可以不实施图3中列出的一些API。

[0068] 在各实施方案中，针对SPS或SMS，也可实施或改为实施不同于API的编程接口。此类接口可包括图形用户接口、web页或web站、命令行接口等。在一些情况下，基于web的接口或GUI可以将API用作构建块，例如，基于web的交互可导致在SMS或SPS的控制部件处调用一个或多个API。图4示出根据至少一些实施方案的示例性基于web的接口，所述接口可以被实施以使得SPS客户端能够生成流处理阶段的图形。如图所示，所述接口包括web页400，所述web页带有消息区402、图形菜单区404以及图形设计区403。

[0069] 在消息区402中，用户可被提供有关构建流处理图形的一般指令，以及链接，以使得用户学习关于流概念和原语的更多内容。在菜单区404中，许多图形图标可作为流处理图形工具箱的一部分被提供。例如，客户端可被允许，作为各个SPS处理阶段的输入或输出，将持久流451、短暂流452或者连接器453指示给第三方处理环境。至于实施基于web的接口的SPS/SMS，持久流451可被定义为数据记录被存储在永久存储装置(例如，磁盘、非易失性RAM或SSD)上的流，并且短暂流452可被定义为数据记录无需存储在永久存储装置上的流。例如，可从SPS阶段的输出创建短暂流，所述输出预期用作实施最大努力恢复策略的不同SPS阶段的输入。

[0070] 在示例性SPS图形构建web页400中，支持两种类型的处理阶段：阶段455，其中使用基于检查点的工作者节点恢复(例如，每个工作者节点不时地保存进程记录，并且在特定工作者节点出现故障的情况下，置换节点参考出现故障的节点的进程记录，以确定开始处理哪些数据记录)，以及阶段456，其中使用最大努力恢复(例如，置换工作者节点并不参考进程记录，而只是开始处理接收到的新数据记录)。通过点击图形构建区403中的对应图标，可输入有关在每个阶段执行的处理操作的细节，如消息区402中的指令所示。除了用于流、连接器和处理阶段的图标之外，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，以及指示可在提供者网络处实施的存储服务的节点的图标类型460，所述提供者网络的资源正用于处理阶段。

[0071] 在图4所示的示例性情形中，客户端已构建图形405，所述图形包括图形设计区403内的三个处理阶段412、415和416。被配置成使用基于检查点的恢复的处理阶段412将持久流411用作输入。阶段412的处理输出或结果被发送到两个目的地：采用形成阶段415的输入的不同持久流413的形式，以及采用形成阶段416的输入的短暂流414的形式。阶段415和416都将最大努力恢复策略用于它们的工作者节点。阶段415的输出采用短暂流的形式发送到存储服务节点419。阶段415的输出经由连接器417发送到第三方处理系统418。“保存图形”按钮420可用来保存处理阶段图形的表示，例如，采用任何适当的格式，例如，JSON(JavaScript对象表示法)、XML(可扩展标记语言)或者YAML。在各实施方案中，任意复杂的处理工作流程可使用类似于图4所示的那些工具构建而成。使用此类工具创建的工作流程随后可被激活，并且此类激活可导致调用SMS API，例如，以获取数据记录用于诸如图4的阶段412等处理阶段，在流411上可以调用getIterator和/或getNextRecords接口。

[0072] 图5示出根据至少一些实施方案的可以在SMS处实施的编程记录提交接口和记录检索接口的实例。在所述实施方案中，数据记录，例如所示DR 110K和110Q，可经由各种类型的编程摄取接口510提交到SMS。在一些实施方案中，DR 110可包括四种类型的元素：流标识符，例如，501A(用于流“S1”)或501B(用于流“S2”)；记录的数据或主体的指示；任选的分区

键504(例如,504A或504B);以及任选的排序偏好指示符506(例如,排序偏好指示符506A和506B)。数据本身可以在一些数据记录中以串联的方式提供(例如,DR 110K的串联数据502),而对于其他数据记录而言,可以提供指针或地址503,从而向SMS表明网络可访问的位置(或者不需要网络传输的本地装置的地址)。在一些实施方案中,给定的流可以支持串联和按引用(基于地址)的数据记录提交两者。在其他实施方案中,给定的流可以要求数据生产者以串联方式供应所有数据或者以引用方式供应所有数据。在一些实施中,数据记录提交可包括用于记录的分区标识符。

[0073] 在所述实施方案中,基于分区策略,输入数据记录110可以指向相应的摄取和/或存储节点。类似地,记录检索也可基于分区,例如,一个或多个检索节点可被指定响应于指向给定分区的记录的读取请求。对于一些流而言,可以要求数据生产者为每个数据记录写入请求提供明确的分区键。对于其他流而言,SMS能够根据分区方案来分布数据记录,所述分区方案依赖于元数据或者不同于明确供应的分区键的属性,例如,涉及提交数据生产者的标识信息可用作分区键,或可以使用提交数据生产者的IP地址的一部分或全部,或者可以使用被提交的数据的一部分。在一些实施中,例如,可将散列函数应用于分区键,以获取一定大小的整数值,例如,128位整数。该大小的正整数的总范围(例如,从0到2^128-1)可被分成N个连续的子范围,其中每个子范围表示相应的分区。因此,在此类实例中,针对数据记录确定或供应的任何给定的分区键将散列到对应的128位整数,并且该整数所属的128位整数的连续子范围可以指示数据记录所属的分区。下文参考图15提供有关分区策略及其使用的进一步细节。

[0074] 负责摄取或接受特定分区的数据记录、存储数据记录以及响应于针对特定分区的读取请求的节点的集合统称为ISR(摄取、存储和检索)节点用于图5中的分区。符号Sj-Pk用来指示流Si的第k个分区。在所示实施方案中,ISR节点520A被配置成摄取、存储和检索分区S1-P1的记录,ISR节点520B被设置用于分区S1-P2的记录,ISR节点520C被设置用于分区S1-P3的记录,ISR节点520K被设置用于分区S2-P1的记录,并且ISR节点520L被设置用于分区S2-P2的记录。在一些实施方案中,摄取子系统、存储子系统或检索子系统的给定节点可被配置成处理一个以上分区(或者一个以上流的一个以上分区)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被一个以上节点摄取、存储或检索。在至少一些情况下,指定用于给定分区Sj-Pk的摄取节点的数量可以不同于指定用于不同分区Sj-P1的摄取节点的数量,并且也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。至于摄取和/或检索,在一些实施方案中,SMS控制节点可以实施API(例如,getStreamInfo),以允许客户端确定哪些节点负责哪些分区。随着时间的推移,可以更改数据记录与分区之间以及分区与所配置的ISR节点(或控制节点)之间的映射,如下文在有关动态重新分区的论述中有所描述。

[0075] 在一些实施方案中,若干不同的编程接口580可实施用于从给定分区中检索或读取流数据记录。如图5所示,一些检索接口581可实施用于非顺序访问,例如,getIterator(用以在带有指定序列号的数据记录处或之后,将迭代器或读取游标实例化)或getRecord(用以读取带有指定序列号的数据记录)。其他检索接口582可实施用于顺序检索,例如,getNextRecords(该接口请求按照递增序列号的顺序从迭代器的当前位置读取N个记录)。在基于旋转磁盘的存储系统中,如之前所述,顺序I/O在很多情况下可能比随机I/O更有效,

这是因为顺序I/O的平均每个I/O所需的磁头寻道的数量一般可比随机I/O少。在很多实施方案中，给定分区的数据记录可以按序列号顺序写入，并且因此，基于序列号次序的顺序读取请求(例如，使用getNextRecords或类似接口)可比随机读取请求更有效。因此，在至少一些实施方案中，顺序与非顺序检索接口可以设置不同的计费率，例如，客户端可能要为非顺序读取支付更多费用。

[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。在所述实施方案中，摄取操作在逻辑上被分成前端和后端功能，其中前端功能涉及与数据生产者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/后端分隔可具有若干优点，例如，提高存储子系统的安全性，和避免将分区策略细节提供给数据生产者。SMS客户端库602可以被提供安装在各个数据生产者120处，并且数据生产者可以调用包括在库602中的编程接口，以提交数据用于摄取。例如，在一个实施方案中，数据生产者120可包括在提供者网络的数百或数千物理和/或虚拟服务器处实例化的日志或监控代理。此类代理可以收集其相应服务器处的各种日志消息和/或度量，并且定期将收集的消息或度量提交到由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，可以为负载分配器建立一个或多个虚拟IP地址(VIP)，其中数据生产者可以将流数据提交到所述地址。在一个实施中，循环DNS(域名系统)技术可用于VIP，以便从若干等效配置的负载分配器中选择特定的负载分配器，其中数据将由数据生产者120发送到所述负载分配器。

[0078] 在所述实施方案中，接收的数据记录可以指向若干前端节点606(例如，606A、606B或606C)中的任一个。在至少一些实施方案中，负载分配器604可能没有意识到用于数据记录的分区策略650，并且因此，可以使用循环负载平衡(或者某一其他的通用负载平衡算法)而不是基于分区的负载平衡来为给定的数据记录选择前端节点606。前端节点606可以意识到用于各个流的分区策略650，并且可与摄取控制节点660交互，以获取被配置用于给定分区的数据记录的特定后端摄取节点608(例如，608A、608B或608C)的标识。因此，在所述实施方案中，基于数据记录所属的相应分区，前端节点604可各自将数据记录传输到多个后端节点606。如之前所述，可基于各种因素的任何组合来确定数据记录所属的分区，例如，数据生产者供应的分区键、诸如数据生产者的标识或地址等一个或多个其他属性或者数据的内容。

[0079] 后端节点606可各自接收属于一个或多个流的一个或多个分区的数据记录，并且将数据记录传输到存储子系统的一个或多个节点。在经由HTTP(超文本传输协议)“PUT”web服务API提交数据的一些实施方案中，后端节点可被称为“PUT服务器”。通过将查询提交到控制节点660，给定的后端节点可确定其数据记录将被传输到的存储子系统节点的集合(在由单独节点集合处理不同子系统的控制功能的实施方案中，所述控制节点继而可将对应的查询提交到存储子系统的控制节点)。

[0080] 在至少一些实施方案中，可以支持很多不同的摄取确认策略652，例如，至少一次摄取策略或最大努力摄取策略。在至少一次策略中，数据生产者120可以要求对提交的每个数据记录进行肯定确认，并且可以重复提交相同的数据记录(如果未接收到首次提交的确认的话)，直到最终接收到确认为止。在最大努力摄取策略中，可以不要求对至少一些提交

的数据记录进行肯定确认(但摄取子系统仍可提供偶尔的确认,或者可以响应于数据生产者对确认的明确请求)。在摄取子系统204被要求将确认提供给数据生产者的一些实施方案中,生成确认之前,负责给定数据记录的后端摄取节点608可以等到已经在存储子系统处成功创建数据记录的所需数量的复本为止(例如,根据针对所述流建立的持久性策略)。在各实施方案中,摄取子系统可以为每个接收到的数据记录生成序列号,例如,表明相对于同一分区或流的其他记录而言,摄取所述记录的顺序,并且此类序列号可以作为确认或确认的一部分而返回到数据生产者。下文参考图13a和图13b提供有关序列号的进一步细节。在一些实施中,确认和/或序列号可经由前端节点606传输回到数据生产者。在至少一个实施中,可以在摄取子系统本身的前端与后端节点之间实施至少一次策略,例如,给定的前端节点606可以将数据记录重复提供到合适的后端节点608,直到后端节点提供确认为止。

[0081] 除了其他功能外,摄取控制节点660可负责将前端和后端节点实例化、监控节点的健康和工作负载水平、根据需要来协调故障转移、提供对有关哪些节点负责给定分区的查询的响应或者对策略相关查询的响应,以及负责因流的动态重新分区而造成的摄取相关配置操作。在一些实施方案中,指定用于一个或多个流的给定集合的摄取控制节点的数量本身可随着时间的推移而改变,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。在为摄取前端或后端节点设置冗余组的一些实施方案中,如下文参考图9和图10进一步详细地描述,控制节点660可负责:记录哪些节点是主要节点以及哪些是不主要的;检测故障转移的触发条件;以及在需要故障转移时选择置换。应注意,在一些实施方案中,可以不实施图6所示的多层摄取子系统架构,例如,在一些情形下,可以只配置摄取节点的单个集合。

[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,在由不同节点集合承担前端和后端摄取责任的实施方案中,是后端摄取节点)可将流的一个或多个分区的数据记录传输到被配置用于这些分区的相应存储节点702。例如,分区S1-P1的数据记录110A被发送到存储节点702A,分区S2-P3的数据记录110B被发送到存储节点702B和702C,分区S3-P7的数据记录110C被发送到存储节点702D,以及分区S4-P5的数据记录110D最初被发送到存储节点702E。在所述实施方案中,存储控制节点780可负责:执行应用于不同流的数据记录的持久性策略750;根据需要来配置和重新配置存储节点;监控存储节点状态;管理故障转移;响应于存储配置查询或存储策略查询;以及各种其他管理任务。

[0084] 在不同的实施方案中,持久性策略750彼此在各个方面可存在差异。例如,应用于流Sj的持久性策略P1与应用于Sk的策略P2的不同之处可在于:(a)将被存储的每个数据记录的复本数量;(b)用于存储复本的存储装置或系统的类型(例如,复本是否存储在易失性存储器、非易失性缓存、基于旋转磁盘的存储设备、固态驱动器(SSD)、各种存储用具、各种RAID(廉价磁盘冗余阵列)中,存储在数据库管理系统中,存储在提供者网络实施的存储服务的节点处等等);(c)复本的地理分布(例如,通过将复本放在不同的数据中心,流数据是否能从大规模故障或某些类型的事故中复原);(d)写入确认协议(例如,如果要存储N个复本,那么在向摄取节点提供确认之前,必须要成功写入N个复本中的多少);和/或(e)在存储数据记录的多个复本的情况下,应并行创建还是按顺序创建复本)。在将存储多个复本的一

些情况下，如在数据记录110D的情况下，给定的存储节点可以将数据记录传输到另一存储节点(例如，存储节点702E发送数据记录110D以便进一步复制到存储节点702F，并且存储节点702F继续将它发送到存储节点702G)。在使用多复本持久性策略的其他情况下，如在将存储两个存储器内复本的数据记录110B的情况下，摄取节点可以并行开始多次复制。在至少一些实施方案中，客户端选择的持久性策略可以不指定将用于流数据记录的存储位置的类型；相反，SMS可以基于各种准则来选择适当类型的存储技术和/或位置，例如，基于成本、性能、与数据源的接近度、耐久性要求等等。在一个实施方案，针对给定流的不同分区或针对不同的流，客户端或SMS可以决定使用不同的存储技术或存储位置类型。

[0085] 在图7所示的实例中，应用于流S1(或至少流S1的分区S1-P1)的持久性策略是单复本存储器内策略，而对于流S2而言，应用并行双复本存储器内策略。因此，在存储节点702A处创建数据记录110A的存储器内复本704A，而在存储节点702B和702C处并行地创建对应于数据记录110B的两个存储器内复本705A和705B。针对流S3的数据记录110C，创建单个磁盘复本706A。针对流S4，可应用顺序磁盘三复本策略，并且因此，在存储节点702E、702F和702G处按顺序创建相应的磁盘复本707A、707B和707C。在不同实施方案中，可将各种其他类型的持久性策略应用于数据流。响应于数据消费者调用各种类型的检索API，检索子系统的节点可以获取来自适当存储节点的数据记录。

[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，以及检索子系统与SPS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，例如，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中的每个都可被配置成响应于来自各个客户端或数据消费者130(例如，SPS的工作者节点840)的流数据检索请求，如下文所述。在不同的实施方案中，检索节点可以实施多种编程检索接口802，例如，之前所述的非顺序和顺序检索接口。在一些实施方案中，诸如HTTP GET请求等web服务API可用于数据记录检索，并且检索节点802可相应地被称为GET服务器。在所述实施方案中，给定的检索节点802可被配置，例如，被检索控制节点880配置，以获取来自存储子系统节点702(例如，存储节点702A和702B)的适当集合的一个或多个流分区的数据记录。

[0088] 在所述实施方案中，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工作者节点840接收到的检索请求。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)由检索节点802A从存储节点702A中读取，并且分别被提供到工作者节点840A和840K。分区S6-P7的数据记录(例如，110M)由检索节点802B从存储节点702A中读取，并且被提供到工作者节点840K。分区S4-P7的数据记录由检索节点802C从存储节点702B中读取，并且被提供到工作者节点840B，而且还提供到其他数据消费者130(例如，直接调用SMS检索API而非经由SPS来与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中的一些或全部可实施相应的缓存804(例如，检索节点802A处的缓存804A、检索节点802B处的缓存804B，以及检索节点802C处的缓存804C)，其中预期未来的检索请求，可临时保留各个分区的数据记录。检索控制节点880可负责实施一些检索策略882，例如，包括缓存策略(例如，应针对给定的分区配置多大的缓存，应缓存多长的数据记录)、存储节点选择策略(例如，在存储数据记录的多个复本的情形下，

应先接触哪个特定的存储节点来获取给定的数据记录)等等。此外,检索控制节点可负责:实例化和监控检索节点802;响应于有关哪些检索节点负责哪些分区的查询;启用或响应于重新分区操作等等。

[0090] 在所示实例中,SPS 290包括两个处理阶段215A和215B。SPS控制节点885可负责将在各个处理阶段215处的工作者节点804实例化,例如,将工作者节点840A实例化以便处理分区S4-P5的记录、将工作者节点840B实例化以便处理分区S4-P7的记录,以及将工作者节点840K实例化以便处理分区S5-P8和S6-P7的记录。SPS控制节点885可以实施编程接口(例如,图3和图4所示的那些),从而使得SPS客户端能够设计处理工作流程。针对不同的处理阶段或工作流程,可以实施各种检查点策略850,表明工作者节点何时或是否存储指示它们处理其相应分区的程度的进程记录、用于进程记录的存储装置的类型等等。故障转移/恢复策略852可表明导致用不同的节点置换工作者节点的触发条件或阈值,以及针对给定的处理阶段,使用最大努力恢复还是使用基于检查点的恢复。在至少一些实施方案中,SPS控制节点885可以与各种类型的SMS控制节点交互,例如,以识别从中获取给定流的数据记录的检索节点、以建立特定处理工作流程可能需要的新的短暂流或持久流等等。在至少一个实施方案中,客户端可以与SPS控制节点交互,以对流进行实例化,例如,替代于使用SMS控制接口,一些客户端可能希望只调用更高级的SPS接口。应注意,尽管图6、图7和图8中示出用于SMS摄取、存储和检索子系统以及用于SPS阶段的控制节点的单独集合,但在至少一些实施方案中,给定的控制节点可用于所述子系统和/或SPS中的若干个。

[0091] 节点冗余组

[0092] 在至少一些实施方案中,节点的冗余组可被配置用于SMS的一个或多个子系统。也就是说,例如,替代于为检索流分区Sj-Pk的数据记录配置一个检索节点,可以为此类检索建立两个或更多节点,其中一个节点在给定时间点处被授予“主”角色或活跃角色,而另一或其他节点被指定为“非主”节点。当前的主节点可负责响应于工作请求,例如,从客户端或其他子系统的节点接收的请求。一个或多个非主节点可保持休眠,直到(例如)因故障、与主节点失去连接或其他触发条件而触发故障转移为止,此时,控制节点可以通知所选择的非主节点来接管先前的主节点的责任。因此,在故障转移期间,主角色可从当前的现任主节点撤回,并被授予当前的非主节点。在一些实施方案中,当确定发生故障转移时,例如,可能无需明确的通知,非主节点本身便可以接管作为主节点。在各实施方案中,此类节点的冗余组可设置用于SMS处的摄取、存储、检索和/或控制功能,并且在至少一些实施方案中,针对SPS处的工作者节点,也可采用类似的方法。在一些实施方案中,包括用于给定功能的至少一个主节点和至少一个非主节点的此类组可被称为“冗余组”或“复制组”。应注意,存储节点的冗余组可以独立于存储的数据记录的物理复本的数量来实施,例如,将要存储的数据记录的复本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冗余组的实例,所述冗余组可以针对SMS或SPS的节点而设立。在所述实施方案中,对于给定的流分区Sj-Pk而言,针对摄取节点、存储节点、检索节点和控制节点来设置相应的冗余组(RG)905、915、925和935。在所示实施方案中,实施控制节点的公共RG 935,但在一些实施方案中,可实施用于摄取控制节点、存储控制节点或检索控制节点的单独RG。每个RG都包括主节点(例如,主摄取节点910A、主存储节点

920A、主检索节点930A以及主控制节点940A)和至少一个非主节点(例如,非主摄取节点910B、非主存储节点920B、非主检索节点920C以及非主检索节点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)和942(用于控制节点),主角色可被撤回并授予当前的非主节点。例如,故障转移策略可管理:导致主状态改变的触发条件、主节点或非主节点的健康状态是否和如何被监控、将在给定的冗余组中被配置的非主节点的数量等等。在至少一些实施方案中,单个RG可被建立用于多个分区,例如,RG 905可负责处理分区Sj-Pk以及Sp-Pq的记录的摄取。在一些实施中,被指定为一个分区的主节点的节点同时可被指定为另一分区的非主节点。在一个实施方案中,多个节点可同时被指定为给定RG内的主节点,例如,给定分区的摄取相关工作负载可以分布在两个主节点之中,其中如果任一主节点出现故障,则一个节点被指定为非主节点。在给定的RG中实例化的节点的数量可取决于对应功能所需的可用性水平或复原水平(例如,取决于所述组能够承受多少同时出现或重叠的故障)。在一些实施方案中,除了用于SMS节点之外或作为替代,冗余组可设置用于SPS处理阶段的工作者节点。给定RG的成员有时可在地理上分布,例如,分布在若干数据中心,如图10所示。在一些实施方案中,选择的控制节点可被配置成检测故障转移触发条件,例如,使用心跳机制或其他健康监控技术,并且通过选择适当的非主节点来置换出故障的主节点、通知/激活选择的置换节点等等,此类控制节点可以协调故障转移。

[0094] 在一些实施方案中,提供者网络可被组织成多个地理区域,并且每个区域可包括一个或多个可用性容器,所述可用性容器在本文中也可称为“可用性区域”。可用性容器又可包括一个或多个不同的位置或数据中心,其被设计成(例如,带有独立基础设施部件,例如,电力相关设备、冷却设备、物理安全部件)使得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。预期一个可用性容器中的故障可不导致任何其他可用性容器出现故障,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不同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。在单个位置,通过将多个应用实例放入相应的可用性容器,或者(在一些SMS和SPS的情况下)将给定冗余组的节点分布到多个可用性容器,可防止各种类型的应用出现故障。同时,在一些实施中,可以在位于同一地理区域内的资源(例如,用于SMS和SPS节点的主机或计算实例)之间提供廉价和低等待时间网络连接,并且同一可用性容器的资源之间的网络传输甚至可能更快。一些客户端可能希望指定预留和/或实例化它们的流管理或流处理资源的位置,例如,在区域水平、可用性容器水平或者数据中心水平,以便维持对运行应用的各个部件的准确位置的所需控制程度。其他客户端可能对预留或实例化其资源的准确位置不太感兴趣,只要资源满足客户端需求即可,例如,对性能、高可用性等的需求。在一些实施方案中,位于一个可用性容器(或数据中心)中的控制节点能够远程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,特定可用性容器或数据中心可能不需要让本地控制节点来管理SMS/SPS节点。

[0095] 图10示出根据至少一些实施方案的提供者网络环境,其中给定冗余组的节点可以分布在多个数据中心之间。在所述实施方案中,提供者网络1002包括三个可用性容器1003A、1003B和1003C。每个可用性容器都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SMS和/或SPS节点的许多不同的冗余组1012。一些

RG 1012可以完全实施在单个数据中心内，如位于数据中心1005A内的RG 1012A的情况。其他RG可以使用给定可用性容器内的多个数据中心的资源，例如，跨可用性容器1003A的数据中心1005A和1005B的RG 1012B。而其他RG可使用分散在不同可用性容器的资源来实施。例如，RG 1012C使用分别位于可用性容器1003A和1003B的数据中心1005B和1005C的资源，并且RG 1012D利用分别在可用性容器1003A、1003B和1003C的数据中心1005B、1005C和1005D的资源。在一个示例性部署中，如果RG 1012包括一个主节点和两个非主节点，那么三个节点中的每个可位于不同的可用性容器中，从而确保即使在两个不同的可用性容器处同时出现大规模故障事件，至少一个节点也很可能保持功能。

[0096] 在所述实施方案中，分别与SMS和SPS相关联的控制台服务1078和1076可提供容易使用的基于web的接口，以用于配置提供者网络1002中的流相关设置。使用分散在一个或多个数据中心或者一个或多个可用性容器的资源，可在提供者网络1002中实施许多额外服务，其中至少一些服务可被SMS和/或SPS使用。例如，可实施虚拟计算服务1072，从而使得客户端能够利用所选量的计算能力，所述计算能力被打包作为各个不同能力水平的计算实例，并且此类计算实例可用来实施SMS和/或SPS节点。可实施一个或多个存储服务1070，从而使得客户端能够(例如)经由块装置卷接口或经由web服务接口来存储和访问具有所需的数据耐久性水平的数据对象。存储对象可附加到服务1072的计算实例或者可从中访问，并且在一些实施方案中，可用来在SMS存储子系统处实施各种流持久性策略。在一个实施方案中，一个或多个数据库服务例如高性能键值数据库管理服务1074或关系型数据库服务可以在提供者网络1002处实施，并且此类数据库服务可用于由SMNS存储子系统来存储流数据记录，和/或用于存储控制子系统、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，可以为SMS和/或SPS的用户提供数据流的许多安全相关选项，从而使得客户端能够选择用于各种功能类别(例如，摄取、存储、检索、处理和/或控制)的资源(例如，虚拟机或物理机)的安全配置文件。例如，此类选项可包括有关用于各个节点的资源的物理位置的类型的选择(例如，使用提供者网络设施还是使用客户端拥有的设施，所述客户端拥有的设施可与提供者网络设施具有不同的安全特性)、有关流数据的加密的选择和/或流处理基础设施的各个部分中的网络隔离选择。例如，一些客户端可能关心入侵者或攻击者获取对有价值的专有业务逻辑或算法的访问权的可能性，并且可能希望使用客户端拥有的驻地内的计算装置来实施流处理工作者节点。将用于实施SMS和/或SPS节点的集合的资源类型在本文中可称为那些节点的“放置目的地类型”。图11示出根据至少一些实施方案的多个放置目的地类型，所述放置目的地类型可被选择用于SMS或SPS的节点。

[0099] 在所述实施方案中，针对一些类型的SMS/SPS功能类别(例如，摄取、存储、检索、控制或处理)，可以在提供者网络1102内选择放置目的地，而针对其他类型的SMS/SPS功能类别，可在提供者网络1102的外部进行选择。在提供者网络1102内，可使用多租户实例主机1103来实施一些资源，例如，计算实例、存储实例或者数据库实例。此类多租户实例主机可形成放置目的地类型的第一类别“A”，其中可在每个所述实例主机处将用于一个或多个客户端的SMS或SPS节点实例化。为了避免与其他客户端共享物理资源，一些客户端可请求使用被限于单个客户端的实例主机来实施其SMS/SPS节点。此类单租户实例主机可形成放置

类别类型“B”。出于若干原因，从一些客户端的角度，可优选单租户实例主机。由于多租户实例主机可包括属于其他客户端的计算实例，因此，与在单租户实例主机中相比，来自多租户实例主机中的另一客户端实例的安全攻击的概率可能更高。此外，当使用单租户实例主机时，也可避免“邻居吵扰”的现象，其中在多租户主机上运行的一个客户端的计算实例CI1经历工作负载的激增并且开始消耗一大部分的主机计算循环或其他资源，从而有可能影响在不同的计算实例CI2上运行的另一客户端应用的性能。

[0100] 在所述实施方案中，隔离虚拟网络(IVN)1106，例如，IVN 1106A和1106B，可以表示放置目的地类型的另一类别“C”。在一些实施方案中，可根据提供者网络客户端的请求来创建IVN 1106，所述IVN1106作为专用网络的逻辑等效物，是使用提供者网络资源构建的，但具有主要被客户端控制的网络配置。例如，客户端可决定用在IVN 1106内的IP地址，而无需关心复制可能已在IVN之外使用的IP地址的可能性。在所述实施方案中，在一个或多个IVN中实施各种类型的SMS和SPS节点可将额外的网络安全水平添加到客户端流数据的管理和/或处理。在一些情况下，给定的客户端可能希望将SMS/SPS节点的一个功能类别放在一个IVN 1106中，并且将不同的功能类别放在不同的IVN中。在各实施方案中，给定的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，使用提供者网络的资源的放置目的地类型选择(或安全配置文件选择)的另一集合(图11中未示出)可用于至少一些客户端。在针对流相关操作客户端可获取并使用来自提供者网络的虚拟化计算服务的计算实例的实施方案中，计算实例可采用两个模式中的一个进行使用。在一个模式中，客户端可将一个或多个可执行程序提供到SPS或SMS，以便在被配置成SPS工作者节点的计算实例处运行(或者在摄取、存储或检索节点处运行)，并且让SMS或SPS运行程序和管理节点。此第一模式可被称为使用用于流操作的计算实例的“流服务管理”模式。在另一模式中，客户端可能希望运行可执行程序并且管理计算实例，其中SPS或SMS的支持较少。此第二模式可被称为使用用于流操作的计算实例的“客户端管理”模式。因此，对于客户端可选择的放置目的地类型或安全配置文件而言，这两种操作模式可以表示额外的选择。例如，如果可执行程序可能会要求由客户端组织的主题专家可最好地执行的除错(包括单步)，那么客户端可以选择客户端管理模式，而流服务管理模式可以是不太可能要求除错的更成熟代码的合理选择。在一些实施方案中，可将不同的价格策略应用于这两种模式。

[0101] 在图11所示的实施方案中，提供者网络外部的设施处可以支持许多放置选项。例如，安装有SMS库1171和/或SPS库1172的主机1160可用于来自客户端设施(例如，客户端拥有的数据中心或驻地)1110A或1110B的流管理或处理，其中两种类型的客户端设施的不同之处是连接到提供者网络的方式。客户端设施1110A经由至少一些共享的互联网链路1151链接到提供者网络1102(即，其他实体的网络流量也可流经客户端设施1110A与提供者网络1102之间的一些链路)。相反，一些客户端设施(例如，1110B)可经由特殊的独享专用物理链路1106(有时可称为“直接连接”链路)链接到提供者网络。这两种不同类型的客户端驻地包括分别使用图11中使用的术语的放置目的地选项“D”和“E”。在一些实施方案中，SMS和/或SPS的部分也可在第三方设施(例如，由SMS/SPS的客户端使用但不被其拥有或管理的数据中心)处实施，并且此类第三方驻地可被指定为放置目的地类型“F”。在至少一些客户端和/或第三方驻地中，SMS和/或SPS库可能需要从提供者网络获取，并且安装在主机上，以用于SMS/SPS节点。在至少一个实施方案中，在适当的库的帮助下，可以在提供者网络的外部实

施所有不同功能类别的节点。

[0102] 在不同的实施方案中，不同放置目的地类型彼此之间的不同可在于各个安全相关方面，例如，实施的网络隔离特征、支持的入侵检测功能、实施的物理安全策略、支持的加密级别等等。因此，各个目的地类型中的每个都可被视作具有相应的安全配置文件，它与其他放置目的地的安全配置文件的不同可能在于一个或多个方面。在一些实施方案中，SMS和/或SPS的客户端可以编程方式为不同的子系统或节点集合选择相应的放置目的地类型，例如，通过将请求发送到SMS或SPS的一个或多个控制节点，如图12a和图12b所示。应注意，在一些实施方案中并且对于某些类型的流应用而言，不但出于安全原因，而且出于性能和/或功能原因，客户端可能希望控制放置目的地类型。例如，通过使用专用的客户端驻地资源或单租户实例主机，可以避免上述邻居吵扰的现象。在一些实施方案中，客户端可具有它们想要用于SPS阶段或SMS节点的专用或专有硬件和/或软件，其中使用此类部件可达到的功能能力或性能水平不容易在提供者网络处复制，或者提供者网络根本不支持。客户端可以在外部数据中心处访问具有超级计算机级别的处理能力的计算机服务器，例如，与单独使用提供者网络资源可能达到的速率相比，所述计算机服务器能够以更高的速率来执行SPS处理。使得客户端能够为各个节点选择放置目的地可允许使用此类专用装置或软件。

[0103] 图12a和图12b示出根据至少一些实施方案的可分别由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，其中对于带有标识符1210的一个或多个处理阶段而言，客户端表明针对所述阶段的控制节点请求的放置目的地类型(PDT)(元素1212)以及针对工作者节点请求的PDT(元素1214)。在至少一个实施方案中，客户端也能够提交请求，以针对它们的流数据记录或流处理结果来配置加密设置，例如，通过请求在经由各种网络链路传输之前，使用指定算法或协议对数据记录进行加密，或者请求加密各种控制或管理交互。例如，在图12a中，用于所述阶段的加密设置可指明将应用于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者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包括许多元素，所述元素指明客户端对带有指定标识符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元素1254、1258和1262中指明。对于摄取控制节点、存储控制节点和检索控制节点的PDT偏好可分别由元素1256、1260和1264指明。数据记录的加密偏好可经由元素1266指明，例如，是否和/或如何在数据记录从一个类别的节点传输到另一个时实施加密。使用如图12a和图12b所示那些的安全选项请求，客户端能够选择位置(例如，在提供者网络内或者在提供者网络外部)，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，出于安全性之外的原因，可提供节点放置目的地的选择。例如，出于性能原因，客户端可能希望在单租户主机处实施一些类型的SMS或SPS节点(例如，以避免之前所述的“邻居吵扰”的问题，而不是主要出于安全性原因)。在至少一些实施方案中，放置选择可以在流的使用期限内改变，例如，客户端最初可以允许SMS节点在多租户实例主机处被实例化，但之后可能希望将节点的至少一些子集移动到单租户实例主机。在至少一些实施方案中，可以将不同的价格策略应用于不同的安全相关选项，例如，在IVN处实施特定功能类别的SMS节点可能比在IVN外部的多租户实例主机处实施的成本更

高，或者在单租户实例主机处实施SMS节点可能比在多租户实例主机处实施的成本更高。

[0106] 流记录的顺序存储和检索

[0107] 对于很多类型的流应用而言，在SMS处可以非常高的速率接收来自多个数据生产者120的数据记录，并且数据消费者一般可能希望按照生成记录的顺序来访问存储的数据记录。尤其是在将旋转磁盘用作流数据记录的存储装置的环境中，如之前所述，顺序I/O访问模式(针对读取和写入而言)可具有优于随机I/O访问模式的显著性能优势。在若干实施方案中，流特有或分区特有的序列号可以在SMS接收到数据记录时分配给所述数据记录，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者与SMS的摄取子系统之间的示例性交互。流数据生产者可以将数据记录110提交给摄取子系统，并且在所述实施方案中，摄取子系统可以用针对提交的记录选择的序列号102作出响应。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一部分，例如，在此类实施方案中，根据适用的持久性策略，可以在存储接收到的数据记录之后确定序列号102，并且存储子系统可以针对数据记录生成自己的数字序列指示符，并且提供所述指示符，以包含到由摄取节点分配给数据记录的更大序列号中。

[0108] 在各实施方案中，可以实施序列号，以提供数据记录的稳定一致的次序，并且使得数据消费者能重复迭代记录。分配给特定分区的数据记录的序列号可以随着时间的推移而单调增加，但在至少一些实施中，所述序列号无需连续。在各实施方案中，序列号可被分配有下列语义的至少某一子集：(a)序列号在流内是唯一的，即，给定流的两个数据记录不会被分配相同的序列号；(b)序列号可用作流的数据记录的索引，并且可用来在给定流分区内迭代数据记录；(c)对于任何给定的数据生产者而言，数据生产者成功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)具有给定分区键值的数据记录的序列编号保留动态重新分区操作上的单调递增语义，例如，在重新分区之后分配给带有分区键值K1的数据记录的序列号可各自大于在动态重新分区之前分配给带有该分区键值K1的数据记录的序列号的任一。(下文参考图16进一步详细描述动态重新分区。)

[0109] 在一些实施方案中，数据生产者可能希望影响针对至少一些数据记录选择的序列号102的选择。例如，数据生产者120可能希望在流的分配序列号内界定边界或分隔符，从而使得所述流的数据消费者更容易提交针对流的特定子集的读取请求。在一些实施中，数据生产者120可以与记录一起提交最小序列号的指示，并且SMS可以根据也符合上述序列号语义的请求最小值来选择序列号。

[0110] 图13b示出根据至少一些实施方案的可以针对SMS处的摄取数据记录生成的序列号的示例性元素。在所述实施方案中，序列号可包括四个元素：n1位SMS版本号1302、n2位时间戳或出现时间(epoch)值1304、n3位序列号1306以及n4位分区号1308。在一些实施中，可以使用128位序列号，例如，n1、n2、n3和n4可以分别是4、44、64和16位。版本号1302可只用来避免SMS软件版本推出的混乱，例如，从而容易分清使用SMS软件的哪个版本来生成序列号。在至少一些实施中，预期版本号1302可以不频繁改变。例如，可通过摄取子系统节点从本地时钟源或全球可访问时钟源(例如，实施getCurrentEpoch或getCurrentTime API的提供者网络的状态管理系统)获取时间戳值1304。在至少一些实施中，与公知时间点的偏移量(例如，自1970年1月1日UTC AM 00:00:00起经过的秒数，这可通过调用基于Unix™的操作系统中的各种时间相关系统调用来获取)可用于时间戳值1304。在一些实施方案中，序列号1036

可由存储子系统生成，并且可指明特定分区的数据记录被写入存储装置的顺序。因此，在给定秒内接收到很多数据记录并且时间戳值1304只以约一秒的间隔改变的实施中，序列号1306可以用作数据记录的记录到达(或存储)顺序的指示符，所述数据记录正好在同一秒内到达，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可以唯一识别给定流内的分区。在序列号时间戳(至少大约)表明摄取对应数据记录的时钟时间的至少一些实施中，序列号可用于某些类型的基于时间的检索请求的索引机制。例如，客户端可能希望检索在特定的一天或在指定时间范围期间生成或摄取的流记录，并且序列号可用作隐性辅助索引的键，以检索数据记录的适当集合。因此，在至少一些实施方案中，使用含有用于顺序存储和检索的时间戳的序列号可以具有下列额外效益：将时间索引提供到存储数据记录的集合中。

[0111] 给定分区的数据记录一般可以按序列号顺序写入(例如，写入到磁盘)，通常使用较大的顺序写入操作。在一些实施方案中，如之前所述，可以实施基于迭代器的编程接口，以允许数据消费者按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处的流数据记录的顺序存储和检索的实例。所示分区Sj-Pk(流Sj的第k个分区)的六个数据记录110A到110F按序列号顺序存储。如图所示，在至少一些实施方案中，序列号可以不是连续的，例如，因为将值分配给上述时间戳部分1304或者序列号1306的方式可能并不始终形成这些元素的连续值。

[0112] 在图14所示的实例中，数据消费者请求创建迭代器，指定起始序列号“865”。响应于请求，SMS将Iterator1初始化，所述Iterator1位于具有大于或等于所请求的起始序列号的最近序列号的数据记录处。在这种情况下，已经选择具有序列号870的数据记录110C作为迭代器的起始位置，因为接下来的下面序列(分配给数据记录110B的860)小于消费者请求中的起始序列号。getIterator接口可被视作在分区内的请求位置设置游标的请求的逻辑等效物，并且随后，getNextRecords接口可用来从游标位置开始读取数据记录，例如，以便沿着所述流按序列号顺序移动游标。在所示实例中，数据消费者已经调用getNextRecords接口，其中参数“迭代器”设置为Iterator1并且“maxNumRecords”(返回的数据记录的最大数量)设置为3。因此，SMS检索子系统按所述顺序将数据记录110C、110D和110E返回到数据消费者。在getNextRecords调用完成之后，迭代器Iterator1可以移动到新的位置，例如，移动到数据记录110F，并且用于相同迭代器的随后getNextRecord调用可以返回从110F开始的数据记录。在一些实施方案中，getIterator调用的语义可以不同，例如，在一些实施方案中，替代于将迭代器定位在具有大于或等于指定序列号的最近序列号的数据记录处，所述迭代器可以定位在具有等于或小于所请求的序列号的最高序列号的最近数据记录处。在另一实施方案中，客户端可以指定getIterator调用中的现有序列号，例如，如果流中不存在具有所请求的序列号的记录，则可返回错误。

[0113] 分区映射

[0114] 如之前所述，在各实施方案中，根据各种分区和重新分区策略，与给定流的记录的摄取、存储、检索和处理相关的工作负载可以细分并且分布在若干节点之间。图15示出根据至少一些实施方案的可以针对SMS和SPS节点作出的流分区映射1501和对应配置决策的实例。当(例如)响应于createStream API的客户端调用，创建或初始化特定数据流时，可针对所述流来激活分区策略，从而可用来确定分区，所述流的任何给定数据记录应被视作所述

分区的成员。基于记录的分区，可以选择摄取子系统204、存储子系统206、检索子系统208以及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。在一个实施方案中，也可基于分区来选择用于给定数据记录的控制节点的至少子集。在至少一些实施方案中，例如，响应于策略中指示的触发条件或者响应于明确请求，作为分区策略的一部分，可以支持数据流的动态重新分区。

[0115] 在各实施方案中，针对给定数据记录选择的分区可以取决于所述记录的分区键，所述分区键的值可以由数据生产者直接供应(例如，作为写入或输入请求的参数)或间接供应(例如，SMS可以将元数据用作分区键，例如，数据生产者客户端的标识符或名称、数据生产者的IP地址或者数据记录的实际内容的部分)。在图15所示的实施方案中，一个或多个映射功能1506可以应用于数据记录分区键或属性1502，以确定数据记录分区标识符1510。在一个实施中，例如，给定的分区标识符1510可以表示128位整数值的空间上的连续范围，从而使得所述流的所有分区的范围的结合可以覆盖128位整数可以呈现的所有可能的值。在此类示例性情形中，一个简单的映射功能1506可以从数据记录的分区键值或选择的属性值中生成128位散列值，并且基于散列值所属的特定连续范围，可以确定分区标识符。在一些实施中，连续范围的大小至少最初可以相等；而在其他实施中，不同的分区可以对应于大小彼此可不同的连续范围。在一个实施中，重新分区还可导致范围边界的调整。在不同的实施中，可以使用其他分区功能106。

[0116] 如果数据流经历动态重新分区(如下文进一步详细地论述)，则可以改变带有特定键的记录所映射到的分区。因此，在至少一些实施方案中，SMS和/或SPS控制节点可以在流的使用期限期间记录应用于流的若干不同映射。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控制节点存储，以用于每个分区映射。例如，时间戳有效范围1511可以表明从流的创建时间直到时间T1应用特定映射M1、从T1到T2应用不同的映射M2等等。当响应于指向流的读取请求时，检索节点可能需要首先确定使用哪个映射(例如，根据读取请求中指示的序列号)，并且随后使用所述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控制节点可负责以若干不同的粒度将分区映射到资源。例如，如图15的示例性实施1599所示，在一个实施中，每个摄取、存储、检索或处理(工作者)节点可以实施作为服务器虚拟机(例如，Java™虚拟机(JVM)或计算实例)内的相应执行进程或相应执行线程，并且每个JVM或计算实例可以在特定的物理主机处实例化。在一些实施方案中，多个JVM可以放在单个计算实例内，从而添加另一层资源映射决策。例如，针对给定的分区，一个或多个控制节点可以选择将哪些特定资源用作摄取节点1515、存储节点1520、检索节点1525或者处理阶段工作者节点1530(例如，分别用于阶段PS1或PS2的节点1530A或1530B)。控制节点还可确定这些节点到服务器的映射(例如，摄取服务器1535、存储服务器1540、检索服务器1545或者处理服务器1550)，以及服务器与主机之间的映射(例如，摄取主机1555、存储主机1560、检索主机1565或者SPS主机1570A/1570B)。在一些实施中，分区映射可被视作包括所示各种资源粒度(例如，节点、服务器和主机粒度)中的每个处的识别信息(例如，资源标识符)、用作一个或多个功能1506的输入的数据记录属性的指示以及功能1506本身。控制服务器可将分区映射的表示存储在元数据存储区中，并且在一些实施方案中，可以暴露各种API(例如，getPartitionInfo API)或其他编程接口，以将映射信息提供给数据生产者、数据消费者，或者提供到SMS子系统或SPS的节点。

[0118] 在一些实施方案中，数据记录到分区的映射以及从分区到资源的映射可以因各种因素而进一步复杂，例如：(a)在一些实施方案中，给定的节点、服务器或主机可被指定负责多个分区，或者(b)故障或其他触发条件可导致新的节点、服务器或主机被分配到给定的分区或分区的集合。此外，如上文所述以及下文描述，用于给定流的分区映射可以随着时间的推移而动态更改，而同时流记录继续由SMS和/或SPS节点处理。因此，在一些实施方案中，针对给定的流可至少临时保留映射元数据的若干版本，每个版本都对应于不同的时间段。

[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所示的时间线的时间T1处，创建或初始化流S1。分区映射PM1是针对流S1创建的，并且在T1到T2的时间间隔内仍有效。通过实例示出在T1与T2之间由SMS接收的三个数据记录。提交的数据记录110A(DR110A)具有客户端供应的分区键值“Alice”，提交的DR110B具有客户端供应的分区键值“Bill”，并且提交的DR110C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都映射到具有分区标识符“P1”的相同分区。对于P1数据记录而言，单个节点I1被配置成处理摄取，单个节点S1被配置成处理存储，单个节点R1被配置成处理检索，以及单个工作者节点W1被配置成处置SPS处理。映射PM1的有效范围的起始时间戳被设置成T1。

[0121] 在时间T2处，流S1在图16的示例性时间线中动态重新分区。在所述实施方案中，无论重新分区何时发生，数据记录都持续到达并且被SMS和SPS处理；SMS或SPS都无需离线。启用重新分区的原因可能是很多因素中的任一个，例如，响应于在摄取、存储、检索或处理节点处检测到超负载条件、响应于检测到在各个子系统的不同主机处的工作负载水平之间出现偏斜或不平衡，或者响应于来自数据消费者或数据生产者客户端的请求。在所述实施方案中，新的映射PM2在时间T2处(或T2后不久)生效，如PM2的有效范围起始时间戳设置所示。在至少一些实施中，与重新分区之前所用的相比，不同数据记录属性的集合可用于对数据记录进行分区。在一些情况下，额外的分区属性可以由数据生产者提交(例如，在SMS的请求下)，而在其他情况下，额外的属性可以由SMS摄取节点生成。此类额外属性可被称为“加盐(salted)”属性，并且将额外属性用于重新分区的技术可被称为“加盐”。在一个示例性实施中，超负载的摄取服务器可以向数据生产者(例如，向数据生产者执行的SMS客户端库代码)表明：对于重新分区而言，除了之前使用的分区键之外，还提供随机选择的较小整数值。初始分区键与加盐额外整数的组合随后可用来将摄取工作负载分布在摄取节点的不同集合之中。在一些实施方案中，检索节点和/或数据消费者可以被通知有关用于重新分区的额外属性。在至少一些实施中，此类额外属性可以不用于重新分区。

[0122] 在图16所示的实施方案中，相对于为T2之前的相同键选择的分区而言，新的分区映射导致为T2之后接收到的数据记录中的至少一些选择的不同分区。在T2之后提交的DR110P具有分区键值“Alice”，在T2之后提交的DR110Q具有分区键值“Bill”，以及在T2之后提交的DR110R具有分区键值“Charlie”。在所示的示例性情形中，使用PM2映射，DR110P被指定为分区“P4”的成员，DR110Q被指定为分区“P5”的成员，而DR110R被指定为分区“P6”的成员。在所述实施方案中，所示在T2之后接收的示例性数据记录都没有被指定为先前使用的分区“P1”的成员，相反，在重新分区之后可以使用全新的分区。在一些实施方案中，重新分区之后可以继续使用至少一些先前使用的分区。对于新分区P4、P5和P6中的每个而言，不同

的节点可被指定用于摄取、存储、检索和/或处理。例如，节点I4、S4、R4和W4可被配置用于分区P4，节点I5、S5、R5和P5可被配置用于分区P5，并且节点I6、S6、R6和P6可被配置用于分区P6。在一些实施方案中，重新分区之后，相同的存储节点可用于带有特定分区键或属性的数据记录，就像重新分区之前用于此类记录一样，但是重新分区之后可以使用所述节点内的不同存储位置(例如，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间段期间，检索请求可以继续以检索针对在重新分区之前由SMS摄取和/或存储子系统处理的数据记录。在至少一些情况下，可能需要基于在数据记录被摄取时实行的PM1映射检索所请求的数据记录。因此，如图16所示，出于数据检索的目的，在T2之后的一段时间可以继续使用PM1和PM2两者。在至少一些实施中，随着数据记录的老化，它们最终可从所述流中删除，并且最终也可以舍弃老的分区映射，例如，当所有的对应数据记录本身都已经被删除时。在一些实施方案中，替代于被删除(或在此之前)，可以将流记录存档(例如，基于客户端选择的存档策略)到存储位置或装置的不同集合，从而使得SMS使用的分区映射仍可用来检索存档之后的记录。在此类实施方案中，只要支持指向存档存储的检索请求需要诸如PM1和PM2的分区映射，便可保留所述分区映射。在一些存档实施中，可以使用不需要保留流分区映射的不同检索方法(例如，可以针对存档的数据记录创建新索引)。在一些实施方案中，重新分区之前使用但重新分区之后不再涉及写入的分区(例如，P2)可在重新分区之后的某个点处被“关闭”用于读取，例如，响应于检索请求，可以提供“到达分区末尾”错误消息的等效物。

[0124] 在一些实施中，给定的数据流可以被分成许多(例如，数百或数千)分区。设想一种示例性情况，其中流S1最初被分成1000个分区，P1、P2、……、P1000。在检测到对应于一个分区(比如P7)的超负载条件的情况下，可能值得改变数据记录到P7的初始映射，但其他分区的映射无需改变。在一种方法中，可经由重新分区操作来创建两个新的分区P1001和P1002。重新分区之后接收的记录可以映射到重新分区之后的P1001或P1002，其中所述记录的属性最初(即，基于原始映射)导致它们在P7中的成员资格，因而，将P7的工作负载分布在两个分区之中。剩余的分区，例如P1到P6和P8到P1000，可无需更改。由于只是分区的较小子集受到此类重新分区的影响，因此，在至少一些实施方案中，可以生成和存储组合数据结构，例如，分区条目的有向非循环图(或者分区条目树)。每个条目可指示分区功能输出范围以及有效时间范围(条目的分区信息被视作有效的时间段)。假设在上述实例中，在时间T2处执行涉及P7的重新分区，而在时间T1处创建流S1(及其初始映射)。在此类情形下，有关P7的条目的有效时间段将是“T1到T2”，P1001和P1002的有效时间段将是“T2向前”，而用于剩余分区的有效时间段将是“T1向前”。在至少一些实施中，使用这种组合数据结构可导致用于分区映射元数据的存储器或存储装置的量大幅度减少。在上述实例中，论述了将分区P7分成两个新的分区。在至少一些实施中，分区也可以在重新分区期间合并，例如，其中接收到相对较少检索请求或者提交相对较少记录的两个邻近分区可以合并成单个分区。针对任何给定的时间点，可以使用分区功能和有效时间范围信息来明确地确定数据记录所属的分区。随着时间的推移，由于执行越来越多的分裂和/或合并，组合数据结构可逐渐发展，但分区元数据所需的总空间(当然取决于分裂发生的频率，以及平均有多少分区受到分裂的影响)可以不显著增加。相反，在不同的实施中，每次重新分区发生时，流的未变元数据的整个集合都可被复制并且与受到重新分区影响的分区的条目进行组合。在后一个实施中，分区映射元

数据的存储和存储器需求的增加速率可能更高，尤其是如果老的映射可能需要保留到重新分区之后的至少一段时间，如上所述。

[0125] 在使用包括时间戳值(例如，图13b所示的时间戳值1304)的序列号的至少一些实施方案中，可以针对动态重新分区来实施特殊类型的序列号过渡。例如，假设将基于时间戳的序列号方案(类似于图13b所示)用于流S1，其中每秒都生成新的时间戳值以包含在所述序列号中。在支持动态重新分区的至少一些实施中，动态重新分区之后分配的序列号可以全都使用与动态重新分区之前所用的相比不同时间戳值的集合(从对应于重新分区事件的所选初始时间戳值开始)。例如，如果在动态重新分区实行(即，生效)时使用的时间戳值是Tk，那么实行之后发出的任何新的序列号都可被要求使用Tk+1向前的时间戳值。由于在图13b所用的方案中，序列号值用其高阶位中的至少一些对时间戳值进行编码，因此，确保重新分区事件对应于所述时间戳边界又可以简化在响应于检索请求而识别将使用的映射中所涉及的簿记(bookkeeping)。因此，在此类实施中，当接收到指定特定序列号的检索请求时，可以从所述序列号中提取时间戳值，并且可能容易确定应使用重新分区之后的映射，还是应使用重新分区之前的映射。如果提取的时间戳值低于为重新分区选择的初始时间戳，那么可以使用重新分区之前的映射，并且如果提取的时间戳值等于或大于为重新分区选择的初始时间戳值，那么可以使用重新分区之后的映射。

[0126] 用于流管理和处理的方法

[0127] 图17是示出根据至少一些实施方案的可以被执行以支持流记录摄取和流记录检索的编程接口的相应集合的操作方面的流程图。如元素1701所示，可以接收到创建或初始化数据流的请求，例如，来自SMS客户端或者数据生产者客户端的请求。可以确定将用于所述流的初始分区映射(元素1704)，例如，基于分区策略，可以确定将用来识别特定数据记录所属的分区的函数，以及将用于所述函数的输入参数。如之前所述，在各实施方案中，SMS的控制部件可以负责接收和响应流创建请求。从一个实施方案到另一个实施方案，实施流创建和初始化(以及其他控制平面操作)的方式可以不同。例如，在一个实施方案中，可以建立控制服务器的冗余组，并且通过生成用于新流的适当元数据(例如，初始分区映射，摄取、存储和检索的初始节点集合，等等)并将其存储在永久存储位置，所述冗余组的主控制服务器可以响应于流创建请求。对有关所述流的随后查询(例如，来自前端摄取节点的有关负责给定分区的后端节点的请求)作出的响应可以由主控制服务器使用存储的元数据生成。在SMS控制平面功能的另一实施中，流配置元数据可以存储在可由摄取、存储或检索子系统的至少一些节点直接访问的数据库中。在流被创建和初始化之后，诸如记录提交、存储和检索等数据平面操作便可以开始，并且可以由对应子系统的相应部件处理，一般无需与控制部件的额外交互。

[0128] 在一些实施方案中，数据生产者可被要求与写入请求一起提交明确的分区键，而在其他实施方案中，基于与所述写入请求相关联的元数据，例如，数据生产者的标识、从中接收到数据记录的IP地址或者数据记录的内容本身，可以确定用于分区功能的输入。在至少一个实施中，客户端可以任选地在数据记录提交中供应分区标识符，并且在此类实施中可以不要求额外的分区功能。

[0129] 在确定或配置用于流的摄取、存储和检索功能的初始节点集合时，可以考虑很多不同的因素(元素1707)。例如，下列因素都可影响不同子系统的节点的数量和放置：分区映

射本身(其可确定所述流被分成多少分区,以及分区的相对预期大小);有关预期摄取率和/或检索率的信息(如果此信息可用的话);流数据记录的耐久性/持久性需求;和/或各个子系统的高可用性需求(其可导致设立类似于图9和图10所示的那些冗余组)。此外,在客户端可以指示各类别的节点的放置目的地类型偏好(如图11、图12a和图12b所示)的实施方案中,此类偏好还可起到确定将用于SMS和/或SPS节点的资源的作用。在至少一些实施方案中,能够执行摄取、存储和/或检索功能的相应节点池可以提前设置,并且控制部件可以将此类池的所选成员分配到创建的每个新流。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,新的摄取、存储或检索节点可能需要被实例化。

[0130] 在所述实施方案中的摄取节点处,可以经由针对数据记录提交实施的编程接口集合中的任一个接收到记录(元素1710),例如,包括串联提交接口(其中数据被包括在提交请求中)以及按引用提交接口(其中地址在提交请求中提供,而通过SMS摄取节点或SMS存储节点可以从所述地址检索数据,例如,使用web服务请求或其他接口)。在不同的实施方案中,针对提交记录的每种方式,可以提供许多不同类型的编程接口中的任一个,例如,针对串联与按引用提交,可以支持相应的应用编程接口(API);可以建立web页或web站;可以实施图形用户接口;或者可以开发命令行工具。在至少一些实施方案中,SMS可以将序列号分配给每个摄取的记录,例如,表明摄取或存储所述记录的顺序,并且所述序列号可用于数据消费者的检索请求。在检索子系统节点处,可以经由实施的编程检索接口集合的任一来接收记录检索请求,并且作为响应,可以提供所请求的数据记录的内容(元素1713)。对于非顺序访问而言,例如,所述接口可包括getIterator(基于getIterator调用中指示的序列号,请求在分区内的选定位置处将迭代器实例化)或者getRecordWithSequenceNumber(以获取具有指定序列号的数据记录)。对于顺序访问而言,可以实施接口,例如,getNextRecords(从迭代器的当前位置开始或者从指定序列号开始,按顺序请求一些记录)。在至少一些实施方案中,不同的检索接口可以具有与之相关联的不同计费率,例如,顺序检索的每个记录计费率可以设置成低于非顺序检索的每个记录计费率。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,每个记录的按引用提交可以比串联提交成本更高。

[0131] 随着时间的推移,针对在流管理服务的各个子系统处实施的不同编程接口,控制节点或专用计费服务器可以收集使用度量(元素1716)。例如,所述度量可以包括:不同编程接口的调用计数、摄取或检索的记录的总数(对于至少一些接口而言,所述总数可以不同于调用计数(例如,可以用单个调用来检索多个记录的getNextRecords))、摄取或检索的数据的总量等等。至少部分基于使用度量以及与编程接口相关联的相应计费率,可以任选地生成将向拥有所述流的客户端或者生产和/或消费所述流中的数据的客户端收取的计费量(元素1719)。在至少一些实施方案中,计费活动可以与流摄取/检索操作异步,例如,可在月度计费期结束时,基于在这个月内收集的度量来生成账单。

[0132] 图18a是示出根据至少一些实施方案的可以被执行以配置流处理(SPS)阶段的操作方面的流程图。如元素1801所示,可以实施编程接口,从而使得客户端能够针对流数据记录来配置许多处理阶段。为了配置特定的阶段,例如,客户端可以指明:在所述阶段将对分区流数据记录执行的处理操作、处理操作的输出的分布策略以及其他参数,例如,从中获取将要处理的数据的输入流的标识。在一些实施方案中,可以要求SPS阶段的处理操作是幂等的。在其他实施方案中,针对至少一些阶段,也可以支持非幂等操作。如果在给定阶段执行

的处理是非幂等的，那么在一些实施方案中，通过将工作者节点配置成定期将操作的输出清除(flush)到某一永久性外部位置、记录相对于记录检索序列来执行清除操作的时间以及之后将置换工作者节点配置成在恢复期间重复清除操作，客户端仍能够获取幂等性的恢复相关益处。在至少一些实施方案中，客户端能够配置有向非循环图(DAG)或者处理阶段的其他图，其中若干不同状态在流数据上并行操作并且一些阶段的结果用作其他阶段的输入流。在一些实施方案中，在不同阶段之间可以创建一个或多个短暂流，而非持久流，例如，在作为输入馈送到不同阶段之前，从一个阶段输出的数据记录无需存储在永久性存储装置上。

[0133] 在一些实施方案中，针对SPS阶段可以实施许多不同恢复策略中的任一个，例如，包括基于检查点的恢复策略或者最大努力恢复策略。在一个实施方案中，客户端可以使用编程接口来为不同的SPS阶段选择恢复策略。在使用基于检查点的恢复的阶段，工作者节点可以被配置成不时地存储进程记录或检查点，表明它们在流分区中实现的进度(例如，最近处理的记录的序列号可以存储作为进程的指示符)。之后在故障后的恢复操作期间，可以使用进程记录，如下文参考图19所述。在最大努力恢复策略中，无需存储进程记录，并且响应于故障配置的置换工作者节点可以在接收到新的数据记录时就进行处理。在给定的SPS阶段图形或工作流程内，在一些实施方案中，可以将不同的恢复策略应用于不同的阶段。

[0134] SPS控制服务器可以接收(例如，经由元素1801所示的编程接口中的一个来接收)根据分区策略PPol1将在特定阶段PS1对流S1执行的幂等操作Op1的指示，其中处理结果将根据输出分布描述符DDesc1进行分布(元素1804)。例如，基于各种因素，例如，Ppol1、幂等操作Op1的复杂性以及用于工作者节点的资源的性能能力，可以确定被配置用于状态PS1的工作者节点的数量以及节点所需的虚拟或物理资源(元素1807)。

[0135] 随后可以对工作者节点进行实例化和配置(元素1810)，例如，作为在所选的虚拟机或物理机资源处的进程或线程。在一个简单的实施中，例如，一个工作者节点最初可以被分配用于S1的每个分区。给定的工作者节点可以被配置成：(a)接收来自S1的检索节点的适当子集的数据记录，(b)对接收的数据记录执行Op1，(c)任选地，例如，基于PS1的恢复策略，存储表明已经处理了分区记录的哪个集合的进程记录/检查点，以及(d)将输出传输到DDesc1指示的目的地(例如，作为中间持久流或短暂流的输入，或者直接传输到其他处理阶段或存储系统)。应注意，至少在一些实施方案中，SPS处理可以不必生成基于前进的基础而在其他地方传输的任何输出。例如，一些SPS应用可以只用作数据记录的临时存储库，和/或可以实施查询接口，从而使得用户能够查看数据记录。此类应用可以管理自己的输出，例如，响应于接收的查询而不用根据分布描述符，可以生成输出。日志相关SPS应用可以保留从大规模分布式系统收集的最后一天的日志记录，例如，从而使得客户端能够查看日志数据，以用于除错或分析的目的。因此，在一些实施方案中，针对SPS的至少一些阶段、针对至少一些流或者针对至少一些分区，无需指定输出分布描述符。随后，工作者节点可以根据它们的相应配置设置开始检索和处理数据记录(元素1813)。在至少一些实施方案中，SPS控制节点可以监控工作者节点的健康状况(例如，使用诸如心跳协议等响应检查)以及各种其他度量，例如，用于工作者节点的资源处的资源利用水平(元素1816)。从工作者节点收集的信息可以用来确定是否需要故障转移(例如，工作者节点是否应被置换)，以及实施的恢复策略，如下文所述。

[0136] 在一些实施方案中，可安装的SPS客户端库可被提供到希望在客户端拥有的驻地处和/或在提供者网络的客户端选择的资源处实施SPS工作者节点的那些客户端。客户端库还可以允许SPS客户端选择它们希望使用SPS管理服务的各种控制平面特征的程度，例如，健康监控功能、自动化工作负载监控和平衡、安全性管理、动态重新分区等等。图18b是示出根据至少一些实施方案的可以响应于调用客户端库的部件以配置流处理工作者节点而执行的操作方面的流程图。如元素1851所示，可以提供SPS客户端库(例如，从被配置成执行图18a所示的操作种类的多租户SPS管理服务的web站下载)。所述库可包括许多可执行部件和/或可以链接到客户端应用的部件。一些库部件可以使得客户端能够选择、通过SPS管理服务注册各个工作者节点或者指定各个工作者节点的所需性能，其中将在所述工作者节点处执行一个或多个SPS阶段的流处理操作。例如，一个客户端可能希望将在提供者网络的虚拟计算服务处实施的自己的计算实例集合用于工作者节点，而另一客户端可能希望将位于客户端自己的数据中心处的计算装置(例如，提供者网络不支持的专用装置)用于处理流记录。根据需要，客户端可以将工作者节点在线带到它们自己的驻地，或者如有需要，使用虚拟计算服务的计算实例。除了工作者节点的此类按需实例化之外或作为替代，在一些实施方案中，客户端可以预先配置可在需要时部署的可能重复使用的工作者节点池。在一些实施中，库部件可以被执行或调用，以允许客户端通过SPS管理服务将由客户端实例化的特定进程或线程注册为指定阶段的工作者节点，对于所述工作者节点而言，随后的控制平面操作可以由SPS管理服务处理。在一个实施方案中，针对工作者节点，客户端还能够从不同水平的控制平面责任中选择将由SPS管理服务处理的那些，例如，一个客户端可能希望使用它们自己的自定义模块来监控工作者节点健康状态，而另一客户端可能希望利用SPS管理服务来监控工作者节点健康状态，并且在检测到故障时采取适当的动作。

[0137] SPS管理服务可以接收特定客户端希望将客户端库用于配置工作者节点和/或特定SPS阶段PS1的控制平面操作的指示(元素1854)。(PS1本身可以使用所述库中包括的编程接口进行设计，或者使用SPS管理服务暴露的编程接口(类似于图4所示的基于web的接口)进行设计。)客户端还可以指示其数据被检索以由PS1用作输入的流。任选地，在至少一些实施方案中，客户端可以指示用于PS1的控制平面设置，例如，客户端想要将服务的健康监控能力用于节点，还是愿意使用自定义健康监控工具(元素1857)。根据客户端指示的偏好，可以确定被配置成由客户端使用的SMS和/或SPS的一个或多个节点(元素1860)。在客户端的工作者节点到SMS/SPS节点之间，可以建立网络连接，和/或可以执行其他配置操作，以使得数据记录和处理结果能够根据需要而流动。在接收到检索请求之后，数据记录可被提供到SP1工作者节点，并且可以根据需要执行所需的控制平面操作(如果客户端请求的话)。应注意，至少在一些实施方案中，也可以或改为实施类似方法，使得客户端能够控制它们希望使用SMS管理服务的各个子系统的控制平面功能的程度。

[0138] 图19是示出根据至少一些实施方案的可以被执行以实施流处理的一个或多个恢复策略的操作方面的流程图。如元素1901所示，SPS控制节点可确定已满足用于置换特定工作者节点的触发准则，例如，工作者节点已经变得无响应或者不健康、当前节点的工作负载水平可能已经达到故障转移的阈值、在工作者节点处检测到的错误数量可能已经超过阈值或者可能识别出工作者节点的某一其他意外状态。置换工作者节点可以被识别或实例化(元素1904)。在一些实施方案中，可以设置可用工作者线程的池，例如，其中的一个可以被

选择作为置换，或者可以发起新的线程或进程。

[0139] 如果最大努力恢复策略被用在特定工作者节点活跃的SPS阶段(如在元素1907中确定)，那么置换工作者节点可以在额外的数据记录变得可用时开始处理额外的数据记录(元素1916)，例如，无需检查被置换的工作者节点的进程记录。如果使用基于检查点的恢复策略，那么可以提供位置的指示(例如，存储装置地址或者URL)，在所述位置处，置换工作者节点可以访问被置换的工作者节点所存储的进程记录(元素1910)。置换工作者节点可以检索被置换的节点所存储的最近进程记录，并且使用所述进程记录来确定置换工作者节点应执行阶段的幂等操作所针对的数据记录集合(元素1913)。在此类基于检查点的恢复策略中，根据最后的进程记录与置换工作者节点被实例化时的时间之间的持续时间，以及根据被置换的工作者节点处理所存储的进程记录之后的额外记录的速率，可能不止一次地处理一些数量的数据记录。在至少一些实施方案中，如果执行的操作是幂等的，那么此类重复操作可能不具有负面影响。在基于先前存储的进程记录，置换工作者节点已执行重复恢复操作之后，在至少一些实施方案中，置换工作者线程可以存储表明恢复完成的其自己的进程记录，并且可以开始对新接收的数据记录进行正常的工作者线程操作(元素1916)。

[0140] 图20是示出根据至少一些实施方案的可以被执行以实施数据流的多个安全选项的操作方面的流程图。如元素2001所示，可以实施一个或多个编程接口，使得客户端能够针对数据流管理和处理而从各种安全选项中进行选择，例如，包括用于不同功能类别的节点(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。在安全配置文件的各个方面，放置目的地类型可以彼此不同。在一些实施方案中，从一个目的地类型到另一目的地类型，用于SMS或SPS节点的资源的物理位置可以不同。例如，可以将诸如位于提供者网络数据中心处的实例主机等资源用于节点，或可以使用客户端拥有的设施处的资源，或者可以使用第三方资源。在至少一些实施方案中，从一个目的地类型到另一目的地类型，网络隔离等级或其他联网特性可以彼此不同，例如，一些SMS或SPS节点可以在隔离虚拟网络内被实例化，或者在经由专用隔离物理链路连接到提供者网络的客户端拥有的设施处进行实例化。在一个实施方案中，客户端可以指示将某些类型的SMS或SPS节点建立在提供者网络的单租户实例主机处，而不是使用也可用的多租户实例主机。在至少一些实施方案中，也可以经由安全相关编程接口来选择各种类型的加密选项。

[0141] 经由安全相关编程接口，可以接收有关用于流S1的一个或多个功能类别的节点的客户端安全配置文件选择或偏好。例如，客户端可以为功能类别FC1的节点选择一个安全配置文件(例如，客户端可能希望在客户端拥有的驻地处实施SPS工作者节点)，并且为不同功能类别FC2的节点选择不同的安全配置文件(例如，客户端可能愿意在提供者网络数据中心处实施SMS摄取节点或存储节点)(元素2004)。在一些情况下，客户端可以决定将所有不同功能类别的节点都设置成具有相同的安全配置文件。在一些实施方案中，SMS和/或SPS可以针对各种功能类别来定义默认放置目的地类型，例如，除非客户端另外指示，否则所有功能类别的节点都可以设置在提供者网络的隔离虚拟网络内。

[0142] 随后，基于客户端对安全配置文件和/或位置的偏好(或者基于客户端并未提供偏好所针对的功能类别的默认设置)，可以配置不同功能类别的节点(元素2007)。例如，所述配置可以涉及：选择适当的物理主机或物理机；以及针对不同功能类别的节点，将适当的计算实例、虚拟机、进程和/或线程实例化；并且在节点之间建立适当的网络连接。在一些实施

方案中，作为配置的一部分，可以提供用于不同流管理和处理功能的可执行库部件，以便安装在提供者网络外部的主机处。

[0143] 根据至少一些实施方案，例如，根据客户端表达的加密偏好或者基于默认加密设置，可以在一个或多个类别的节点处激活加密模块(元素2010)。随后可以激活各种功能类别的节点，从而根据客户端需要，流数据被摄取、存储、检索和/或处理(元素2013)。

[0144] 图21是示出根据至少一些实施方案的可以被执行以实施数据流的分区策略的操作方面的流程图。如元素2101所示，可以为数据流确定分区策略。例如，所述策略可包括基于数据生产者供应的键或者基于所提交的数据记录的各种属性的从数据记录到分区的初始映射，以及用于对数据流进行重新分区的一个或多个触发准则。在一些实施方案中，例如，散列函数可以应用于一个或多个分区键，从而产生128位整数散列值。128位整数的可能范围可以被分成N个连续的子范围，每个子范围都表示流的N个分区中的一个。在一些实施方案中，从一个流到另一流，分区的数量和/或子范围的相对大小可以不同。在至少一些实施方案中，客户端(代表客户端，流被配置)可以提供有关将要使用的分区方案的输入，例如，所需的分区数量，或者将要使用的分区功能的所需特性。在至少一个实施方案中，客户端可以提供用于所提交的数据记录的一些子集或全部的分区标识符或名称。

[0145] 当接收到流的数据记录时，可以基于供应的键和/或其他属性来确定相应的分区，并且针对识别的分区，可以选择摄取、存储和检索节点的适当集合(元素2104)。在至少一些实施方案中，针对数据记录，可以生成相应的序列号，例如，表示接收到给定分区的记录的序列(元素2107)。在一些实施中，序列号可以包括许多元素，例如，时间戳值(例如，自诸如1970年1月1日UTC 00:00:00等公知出现时间起经过的秒数)、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，可以将序列号提供到数据生产者，例如，以便确认所提交的数据记录的成功摄取。在一些实施方案中，序列号也可以由数据消费者用来按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，数据记录可以按序列号顺序存储在基于分区策略而数据记录被指向的存储节点处(元素2110)。在使用旋转磁盘存储装置的实施方案中，顺序写入一般可以用来将接收的数据记录保存到磁盘，从而避免磁盘寻道等待时间。在至少一些实施中，在将记录存储到磁盘之前，非易失性缓冲器可以用作写入缓存，例如，以进一步降低磁盘寻道的概率。响应于针对按序列号顺序读取多个数据记录的请求(例如，调用getNextRecords或类似接口)，之后可以使用顺序读取从存储装置中读取数据记录(元素2113)。

[0147] 图22是示出根据至少一些实施方案的可以被执行以实施数据流的动态重新分区的操作方面的流程图。如元素2201所示，可以确定(例如，在SMS或SPS的控制部件处)流将被动态重新分区。很多不同的触发条件可以导致决定对流进行重新分区，例如，检测到摄取、存储、检索、处理或控制节点中的一个或多个处的超负载，或检测到不同节点的工作负载水平不平衡，或者可能从客户端(例如，数据生产者或数据消费者)接收到的重新分区请求。在一些实施中，客户端重新分区请求可以包括所请求的重新分区的具体细节，例如，将要生成的更改映射的各种参数(例如，添加或移除的分区的数量、哪些具体分区应被组合或分裂等等)。在一个实施中，客户端重新分区请求可以表明客户端希望解决的问题状态(例如，负载不平衡)，并且SMS或SPS可以负责将问题状态的描述转译成适当的重新分区操作。在一些情

况下，替代于请求重新分区或者描述问题状态，客户端可以指定将用于重新分区的触发准则。在一些实施方案中，确定改变数据流的数据耐久性需求可以触发重新分区，从而可导致(例如)针对流记录来选择存储装置的不同集合或者不同的存储技术。在一些情况下，检测到数据流的使用模式(例如，生产或消费数据记录的速率)的改变也可能导致重新分区，并且还可能导致使用更适合改变后的使用模式的不同存储技术或者存储装置的不同集合。例如，基于确定，对于给定分区或整个流预期的读取和写入速率而言，SSD可能是比旋转磁盘更合适的存储技术，可以做出重新分区的决定。在一个实施方案中，安排好的或者即将发生的软件和/或硬件版本的变化可以触发重新分区。在一些情况下，如当客户端指示使用不同的分区方法或不同的存储方法可以更有效地满足的预算约束时，价格或计费问题可以触发重新分区。在至少一些实施方案中，改变的性能目标也可以触发重新分区。在图22所示的实施方案中，可以选择将用于重新分区之后分配的序列号的初始时间戳值(例如，从1970年1月1日UTC 00:00:00(在若干操作系统中可经由系统调用而得到的出现时间值)算起的秒数的偏移量)(元素2204)。在一些实施中，在提供者网络处实施的全球状态管理器可以支持getEpochValue API，例如，从而使得SMS和/或SPS的各种部件能够获得一致的时间戳值，以便用于序列号生成。在其他实施中，可以使用其他时间源，例如，可以指定SMS或SPS控制节点将一致顺序的时间戳值提供给其他部件，或者可以使用本地系统呼叫调用。在一些实施方案中，时间戳值不必对应于任何特定主机处的挂钟时间，例如，可以就使用单调递增的整数计数值。

[0148] 针对所述流，可以生成更改的分区映射，所述映射与决定重新分区时使用的映射不同(元素2207)。在至少一些实施方案中，改变的映射可以将具有特定分区键的数据记录映射到不同分区，即，与具有相同键的数据记录在重新分区之前被映射到的分区不同。一些分区(一般是使用率高的分区)可以被分裂，而其他(一般是使用率低的)分区可以被合并，具体取决于重新分区的触发条件和/或观察到的工作负载度量。在一些实施方案中，重新分区之后与重新分区之前可以使用不同的分区功能，例如，不同的散列函数，或者可以使用将散列函数结果细分成分区的不同方法。例如，在分区对应于128位整数的连续范围的一些实施中，128位整数空间在重新分区之后可以被分成不同子范围的集合。在至少一些实施方案中，摄取、存储、检索、处理或控制节点的新集合可以被分配给新创建的分区。在一些实施中，可以使用省空间的组合数据结构来表示初始映射和更改的映射两者(元素2208)。例如，可以存储有向非循环图或树状结构，其中每个条目都含有分区功能输出范围的指示(例如，对应于给定分区的分区散列函数结果的范围)以及有效时间范围，从而作为重新分区的结果，只有对应于更改的分区的记录需要改变。在数据结构中，重新分区期间保持不变的分区的条目无需更改。新节点可以被配置成实施更改的分区映射(元素2210)。在至少一些实施方案中，由于在至少一段时间内可以继续接收到针对基于先前的映射存储的数据记录的检索请求，因此，先前的节点和先前的映射可以保留一段时间。当接收到指定特定序列号或时间戳的读取请求(元素2213)时，可以确定(例如，在控制节点处或在检索节点处)要使用新的分区映射还是先前的分区映射来满足所述读取请求。随后，选择的映射可以用来识别从中获取请求的数据的适当存储节点。

[0149] 图23是示出根据至少一些实施方案的可以被执行以实施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。如元素2301所示，可以实施一个或多个编程接口，以使

得客户端能够针对数据流从若干摄取策略选项中选择记录摄取策略，例如，所述策略包括(a)至少一次策略，根据所述策略，记录提交者一次多次提交记录直到接收到肯定确认为止；或者(b)最大努力摄取策略，根据所述策略，不用为至少一些记录提交提供确认。一些数据生产客户端可不像其他客户端那样担心可能会丢失它们记录的一小部分，并且因此，可以选择最大努力摄取方法。在一些实施中，甚至对于被配置用于最大努力摄取的流而言，SMS仍可为数据记录的一些子集提供确认，或者甚至可能尝试为所有的数据记录提供确认，即使最大努力策略并不要求对每个数据记录进行确认。

[0150] 经由编程接口中的一个可以接收请求，从而表明将用于指定流的特定摄取策略(元素2304)。根据为所述流实行的分区策略，可以将摄取节点实例化(元素2307)。当在摄取节点处接收到相同数据记录的一次或多次提交(元素2310)时，根据实行的摄取策略可以采取不同的动作。如果使用至少一次摄取策略(如在元素2313中确定)，那么针对一次或多次提交中的每个，可以将确认发送到数据生产者，但数据记录在存储子系统处可以只保存一次(2316)。(应注意，根据为所述流实行的持久性策略，在一些实施方案中，可以存储给定记录的N个复本，但如果给定的数据记录被提交M次，那么可以只针对提交中的一个生成复本，即，存储的记录复本的总数将是N，而不是NxM。)如果实行最大努力摄取策略(如也在元素2313中检测到)，那么数据记录仍是只可以在存储装置处保存一次，但无需向数据生产者发送确认(元素2319)。在至少一些实施方案中，至少部分基于所选择的摄取策略，可以任选地确定客户端计费量(元素2322)。如之前所述，在一些实施方案中，可以支持至少一次摄取策略的两个版本。在一个版本中，类似于图23所示，SMS可以负责去除重复数据记录(即，确保，响应于两次或更多提交的集合中的仅一个，数据存储在SMS存储子系统处)。在至少一次摄取的不同版本中，SMS可以准许数据记录重复。后一种方法可用于数据记录重复很少有或没有造成消极后果的流应用，和/或用于本身执行重复消除的流应用。

[0151] 图24是示出根据至少一些实施方案的可以被执行以实施数据流的多个持久性策略的操作方面的流程图。如元素2401所示，可以实施一个或多个编程接口，使得客户端能够针对流数据记录从多个持久性策略中选择一个持久性策略。持久性策略可以在各种方面的任一方面彼此不同：例如，(a)保存的复本的数量可以不同(例如，可以支持N个复本对2个复本对单个复本策略)、(b)将使用的存储位置/装置类型可以不同(例如，旋转磁盘对SSD对RAM对数据库服务或者多租户存储服务)和/或(c)所述策略在从大规模故障复原的预期程度方面可以不同(例如，可以支持多数据中心对单数据中心策略)。可以接收请求，其表明客户端针对指定流的特定持久性策略的选择(元素2404)。在一些实施方案中，客户端选择的持久性策略可以导致针对给定流的相应分区来使用不同的存储位置类型或装置类型。在一个实施方案中，SMS而不是客户端可以在流水平或在分区水平选择存储位置类型或装置类型。在一些实施方案中，当在一些实施方案中选择持久性策略时，客户端可以指明数据耐久性目标和/或性能目标(例如，所需的读取或写入吞吐量或等待时间)，并且这些目标可以由SMS用来选择合适的存储装置类型或位置。例如，如果需要低的等待时间，那么可以使用SSD来存储一个或多个分区或流的数据记录，而不是旋转磁盘。

[0152] 摄取节点的集合可以被确定或配置成从数据生产者接收所选择的流的数据记录，并且存储节点的集合可以被配置成实施所选择的持久性策略(元素2407)。当在摄取节点处接收到数据记录(元素2410)时，基于所选择的持久性策略，可以通过负责数据记录所属的

分区的存储节点而将数据记录的一个或多个复本存储在所选择的存储装置处(元素2413)。在至少一些实施中,基于客户端选择的具体持久性策略,可以任选地(和/或异步)确定计费量(元素2416)。

[0153] 用于流处理的分散工作负载管理

[0154] 在一些实施方案中,SPS的控制平面功能的主要部分或全部都可以分散方式实施,例如,通过给定SPS阶段内的工作者节点经由共享数据结构(例如,数据库表)来协调各个控制操作(例如,对工作者节点的分区分配、对动态重新分区的响应、健康监控和/或负载平衡)。给定的工作者节点W1可以检查共享数据结构内的条目,例如,以确定所述阶段的输入流(如果有的话)的哪些分区当前未被处理。如果找到这样的分区P1,则W1可以更新共享数据结构中的条目,以表明W1将对P1的记录执行所述阶段的处理操作。其他工作者节点可以了解到W1被分配来处理P1记录,并且因此,可以将不同的分区分配给自己。工作者节点可以定期或偶尔将查询提交给SMS控制平面,以确定为输入流实行的当前分区映射,并且根据需要来更新共享数据结构,以表明映射变化(例如,作为重新分区的结果)。在各实施方案中,负载平衡和其他操作也可经由共享数据结构进行协调,如下文描述。在一些此类分散实施中,SPS可以不要求专用控制节点,从而减少实施SPS工作流程所需的开销。此类分散的SPS控制平面实施可尤其受关心预算的客户欢迎,所述客户利用SPS客户端库来实施流处理的各个方面,例如,在被分配给客户的提供者网络内的计算实例处实施,或者在提供者网络外部的位置处实施。分散的SPS控制平面技术也可以用在不使用客户端库的实施方案中,例如,当用于SMS和SPS的所有资源都被配置在提供者网络内时。工作者节点针对至少一些处理阶段来实施SPS控制平面功能中的一些或全部所处的SPS在本文中可被称为“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者节点使用数据库表来协调它们的工作负载。在分散控制SPS 2590中,界定两个阶段215A和215B,每个都具有相应的工作者节点集合。阶段215A包括工作者节点2540A和2540B,而阶段415B包括工作者节点2540K和2540L。针对每个阶段215A和215B,在数据库服务2520处创建对应的分区分配(PA)表2550,例如,用于阶段215A的PA表2550A以及用于阶段215B的PA表2550B。例如,响应于客户端库部件或功能的调用,在一些实施方案中,可以在阶段初始化期间创建给定阶段的PA表2550。每个PA表2550可以被填充表示所述阶段的输入流的未分配分区(即,当前未被分配工作者节点的分区)的条目或行的初始集合。图26示出并且下文描述PA表条目的示例性列或属性。针对所述阶段发起的工作者节点2540(例如,在计算实例或其他服务器处发起的进程或线程)可以被授予所述阶段的PA表的读取/写入访问权。在图25中,从工作者节点指向PA表的读取和写入由分别用于工作者节点2540A、2540B、2540K和2540L的箭头2564A、2564B、2564K和2564L表示。

[0156] 给定的工作者节点2540可以被配置成,通过检查PA表中的条目,选择执行所述阶段的处理操作所针对的特定分区。在一个实施中,工作者节点2540A可以扫描PA表2550A中的条目,直到它发现未分配分区Pk的条目为止,并且可以尝试通过更新条目来将分区Pk分配给自己,例如,通过将工作者节点的标识符插入到条目的一列中。此类插入可被视作类似于工作者节点锁定所述分区。根据所用的数据库服务的类型,可以使用不同方法来管理可能同时对PA表条目的写入(例如,由正好几乎同时识别未分配的分区的两个或更多工作者

节点写入)。

[0157] 在一个实施方案中,可以使用提供者网络的非关系型多租户数据库服务,它支持强一致性和条件型写入操作,但不必支持关系型数据库事务语义。在此类情况下,条件型写入操作可以用于工作者节点进行的更新。设想一个实例,其中列“工作者节点ID(worker-node-ID)”用来指示分配给PA表中的分区的特定工作者节点的标识符,并且如果没有将工作者节点分配给所述分区,则所述列的值被设置为“空(null)”。在此类情形下,具有标识符WID1的工作者节点可以请求下列内容的逻辑等效物:“如果在分区Pk的条目中,工作者节点ID为空,那么将所述条目的工作者节点ID设置为WID1”。如果此类条件型写入请求成功,那么具有标识符WID1的工作者节点可以假设分区Pk被分配给它。工作者节点随后可以开始检索分区Pk的数据记录,例如,使用SMS检索子系统206的记录检索接口,如箭头2554所示(例如,分别用于工作者节点2540A、2540B、2540K和2540L的箭头2554A、2554B、2554K和2554L),并且对检索的记录执行处理操作。如果条件型写入失败,那么工作者节点可以重新开始搜索不同的未分配分区。在其他实施方案中,可以使用支持事务的数据库服务(例如,关系型数据库),并且事务功能可以用来实施条件型写入操作的等效物,例如,以确保在将分区分配给工作者节点的多次同时(或近乎同时)尝试中,只有一次成功,以及确保此类同时尝试中涉及的工作者节点被可靠地通知它们的成功或失败。在一些实施方案中,可以使用既不依赖于条件型写入也不依赖于事务支持的同步技术。在一些实施中,可以不使用数据库服务;相反,工作者节点可以使用锁定服务来获取独占访问,以更新类似于PA表的持久性数据结构中的条目。

[0158] 其他工作者节点2540可以检查PA表中的条目、确定哪些分区未被分配并且可最终成功将一个或多个分区分配给自己。通过这种方式,用于所述阶段的一个或多个输入流的分区的处理工作负载最终可以由所述阶段的工作者节点分布在它们自己之中。

[0159] 例如,由于之前所述的动态重新分区操作,任何给定的流的初始分区映射可以随着时间的推移而改变。因此,在图25所示的实施方案中,工作者节点2540中的一个或多个可以偶尔(或者响应于触发条件,如下文所述)将请求提交到它们阶段的输入流的SMS控制子系统210,以获取当前的分区元数据。在一些实施中,此类请求可以包括SMS控制平面API的调用,例如,箭头2544A、2544B、2544K和2544L所示的getStreamInfo API的调用。例如,SMS控制子系统可以用流的最新分区列表和/或其他细节作出响应,例如,分区的有效时间段。如果SMS控制子系统210提供的分区信息并不匹配PA表中的条目,那么PA表可以由工作者节点更改,例如,通过插入或删除一个或多个分区的条目进行更改。在至少一些实施方案中,对SMS控制子系统的此类请求2554一般没有记录检索请求2554(和/或数据库读取或写入操作2564)那么频繁,如标记“不频繁”的箭头2554A所示。例如,一旦被分配了分区,工作者节点一般可以不断地检索和处理所述分区的数据记录,直到完全消费分区数据为止(例如,如果流的拥有者关闭所述流的话,或者如果分区因动态重新分区而被关闭的话),或者直到遇到某一其他低概率状况为止(例如,如果由于检测到负载不平衡,不同的工作者节点请求转移所述分区的话,如下文论述)。因此,在各实施方案中,与调用getStreamInfo或类似API相关联的开销一般可非常小,即使响应于任何给定的调用而提供大量信息(如在为阶段的输入流界定数百或数千分区时可能出现的情况),也是如此。

[0160] 在图25所示的实施方案中,分散控制SPS环境的关键工作负载管理操作中的一些

因而可总结如下：(a)至少部分基于流处理阶段的第一工作者节点访问数据库表，选择所述流处理阶段的输入数据流的特定分区，以便实施针对所述阶段定义的处理操作的集合；(b)将所述特定分区分配到第一工作者节点的指示符写入到所述表中存储的特定条目中；(c)通过第一工作者节点，使用在多租户流管理服务处实施的编程记录检索接口来检索特定分区的记录；(d)通过第一工作者节点，对特定分区的记录实施处理操作的集合；(e)至少部分基于特定数据库表中的特定条目，由第二工作者节点确定分配第一工作者节点来对特定分区执行处理操作的集合；以及(f)由第二工作者节点选择加以执行处理操作的集合的不同分区。如果并且当工作者节点确定没有更多的记录留在被分配到所述节点的分区中时，工作者节点可以请求来自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的话，可以更新PA表。

[0161] 图26示出根据至少一些实施方案的可以存储在分区分配表2550中的示例性条目，所述分区分配表用于工作负载协调。如图所示，表2550可以包括四列：分区标识符列2614、分配的工作者节点标识符列2618、工作者节点健康指示符列2620以及工作负载水平指示符列2622。在其他实施中，可以实施其他列集合，例如，在一些实施方案中可以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负载水平指示符列。

[0162] 应注意，在一些实施方案中，至少在某些时间点，由SMS控制系统维护的分区列表2650(例如，作为分区条目树、图形或其他之前所述的组合数据结构的一部分)包括比PA表2550中包括的更多的分区。在所述实例中，分区列表2650包括分区P1、P2、P3、P4和P5，其中所示P1和P4因重新分区而处于关闭状态，而P2、P3和P5被示为活跃(即，分区的数据记录当前正被检索和处理)。在所述实施方案中，PA表2650包括用于活跃分区的条目，但不包括用于关闭的分区的条目(例如，当工作者节点在重新分区发生之后获取getStreamInfo调用的响应时，它们可能已经删除了所述关闭的分区)。至少在一些实施中，在给定时间点处，并不是所述流的所有当前开放的分区都要具有PA表中的相应条目；相反，例如，可以只呈现当前被分配或正被处理的这些分区的子集。

[0163] 在图26所示的示例性情形中，分区P1和P2被分配给分别带有标识符W7和W3的工作者节点，而P5当前未被分配。在不同的实施中，健康指示符列2620可以存储不同类型的值。在一些实施中，工作者节点可以负责定期(例如，每N秒一次，或者根据基于探试法的某一设置的计划)更新它们的分配分区的PA条目中的健康指示符列的内容，以表明工作者节点活跃并且能够继续检索和处理操作。在图26中，可以存储用于所述条目的工作者节点更新健康指示符列的最近时间的指示(“最后修改时间”)，例如，所示工作者W7在2013年12月1日的02:24:54和53秒修改了条目。在一些实施方案中，其他工作者节点可以使用最后修改时间值来确定分配的工作者节点是否健康，例如，如果经过X秒或分钟，如所述阶段的故障转移策略定义，那么分配的工作者节点可以被假定为不健康或无法访问，并且分区可以重新分配。在其他实施中，可以将计数器用作健康指示符(例如，如果计数器值在Y秒内没有改变，那么分配的工作者节点可以被视作故障转移的候选)，或者可以使用表明分配的工作者节点最后何时读取条目的“最后读取时间”值。

[0164] 在至少一些实施方案中，工作负载水平指示符值2622可以(例如)由分配的工作者节点存储在条目中，例如，某一最近时间间隔期间(例如，最后修改时间前五分钟内)处理的记录的数量、工作者节点的最近性能相关度量，诸如CPU利用率、存储器利用率、存储利用率

等。在一些实施方案中，此类工作负载水平指示符值可以由工作者节点用来确定是否存在负载不平衡，如下文参考图29所述，并且用来响应于检测到不平衡而采取行动。例如，工作者节点Wk可以确定它的工作负载水平高于平均工作负载水平，并且可以不分配分区中的一个，或者可以请求动态重新分区；或者，工作者节点Wk可以确定它的工作负载相对于其他工作者节点或分区来说太低，并且可以将额外的分区分配给自己。因此，使用图26中所示的PA表的列，在所述实施方案中，工作者节点可以执行一些相同类型的控制平面功能，而在集中控制SPS实施中，所述控制平面功能一般可由专用SPS控制节点执行。

[0165] 图27示出根据至少一些实施方案的可以由流处理阶段的工作者节点执行以便选择分区的操作方面，其中在所述分区上执行处理操作。如元素2701所示，针对分散控制SPS处理阶段SP1，可以在数据库服务处将PA表PAT1初始化。例如，所述表可以在SPS客户端库部件被调用时创建，例如，创建自客户端设施处的主机，或者创建自提供者网络数据中心处的计算实例。客户端库可以用于各种目的：例如，提供可执行部件，例如，用于将要在SPS阶段实施的特定处理操作的JAR(Java™档案)文件；指示可用来识别工作者节点的标签(例如，程序名、进程名或者计算实例名)；指示将用作所述阶段的输入的流；指示所述阶段的输出目的地(如果有的话)；等等。在一些实施方案中，PAT1最初可以被填充分区{P1,P2,…}的至少子集的条目或行，所述分区被定义用于所述阶段的输入流。在一些实施中，所述表最初可以是空的，并且工作者节点中的一个或多个可以用未分配分区的行来填充所述表，例如，作为获取来自SMS控制子系统的分区元数据的结果。例如，在提供者网络内的各个计算实例处或者在客户端拥有的计算装置处，可以启动工作者节点的初始集合{W1,W2,…}(元素2704)。在所述实施方案中，工作者节点可以被授予PAT1的读取和写入访问权。

[0166] 随着工作者节点在线，它们可以各自访问PAT1，以尝试找到未被分配的分区。例如，工作者节点W1可以检查PAT1，并且发现分区P1未被分配(元素2707)。例如，根据所用的数据库服务的类型，使用条件型写入请求或事务更新请求，W1随后可以更新PAT1中的P1的条目，以表明P1被分配给W1(元素2710)。更新所述表之后，W1可以使用SMS检索子系统接口开始检索P1的数据记录(元素2713)，并且可以对检索的记录执行阶段PS1的处理操作。

[0167] 同时，在某一时间点，不同的工作者节点W2可以访问PAT1，自己尝试找到未被分配的分区(元素2716)。基于W1之前的更新，W2可以确定P1已经被分配，但不同的分区P2没有被分配。在一些实施方案中，W2确定P2的当前分配工作者节点不健康或不活跃(例如，基于P2条目中的健康指示符列)也可能导致W2选择P2。因此，在至少一些实施方案中，未分配状态或者确定当前工作者节点的不健康状态都可以用来选择给定的分区，以便重新分配(或初始分配)。W2随后可以尝试更新PAT1，以将P2分配给自己(元素2719)。如果更新成功，那么W2可以使用SMS检索接口开始检索P2记录(元素2722)，并且执行针对所述阶段定义的适当处理操作。

[0168] 如之前所述，分散控制SPS中的工作者节点可以(一般是频繁地)获取来自SMS的分区映射信息，并且如有需要，使用此类信息来更新PA表。图28示出根据至少一些实施方案的可以由流处理阶段的工作者节点执行，以便基于从流管理服务控制子系统获取的信息来更新分区分配表的操作方面。如元素2801所示，在工作者节点初始化期间或者响应于各种触发条件，例如，关闭了分配给它的分区中的一个，工作者节点W1可以将请求提交到SMS控制子系统，以便获取最近或当前的分区列表或者活跃分区的列表。在一些实施中，可以为此目

的来调用getStreamInfo或类似的API。在一些实施方案中，可以使用其他触发条件：例如，工作者节点可以各自被配置成在随机的时间量之后或者响应于工作负载水平意外下降或增加而获取新的分区列表。SMS返回的分区列表可以与所述分区的PA表中条目相比较(元素2807)。在所述实施方案中，如果发现差异(例如，如果新近获取的分区列表中的一些分区不在PA表中，或者如果PA表中有条目不在SMS列表中)，那么工作者节点可以插入或删除PA表中的条目，以解决差异(元素2810)。(在一些实施中，如果当前目标为被删除的条目具有分配的工作者节点，则可能需要额外的协调，例如，可以直接或经由PA表本身来通知分配的工作者节点。)

[0169] 在调整差异之后，或者如果没有检测到差异，那么工作者节点W1可以选择其应执行阶段的处理操作所针对的分区的集合(元素2813)，并且可以相应地更新PA表。在一些情况下，根据导致检索分区列表的触发条件，W1可能已经有分配给它的一个或多个分区，并且可能不需要改变它的分配或者更新PA表。W1随后可以前进到检索所分配的一个或多个分区的数据记录并且处理所述记录，而无需与SMS控制子系统交互或者改变PA表中的条目的数量(元素2816)。最终，当检测到触发条件时(例如，当针对检索请求接收“到达分区末尾”响应的等效物，表明分区被关闭时)，W1可以再次将针对新分区信息的请求发送到SMS控制子系统，并且可以重复元素2801向前的操作。

[0170] 图29示出根据至少一些实施方案的可以由流处理阶段的工作者节点执行的负载平衡操作的方面。如元素2901所示，工作者节点W1可以确定在检测到各种触发条件中的任一个之后，例如，检测到高的资源利用水平，或者基于可配置的调度，在它的阶段上执行负载平衡分析。W1可以检查PA表中的条目(元素2904)，以确定所述阶段的各种工作负载度量。此类度量可包括分配给工作者节点的分区的平均数、工作者节点或不同分区的平均工作负载水平(在工作负载水平指示符被保存到所述表的实施方案中)、每个工作者节点的工作负载的范围或分布等等。

[0171] W1随后可以将自己的工作负载(例如，基于分配给W1的分区数量和/或每个分区的工作负载水平指示符)与度量中的一些或全部进行比较。一般来说，可以得出三种类型的结论中的任一个：W1超负载、W1负载不足或者W1的工作负载既不太高也不太低。在一些实施方案中，“太高”或“太低”的工作负载水平可以由客户端选择的策略定义，所述阶段代表所述客户端而被配置，或者在其他实施方案中，可以使用探试法的某一默认设置来定义。如果W1确定它的工作负载太低(元素2907)，例如，低于某一最小负载阈值T1，那么可以识别更忙或更高负载的工作者节点Wk(元素2910)。例如，通过尝试更改PA表中的Pm条目、请求此类更改(这可导致针对Wk生成通知)或者直接请求Wk，W1随后可以开始将一个或多个分区Pm从Wk转移给自己的过程(元素2913)。

[0172] 如果W1确定它的工负载太高(元素2916)，例如，高于最大阈值T2，那么它可以确定放弃分配分区Pn中的一个或多个(即，释放分区，以供其他工作者节点分配)(元素2919)。例如，通过从Pn的条目的受让列中移除标识符，W1随后可以更改PA表中的适当条目(元素2922)。如果W1的工作负载既不太高也不太低，或者在W1采取上述动作以增加或减少工作负载之后，W1可以重新开始处理它被分配到的分区的记录(元素2925)。当并且如果满足触发另一负载平衡分析的条件时，可以重复对应于元素2901之后的操作。应注意，在图29所示的操作中，所示W1只有在检测到自己的工作负载不平衡时才开始改变工作负载。在其他实施

方案中，如果在本身之外的其他工作者节点之中检测到不平衡，例如，如果确定W2的工作负载水平远低于W3，那么W1可以开始再平衡动作。在一些实施中，如果并且当检测到工作负载不平衡时，W1可以请求或开始动态重新分区（例如，通过调用诸如图3所示的repartitionStream SMS API或其等效物）。在一些实施方案中，图29所示的操作种类可以由新配置的工作者节点执行，例如，在一个阶段已经操作一段时间之后，有新的节点添加到所述阶段时，通过从高负载的现有节点请求分区的重新分配，新的节点可以间接通知现有节点它们的存在。在一些实施方案中，类似于上文针对SPS工作者节点描述的那些分散控制技术也可以或者改为用在一个或多个SMS子系统处，例如，摄取、存储或检索子系统的节点可以使用类似于PA表的共享数据结构来协调它们的工作负载。

[0173] 应注意，在各实施方案中，除了图17到图24和图27到图29的流程图所示的那些之外的操作可以用来实施上述流管理服务和/或流处理功能。在一些实施方案中，所示操作中的一些可以不实施，或可按照不同的顺序实施，或者并行实施而不是相继实施。还应注意，在各实施方案中，至于支持编程接口的SMS和SPS功能中的每个，一个或多个技术的任何组合可以用来实施接口，包括使用web页、web站、web服务API、其他API、命令行工具、图形用户接口、移动应用(app)、平板app等。

[0174] 使用情况

[0175] 建立用于收集、存储、检索和阶段化处理流数据记录的可扩展型基于分区的动态配置管理多租户服务的上述技术可以用于很多情形。例如，大型提供者网络可以包括同时为数万客户端实施很多不同的多租户或单租户服务的服务实例的数千实例主机。安装在各个实例和主机上的监控和/或计费代理可以迅速生成数千度量记录，所述记录可能需要存储和分析以产生准确的计费记录、确定提供者网络的数据中心的有效供应计划、检测网络攻击等。监控记录可以形成SMS的输入流以便可扩展地摄取和存储，并且所述SPS技术可以实施用于分析所收集的度量。类似地，收集和分析来自很多日志源(例如，来自分布式应用的节点的应用日志，或者来自数据中心处的主机或计算实例的系统日志)的大量日志记录的应用也能够利用SMS和SPS功能。在至少一些环境中，SPS处理操作可以包括实时ETL(提取-转换-加载)处理操作(即，实时将接收的数据记录转换以加载到目的地而不是进行离线转换的操作)，或者用于插入到数据仓库中的数据记录的转换。在数据可以插入到仓库中以供分析之前，将SMS/SPS组合用于实时将数据加载到数据仓库中可以避免清理和精确一个或多个数据源的数据一般所需的延迟。

[0176] 使用SMS和SPS技术还可以建立许多不同的“大数据”应用。例如，使用流可以有效执行对社交媒体交互的各种形式的趋势进行分析。从移动电话或平板计算机收集的数据，例如，用户的位置信息，可以作为流记录进行管理。例如，从一组监控摄像机收集的音频或视频信息可以表示可以用可扩展的方式收集和处理的另一类别的流数据集，从而可能有助于防止各种攻击。要求分析日益增长的数据集(例如，从气象卫星、海洋传感器、森林传感器、天文望远镜收集的数据集)的科学应用也可受益于本文所述的流管理和处理能力。基于策略的灵活配置选项和价格选项可以帮助不同类型的用户自定义流功能，以适应具体的预算和数据耐久性/可用性需求。

[0177] 本发明的实施方案可以根据以下条款进行描述：

[0178] 1.一种系统，其包括：

[0179] 一个或多个计算装置，其被配置成：

[0180] 通过多租户流管理服务的一个或多个控制部件来确定(a)记录摄取子系统、(b)记录存储子系统和(c)记录检索子系统的相应节点集合，其中所述一个或多个控制部件被分配给特定数据流，所述特定数据流包括由一个或多个数据生产者生成的多个数据记录，其中所述记录摄取子系统、所述记录存储子系统和所述记录检索子系统中的每个子系统包括由所述一个或多个控制部件基于包括分区策略的一个或多个策略而动态配置的一个或多个节点；

[0181] 接收经由在所述记录摄取子系统处实施的一个或多个编程记录提交接口提交的数据记录，其中所述一个或多个编程记录提交接口包括支持数据记录的串联提交的第一提交接口以及使得能够通过引用存储数据的网络地址来进行数据记录提交的第二提交接口；

[0182] 响应于经由在所述记录检索子系统处实施的一个或多个编程记录检索接口接收的数据记录检索请求，提供数据记录的内容，其中所述一个或多个编程记录检索接口包括实现非顺序访问模式的第一检索接口以及实现顺序访问模式的第二检索接口，并且其中与使用所述第一检索接口相关联的计费率与使用所述第二检索接口相关联的计费率不同；

[0183] 至少部分基于所述多个记录检索接口和所述多个记录提交接口的相应使用计数度量，生成与特定数据流相关联的客户端计费量。

[0184] 2.根据条款1所述的系统，其中所述一个或多个计算装置还被配置成：

[0185] 根据所述分区策略，至少部分基于与所述特定数据流的特定数据记录相关联的键，将所述特定数据记录分配到所述特定数据流的分区，其中所述键由对应于所述特定数据记录的写入请求指示。

[0186] 3.根据条款1所述的系统，其中所述一个或多个计算装置还被配置成：

[0187] 至少部分基于特定数据记录被分配到的分区，选择下列项中的一个或多个：(a)负责接受所述特定数据记录的所述记录摄取子系统的特定节点，(b)负责存储所述特定数据记录的至少一个复本的所述记录存储子系统的特定节点，以及(c)负责响应于读取请求而从所述记录存储子系统获取所述特定数据记录的所述检索子系统的特定节点。

[0188] 4.根据条款1所述的系统，其中所述记录摄取子系统、所述记录存储子系统和所述记录检索子系统中的至少一个子系统包括被所述一个或多个控制部件配置成冗余组的成员的多个节点，其中所述冗余组包括：(a)被分配来对数据记录的集合执行操作的一个或多个主节点，以及(b)被配置成响应于检测到一个或多个触发事件而承担主节点角色的一个或多个非主节点。

[0189] 5.根据条款1所述的系统，其中所述一个或多个计算装置还被配置成：

[0190] 生成对应于特定数据记录的特定序列号，所述特定序列号表示相对于所述特定数据记录所属的分区的其他数据记录而言，在所述记录摄取子系统处接收到所述特定数据记录的顺序；

[0191] 至少部分基于针对数据记录生成的相应序列号，由所述记录存储子系统按一定顺序存储多个数据记录，所述多个数据记录包括所述特定数据记录；以及

[0192] 响应于以所述特定序列号作为参数调用所述第一检索接口的读取请求，从所述记录存储子系统中检索所述特定数据记录。

[0193] 6.一种方法，其包括：

[0194] 由一个或多个计算装置执行：

[0195] 针对包括多个数据记录的特定数据流，确定节点集合，所述节点集合可由一个或多个控制部件配置成基于包括流分区策略的一个或多个策略来执行流管理操作；

[0196] 响应于经由一个或多个编程记录检索接口接收的数据检索请求，提供数据记录，其中所述一个或多个编程记录检索接口包括实现非顺序访问模式的第一检索接口以及实现顺序访问模式的第二检索接口，并且其中与使用所述第一检索接口相关联的计费率与使用所述第二检索接口相关联的计费率不同；以及

[0197] 至少部分基于所述多个记录检索接口的相应使用计数度量，生成与所述特定数据流相关联的客户端计费量。

[0198] 7.根据条款6所述的方法，其还包括由所述一个或多个计算装置执行：

[0199] 根据所述分区策略，至少部分基于与所述特定数据流的特定数据记录相关联的键，将所述特定数据记录分配到所述特定数据流的第一分区，其中所述键由对应于所述特定数据记录的写入请求指示。

[0200] 8.根据条款6所述的方法，其还包括由所述一个或多个计算装置执行：

[0201] 至少部分基于特定数据记录被分配到的分区，选择下列项中的一个或多个：(a)负责接受所述特定数据记录的记录摄取子系统的特定节点，(b)负责存储所述特定数据记录的至少一个复本的记录存储子系统的特定节点，以及(c)负责响应于读取请求而从所述记录存储子系统获取所述特定数据记录的检索子系统的特定节点。

[0202] 9.根据条款6所述的方法，其还包括由所述一个或多个计算装置执行：

[0203] 接收经由一个或多个编程记录提交接口提交的数据记录，其中所述一个或多个编程记录提交接口包括支持数据记录的串联提交的第一提交接口以及使得能够通过引用下列中的一个来提交数据记录的第二提交接口：(a)提供者网络实施的存储服务处的对象地址，(b)通用记录定位符，(c)数据库记录。

[0204] 10.根据条款6所述的方法，其还包括由所述一个或多个计算装置执行：

[0205] 至少部分基于下列项中的一个，从数据存储子系统的特定节点移除特定数据记录：(a)针对所述特定数据流配置的数据去除重复窗口，(b)与所述特定数据流相关联的数据存档策略，(c)客户端指明的数据保留策略，(d)移除所述特定数据记录的客户端请求，或者(e)所述特定数据记录已被一个或多个数据消费者处理的指示。

[0206] 11.根据条款6所述的方法，其中针对所述特定数据流配置的(a)记录摄取子系统、(b)记录存储子系统或者(c)记录检索子系统中的至少一个子系统包括被所述一个或多个控制部件配置成冗余组的成员的多个节点，其中所述冗余组包括：(a)被分配来对所述流的数据记录的集合执行操作的一个或多个主节点，以及(b)被配置成响应于一个或多个触发事件而承担主节点角色的一个或多个非主节点。

[0207] 12.根据条款11所述的方法，其中在特定数据中心处将所述一个或多个主节点中的特定主节点实例化，并且其中在不同的数据中心处将所述一个或多个非主节点中的特定非主节点实例化。

[0208] 13.根据条款11所述的方法，其还包括由所述一个或多个计算装置执行：

[0209] 检测所述一个或多个触发事件中的触发事件；以及

[0210] 通知所述一个或多个非主节点中的特定非主节点来承担所述主节点角色。

[0211] 14.根据条款6所述的方法，其中所述一个或多个控制部件包括被配置成冗余组的多个控制节点，包括：主控制节点，其被配置成响应对包括所述特定数据流的一个或多个数据流进行控制平面操作的请求；以及至少一个非主控制节点，其被配置成响应于一个或多个触发事件而承担主节点角色。

[0212] 15.根据条款6所述的方法，其还包括由所述一个或多个计算装置执行：

[0213] 在提供者网络的网络可访问数据库处，存储所述特定数据流的元数据，所述元数据包括根据所述流分区策略生成的分区映射；以及

[0214] 从记录检索子系统访问所述元数据，以响应于特定记录检索请求。

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

[0216] 生成对应于所述数据流的特定数据记录的特定序列号，所述特定序列号表示相对于所述特定数据记录所属的分区的其他数据记录而言，在记录摄取子系统处接收到所述特定数据记录的顺序；

[0217] 至少部分基于针对数据记录生成的相应时间戳，由所述记录存储子系统按一定顺序存储多个数据记录，所述多个数据记录包括所述特定数据记录；以及

[0218] 响应于接收到以所述特定序列号作为参数调用所述第一检索接口的读取请求，从所述记录存储子系统中检索所述特定数据记录。

[0219] 17.根据权利要求6所述的方法，其中所述节点集合中的特定节点包括下列一个：(a)在提供者网络处实施的数据库服务的数据库表的至少一部分，或者(b)在提供者网络处实施的存储服务的存储对象的至少一部分。

[0220] 18.一种存储程序指令的非暂时计算机可访问存储介质，所述程序指令在一个或多个处理器上执行时实施多租户流管理服务的控制节点，其中所述控制节点被配置成：

[0221] 接收将特定数据流初始化的请求，所述特定数据流包括由一个或多个数据生产者生成的多个数据记录；

[0222] 至少部分基于与所述特定数据流相关联的分区策略，确定将用于针对所述特定数据流来配置一个或多个子系统的一个或多个参数，所述一个或多个子系统包括记录检索子系统，其中所述一个或多个参数包括将在记录检索子系统中实例化的节点的初始数量；

[0223] 识别将用于所述记录检索子系统的特定节点的一个或多个资源，其中所述特定节点被配置成实施多个编程记录检索接口，包括实现非顺序访问模式的第一检索接口以及实现顺序访问模式的第二检索接口；以及

[0224] 使用所述一个或多个资源来配置所述记录检索子系统的所述特定节点。

[0225] 19.根据条款18所述的非暂时计算机可访问存储介质，其中所述一个或多个子系统包括：记录摄取子系统，其包括可配置成接收所述特定数据流的数据记录的一个或多个节点；以及记录存储子系统，其可配置成将所述流的数据记录存储在选择的存储位置集合处。

[0226] 20.根据条款19所述的非暂时计算机可访问存储介质，其中所述记录摄取子系统、所述记录存储子系统和所述记录检索子系统中的至少一个子系统包括被配置成冗余组的成员的多个节点，其中所述冗余组包括：(a)被分配来对所述流的数据记录的集合执行操作的一个或多个主节点，以及(b)被配置成响应于一个或多个触发事件而承担主节点角色的一个或多个非主节点。

[0227] 21.根据条款19所述的非暂时计算机可访问存储介质，所述记录摄取子系统、所述记录存储子系统或者所述记录检索子系统中的一个的至少一个节点包括在提供者网络处实施的虚拟化计算服务的计算实例的部件。

[0228] 22.根据条款19所述的非暂时计算机可访问存储介质，其中所述一个或多个数据生产者包括下列一个或多个：(a)计算装置的日志部件，(b)社交媒体应用，(c)计算装置处的监控代理，(d)传感器装置，(e)移动电话，或者(f)平板计算机。

[0229] 说明性计算机系统

[0230] 在至少一些实施方案中，实施本文所述的一种或多种技术(包括实施SMS子系统(例如，摄取、存储、检索和控制子系统)的部件以及SPS工作者和控制节点的技术)的一部分或全部的服务器可以包括通用计算机系统，所述通用计算机系统包括或被配置成访问一个或多个计算机可访问介质。图30示出此类通用计算装置9000。在所示实施方案中，计算装置9000包括经由输入/输出(I/O)接口9030耦接到系统存储器9020的一个或多个处理器9010。计算装置9000还包括耦接到I/O接口9030的网络接口9040。

[0231] 在各实施方案中，计算装置9000可以是包括一个处理器9010的单处理器系统，或者是包括若干处理器9010(例如，两个、四个、八个或者另一合适的数量)的多处理器系统。处理器9010可以是能够执行指令的任何合适的处理器。例如，在各实施方案中，处理器9010可以是实施多种指令集架构(ISA)中的任一个的通用或嵌入式处理器，例如，x86、PowerPC、SPARC或MIPS ISA，或者任何其他合适的ISA。在多处理器系统中，每个处理器9010通常可以但不必实施相同的ISA。在一些实施中，替代于传统处理器或除此之外，可以使用图形处理单元(GPU)。

[0232] 系统存储器9020可以被配置成存储可由处理器9010访问的指令和数据。在各实施方案中，系统存储器9020可以使用任何合适的存储器技术实施，例如，静态随机存取存储器(SRAM)、同步动态RAM(SDRAM)、非易失性/快闪型存储器或者任何其他类型的存储器。在所述实施方案中，实施一个或多个所需功能(例如，上文所述的那些方法、技术和数据)的程序指令和数据被示出作为代码9025和数据9026存储在系统存储器9020中。

[0233] 在一个实施方案中，I/O接口9030可以被配置成协调处理器9010、系统存储器9020以及所述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或者其他外围接口，例如，用来存储数据对象分区的物理复本的各种类型的永久性和/或易失性存储装置。在一些实施方案中，I/O接口9030可以执行任何需要的协议、计时或者其他数据转换，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用于另一部件(例如，处理器9010)的格式。在一些实施方案中，I/O接口9030可以包括支持通过各种类型的外围总线附接的装置，例如，外围部件互连(PCI)总线标准或通用串行总线(USB)标准的变体。在一些实施方案中，功能I/O接口9030可以分成两个或更多单独的部件，例如，北桥和南桥。此外，在一些实施方案中，I/O接口9030的一些功能或全部功能(例如，通往系统存储器9020的接口)可以直接并入到处理器9010中。

[0234] 网络接口9040可以被配置成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之间交换，例如，图1到图29所示的其他计算机系统或装置。在各实施方案中，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进行的通信，例如，以太网的类型。此外，网络接口9040可以支持经由电信/电话网络(例如，模拟语音网络

或数字光纤通信网络)进行的通信、经由存储区域网络(例如,光纤通道SAN)进行的通信,或者经由任何其他合适类型的网络和/或协议进行的通信。

[0235] 在一些实施方案中,系统存储器9020可以是计算机可访问介质的一个实施方案,其被配置成存储程序指令和数据,如上文针对图1到图29所述,以用于实施对应方法和设备的实施方案。然而,在其他实施方案中,程序指令和/或数据可以被接收、发送或者存储在不同类型的计算机可访问介质上。一般来说,计算机可访问介质可以包括磁或光学介质等非暂时存储介质或存储器介质,例如,经由I/O接口9030耦接到计算装置9000的磁盘或DVD/CD。非暂时计算机可访问存储介质还可以包括任何易失性或非易失性介质,例如,RAM(例如,SDRAM、DDR SDRAM、RDRAM、SRAM等)、ROM等,所述易失性或非易失性介质可以被包括在计算装置9000的一些实施方案中,作为系统存储器9020或另一类型的存储器。此外,计算机可访问介质可以包括经由通信介质(例如,诸如可经由网络接口9040实施的网络和/或无线链路)传送的传输介质或信号,例如,电信号、电磁信号或数字信号。在各实施方案中,诸如图30所示的多个计算装置的部分或全部可用来实施所述功能,例如,在多种不同的装置和服务器上运行的软件部件可以合作,以提供功能。在一些实施方案中,除了使用通用计算机系统实施之外或作为其替代,所述功能的部分可以使用存储装置、网络装置或者专用计算机系统来实施。本文中使用的术语“计算装置”是指至少所有这些类型的装置,但不限于这些类型的装置。

[0236] 总结

[0237] 各实施方案还可以包括在计算机可访问介质上接收、发送或存储根据上述说明实施的指令和/或数据。一般来说,计算机可访问介质可以包括磁性或光学介质等存储介质或存储器介质,例如,磁盘或DVD/CD-ROM,易失性或非易失性介质,例如,RAM(例如,SDRAM、DDR、RDRAM、SRAM等)、ROM等,以及经由网络和/或无线链路等通信介质传送的传输介质或者信号,例如,电信号、电磁信号或数字信号。

[0238] 附图示出并且本文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,并且各种元件可以被添加、重新排序、组合、忽略、更改等。

[0239] 受益于本公开的本领域内的技术人员将明白可以进行各种更改和变化。本发明意图涵盖所有此类更改和变化,且因此,上述描述应被视作说明性的,而非限制性的意义。

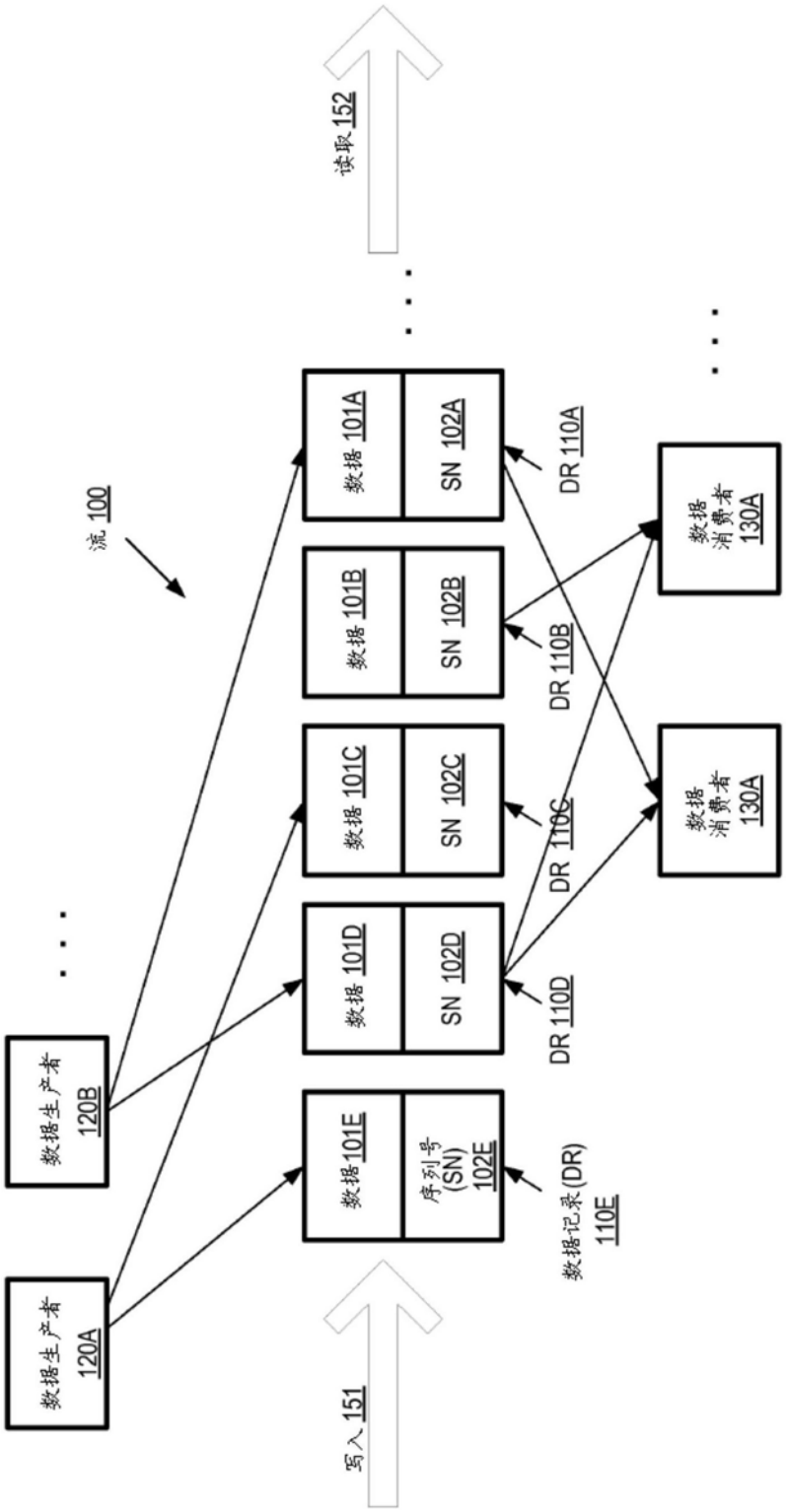

读取 152
流 100
数据 101A
SN 102A
数据 101B
SN 102B
数据 101C
SN 102C
数据 101D
SN 102D
数据101E
序列号 (SN) 102E
DR 110A
DR 110B
DR 110C
DR 110D
数据记录(DR) 110E
数据生产者 120B
数据生产者 120A
数据消费者 130A
数据消费者 130A
写入 151

图1

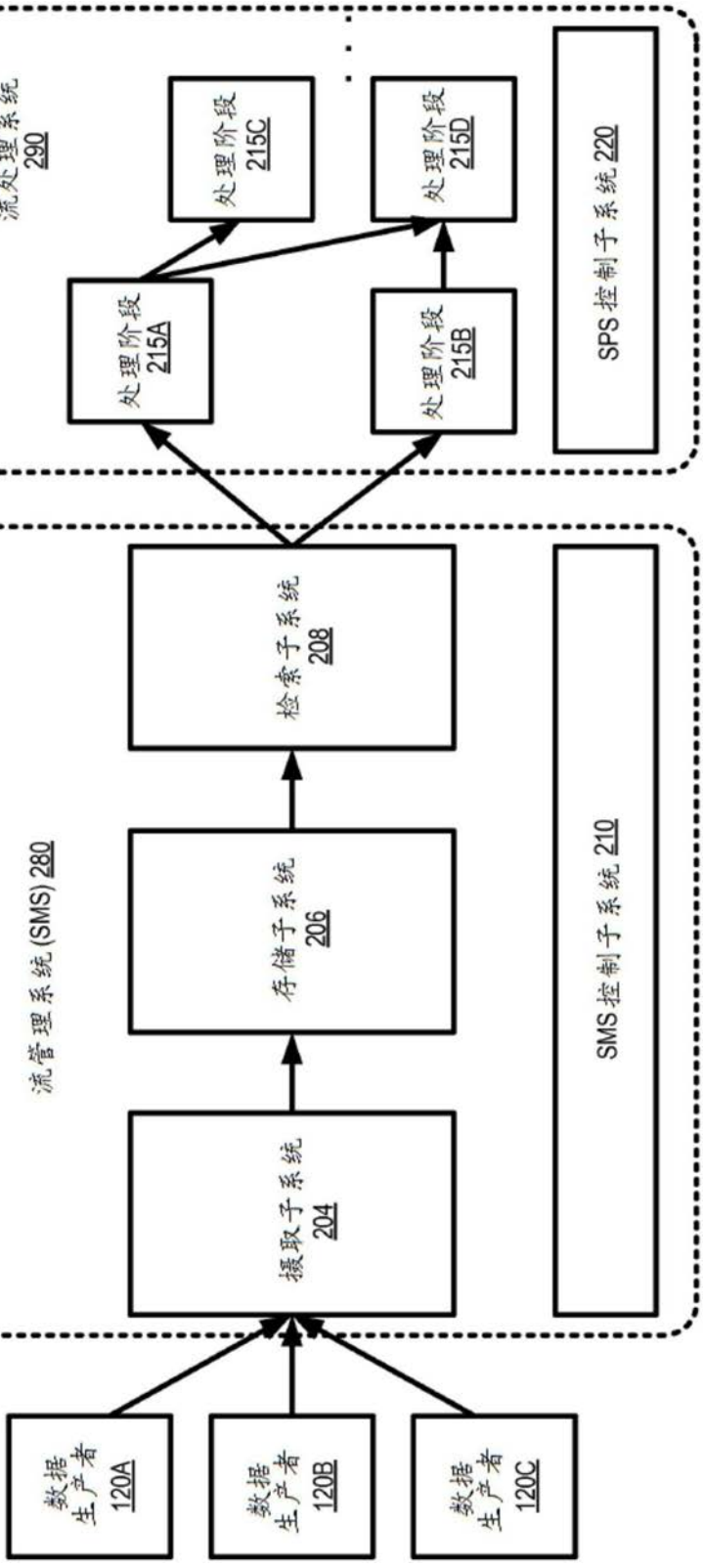

流处理系统
290
处理阶段
215C
处理阶段
215D
SPS 控制子系统 220
处理阶段
215A
处理阶段
215B
流管理系统 (SMS) 280
检索子系统
208
存储子系统
206
摄取子系统
204
SMS 控制子系统 210
数据
生产者
120A
数据
生产者
120B
数据
生产者
120C

图2


SPS客户端 375
SPS 客户端调用SPS 接口
350
SPS 编程接口 305
createStreamProcessingStage(inputStream, IdempotentOperations, outputPolicy, recoveryPolicy),
initializeWorkerNode(processingStage),
saveCheckpoint(processingStage,workerProcess),
setOutputDistribution(processingStage, outputStreamList, outputPartitioningPolicy)
linkStages(sourceStage, destinationStage)
createThirdPartyConnector(sourceStage,destinationThirdPartySystemSpec)
...
SPS调用SMS 接口，
作为较低级操作 352
SMS 编程接口
307
createStream()
deleteStream()
getStreamInfo()
putRecord()
getIterator()
getNextRecords()
repartitionStream()
...
直接调用的
SMS 接口354
SMS 客户端 370

图3

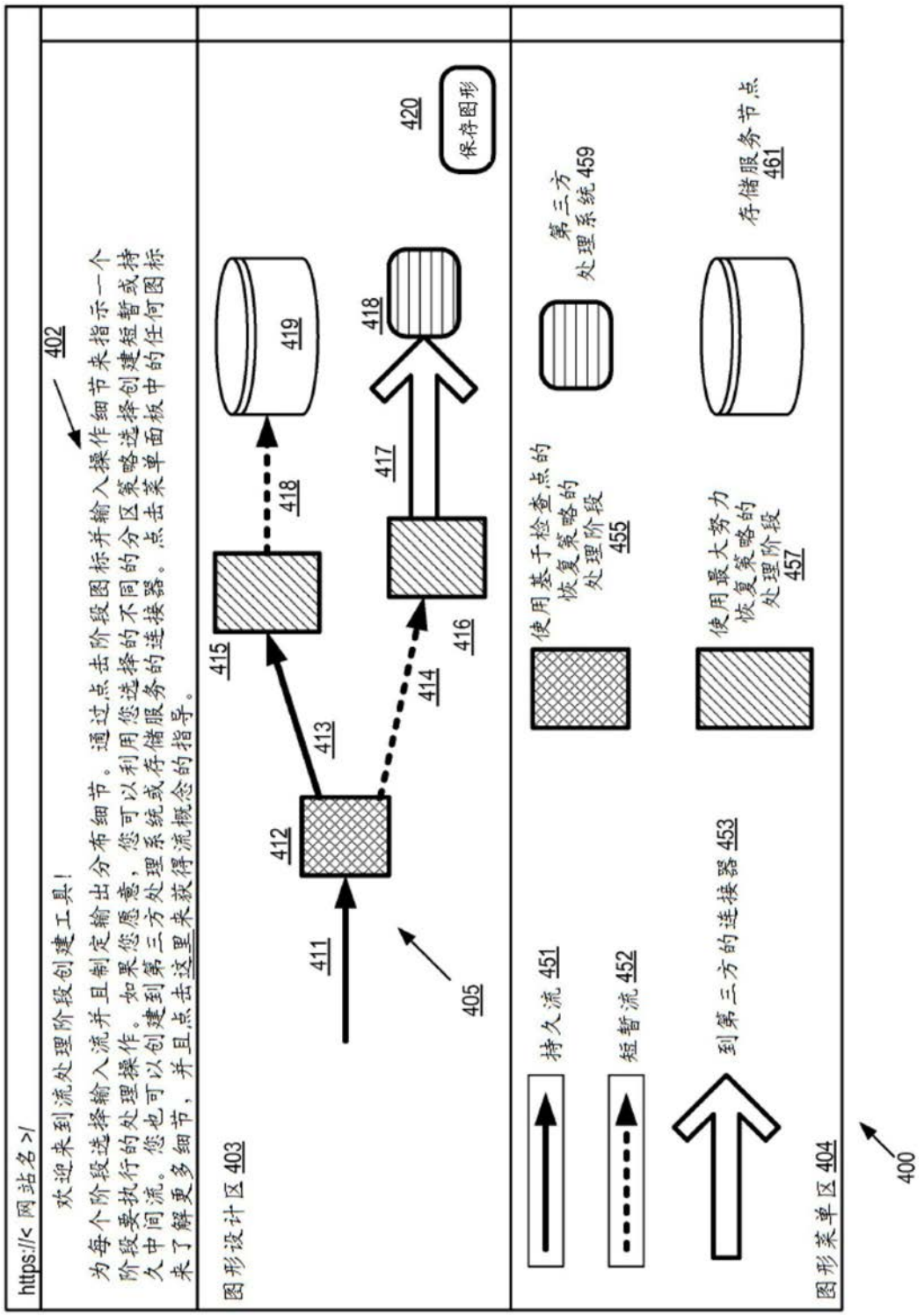

https://< 网站名>/
402
欢迎来到流处理阶段创建工具！
为每个阶段选择输入流并且制定输出分布细节。通过点击阶段图标并输入操作细节来指示一个阶段要执行的处理操作。如果您愿意，您可以利用您选择的不同的分区策略选择创建暂时或持久中间流。您也可以创建到第三方处理系统或存储服务的连接器。点击菜单面板中的任何图标来了解更多细节，并且点击这里来获得流概念的指导。
图形设计区 403
411
412
413
414
415
416
417
418
419
420
保存图形
405
持久流 451
短暂流 452
到第三方的连接器 453
使用基于检查点的恢复策略的处理阶段 455
第三方处理系统 459
使用最大努力恢复策略的处理阶段 457
存储服务节点 461
图形菜单区 404
400

图4

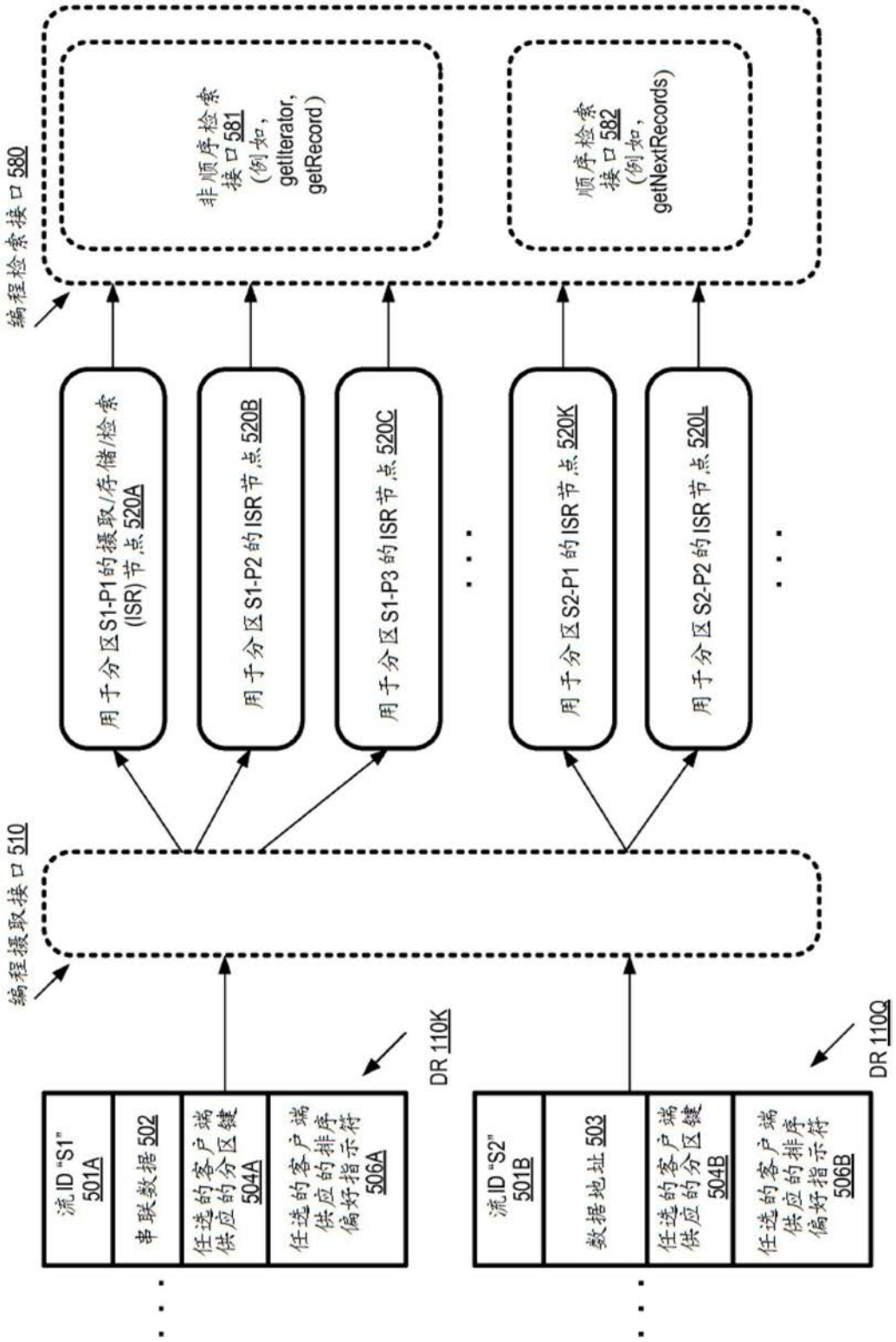

编程检索接口 580
非顺序检索接口 581 (例如, getIterator, getRecord)
顺序检索接口 582 (例如, getNextRecords)
用于分区 S1-P1 的摄取/存储/检索 (ISR) 节点 520A
用于分区 S1-P2 的 ISR 节点 520B
用于分区 S1-P3 的 ISR 节点 520C
用于分区 S2-P1 的 ISR 节点 520K
用于分区 S2-P2 的 ISR 节点 520L
编程摄取接口 510
流 ID "S1" 501A
串联数据 502
任选的客户端供应的分区键 504A
任选的客户端供应的排序偏好指示符 506A
DR 110K
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
DR 110Q

图5

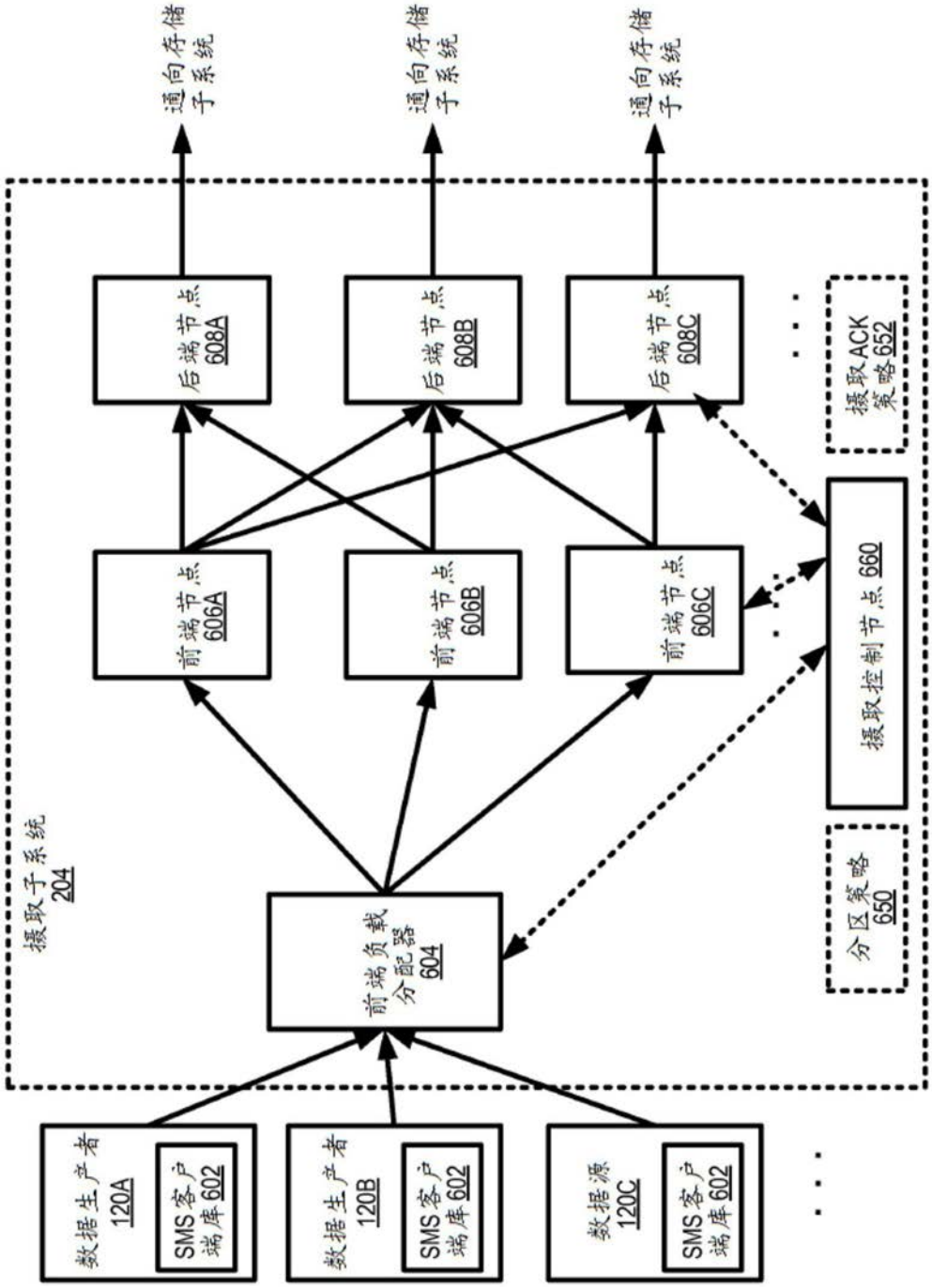

通向存储子系统
通向存储子系统
通向存储子系统
后端节点 608A
后端节点 608B
后端节点 608C
摄取 ACK 策略 652
前端节点 606A
前端节点 606B
前端节点 606C
摄取控制节点 660
摄取子系统 204
前端负载分配器 604
分区策略 650
数据生产者 120A
SMS 客户端库 602
数据生产者 120B
SMS 客户端库 602
数据源 120C
SMS 客户端库 602

图6

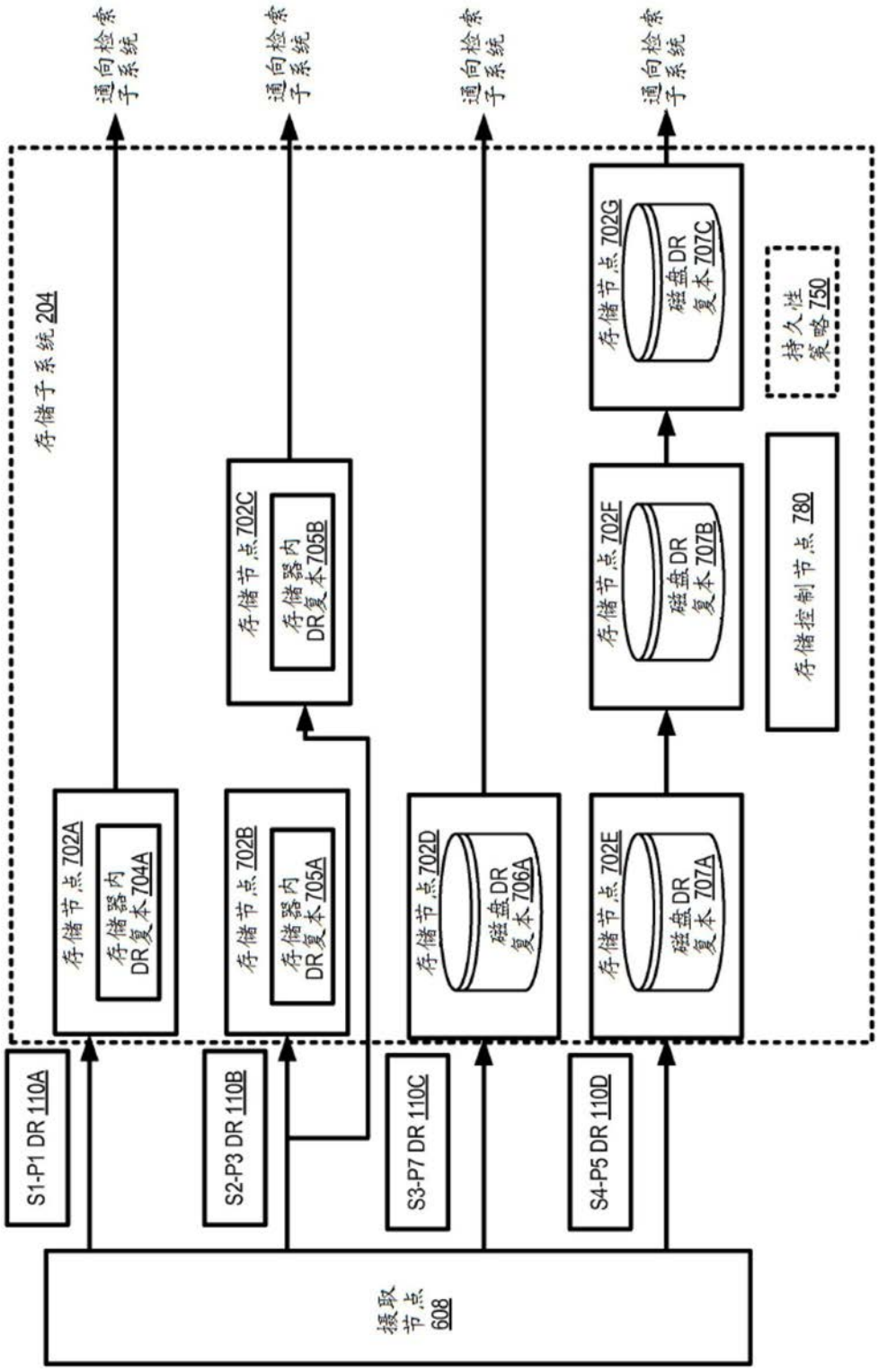

通向检索子系统
通向检索子系统
通向检索子系统
通向检索子系统
存储子系统 204
存储节点 702A
存储器内 DR复本 704A
存储节点 702B
存储器内 DR复本 705A
存储节点 702C
存储器内 DR复本 705B
存储节点 702D
磁盘DR 复本 706A
存储节点 702E
磁盘DR 复本 707A
存储节点 702F
磁盘DR 复本 707B
存储节点 702G
磁盘DR 复本 707C
持久性 策略 750
存储控制节点 780
S1-P1 DR 110A
S2-P3 DR 110B
S3-P7 DR 110C
S4-P5 DR 110D
摄取节点 608

图7

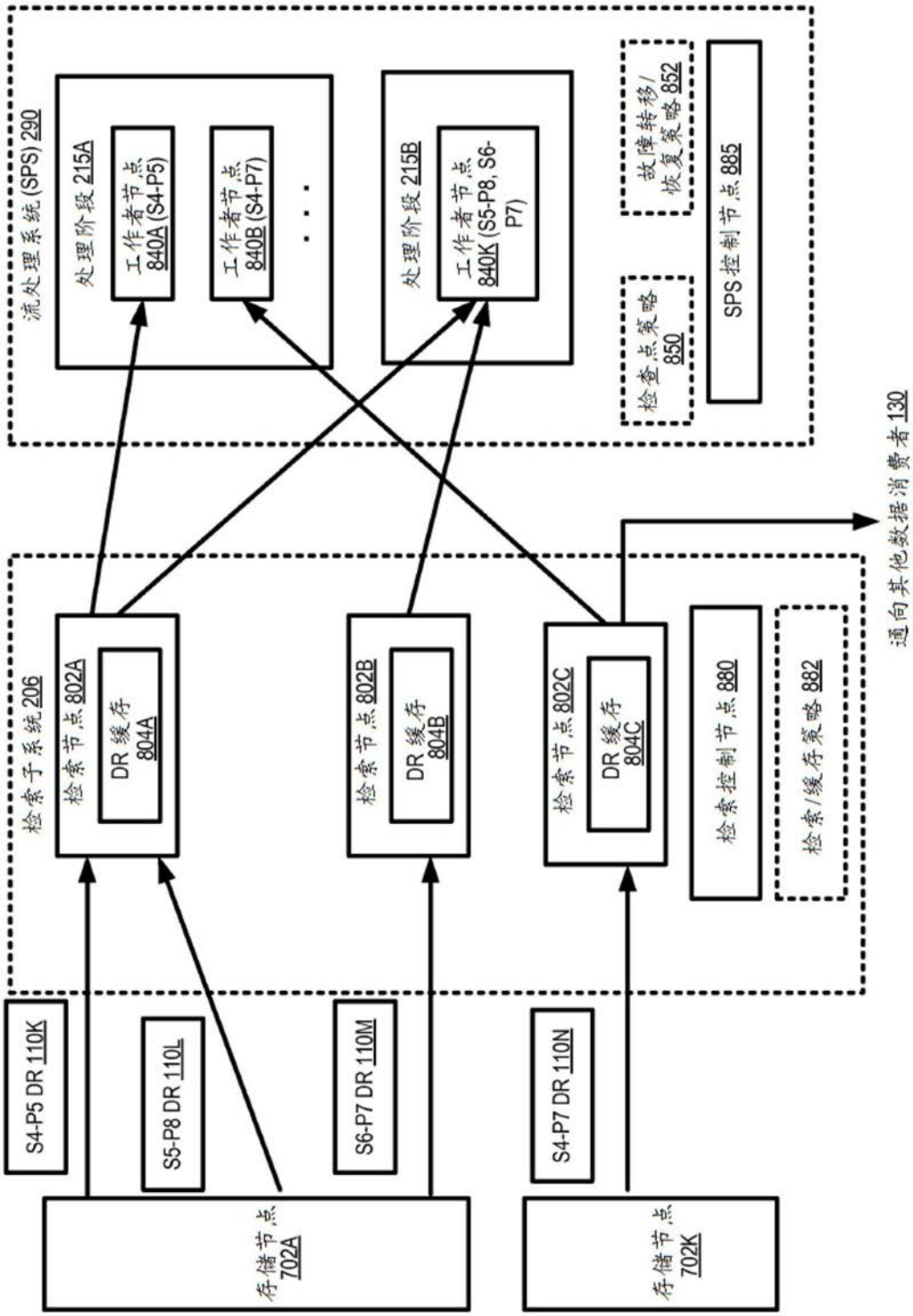

流处理系统(SPS) 290
处理阶段 215A
工作者节点 840A (S4-P5)
工作者节点 840B (S4-P7)
处理阶段 215B
工作者节点 840K (S5-P8, S6-P7)
故障转移/恢复策略 852
SPS 控制节点 885
检查点策略 850
检索子系统 206
检索节点 802A
DR 缓存 804A
检索节点 802B
DR 缓存 804B
检索节点 802C
DR 缓存 804C
检索控制节点 880
检索/缓存策略 882
通向其他数据消费者 130
S4-P5 DR 110K
S5-P8 DR 110L
S6-P7 DR 110M
S4-P7 DR 110N
存储节点 702A
存储节点 702K

图8

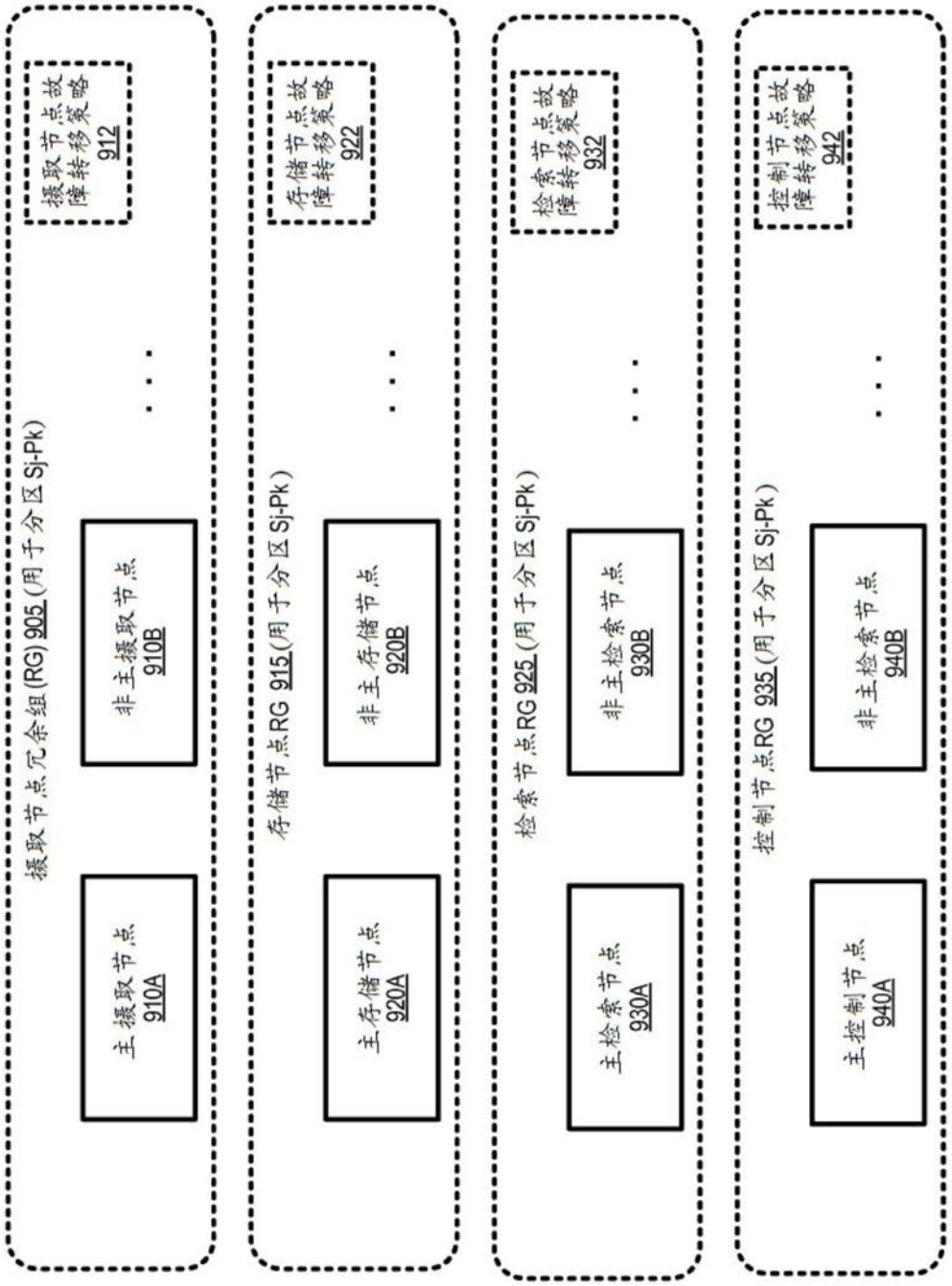

摄取节点冗余组(RG)905(用于分区Sj-Pk)
主摄取节点 910A
非主摄取节点 910B
摄取节点故障转移策略 912
存储节点RG 915(用于分区Sj-Pk)
主存储节点 920A
非主存储节点 920B
存储节点故障转移策略 922
检索节点RG 925(用于分区Sj-Pk)
主检索节点 930A
非主检索节点 930B
检索节点故障转移策略 932
控制节点RG 935(用于分区Sj-Pk)
主控制节点 940A
非主检索节点 940B
控制节点故障转移策略 942

图9

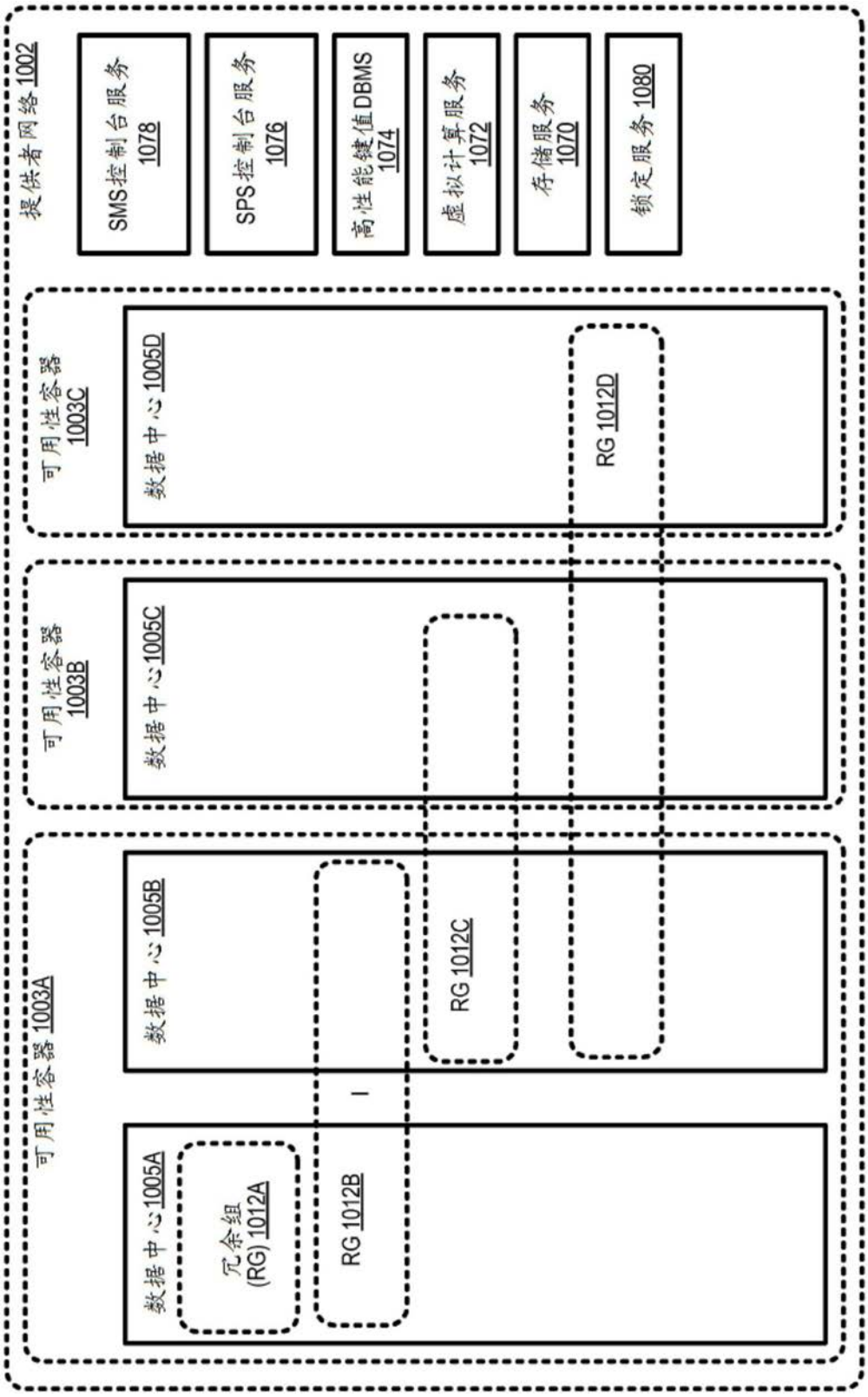

提供者网络 1002
SMS控制台服务 1078
SPS控制台服务 1076
高性能键值DBMS 1074
虚拟计算服务 1072
存储服务 1070
锁定服务 1080
可用性容器 1003C
数据中心 1005D
RG 1012D
可用性容器 1003B
数据中心 1005C
RG 1012C
可用性容器 1003A
数据中心 1005B
数据中心 1005A
冗余组 (RG) 1012A
RG 1012B

图10

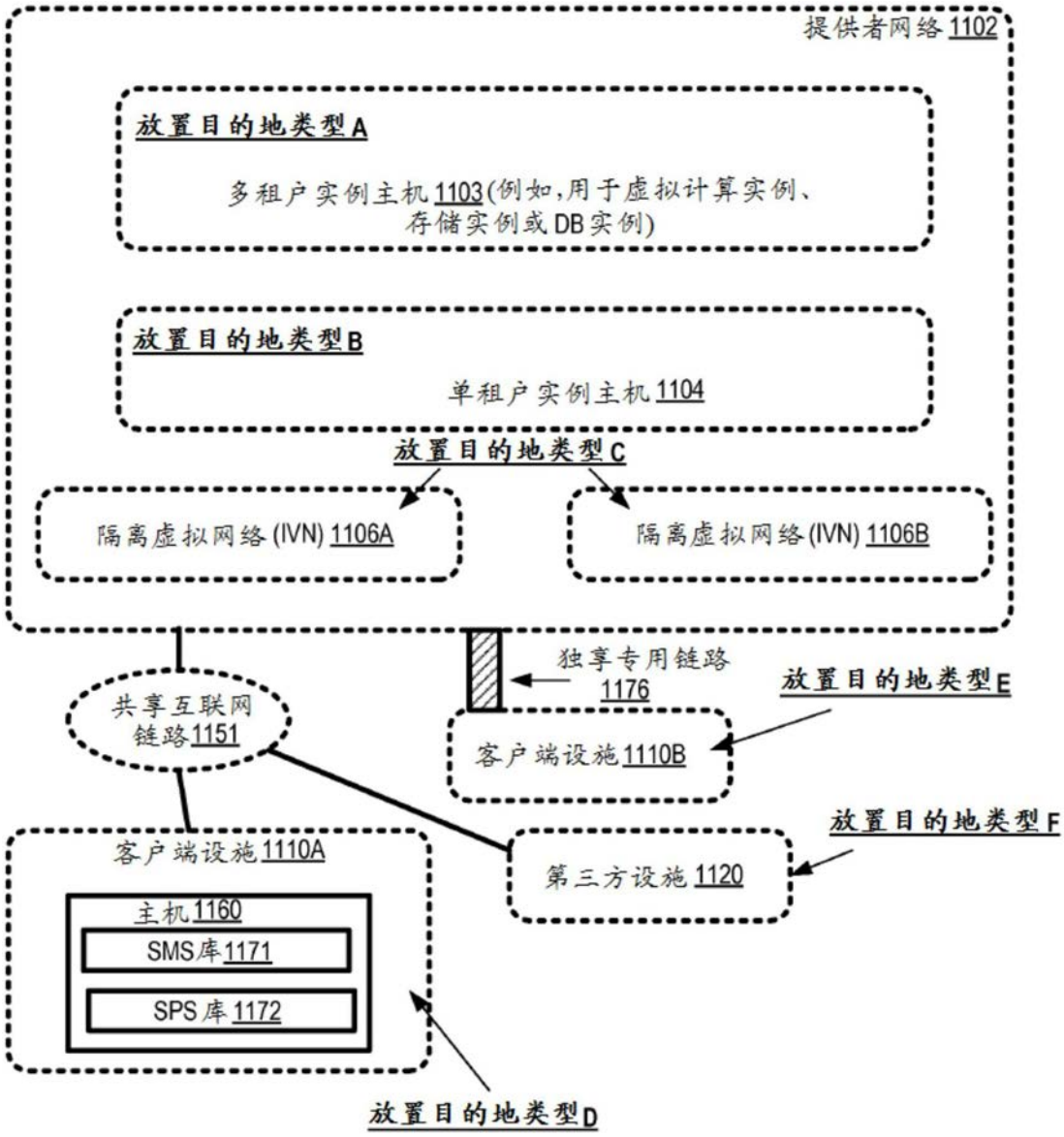

提供者网络 1102
放置目的地类型A
多租户实例主机 1103(例如,用于虚拟计算实例、存储实例或DB实例)
放置目的地类型B
单租户实例主机 1104
放置目的地类型C
隔离虚拟网络(IVN) 1106A
隔离虚拟网络(IVN) 1106B
独享专用链路 1176
放置目的地类型E
共享互联网链路 1151
客户端设施 1110B
放置目的地类型F
客户端设施 1110A
第三方设施 1120
主机 1160
SMS库 1171
SPS库 1172
放置目的地类型D

图11


SPS安全选项请求
1200
阶段ID 1210
SPS控制节点放置目的地类型(PDT) 1212
SPS工作者节点PDT 1214
加密设置 1216

图12a


SMS安全选项请求
1250
流ID 1252
摄取节点PDT 1254
摄取控制节点PDT 1256
存储节点PDT 1258
存储控制节点PDT 1260
检索节点PDT 1262
检索控制节点PDT 1264
加密设置 1266

图12b

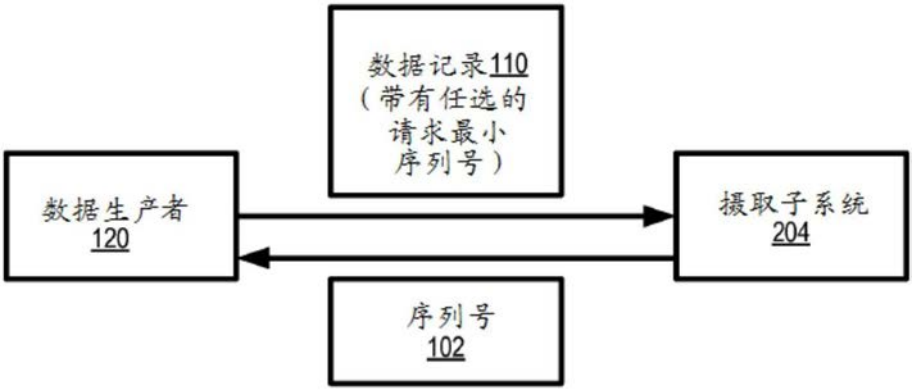

数据记录110
(带有任选的
请求最小
序列号)
数据生产者
120
摄取子系统
204
序列号
102

图13a

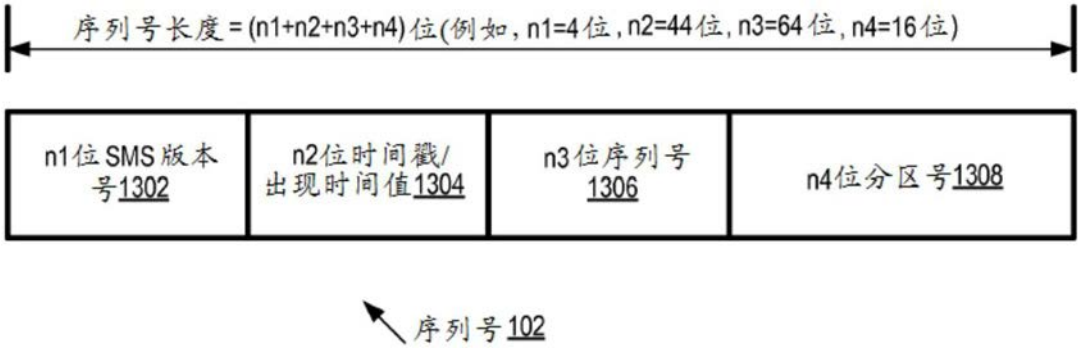

序列号长度=(n1+n2+n3+n4)位(例如,n1=4位,n2=44位,n3=64位,n4=16位)
n1位SMS版本号1302
n2位时间戳/出现时间值1304
n3位序列号1306
n4位分区号1308
序列号102

图13b

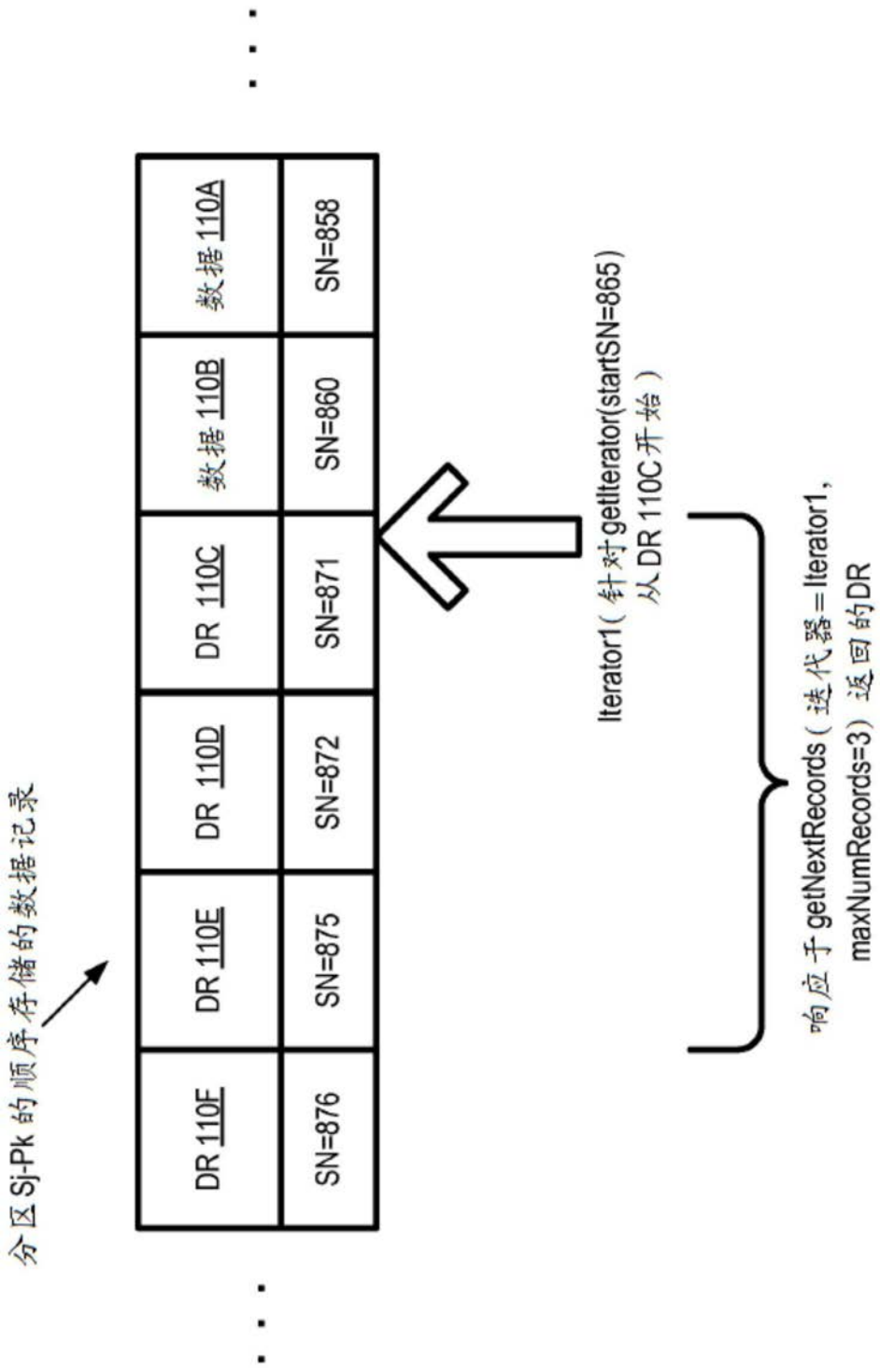

分区Sj-Pk的顺序存储的数据记录
DR 110F
DR 110E
DR 110D
DR 110C
数据 110B
数据 110A
SN=876
SN=875
SN=872
SN=871
SN=860
SN=858
Iterator1(针对getIterator(startSN=865)
从DR 110C开始)
响应于getNextRecords(迭代器=Iterator1,
maxNumRecords=3)返回的DR

图14

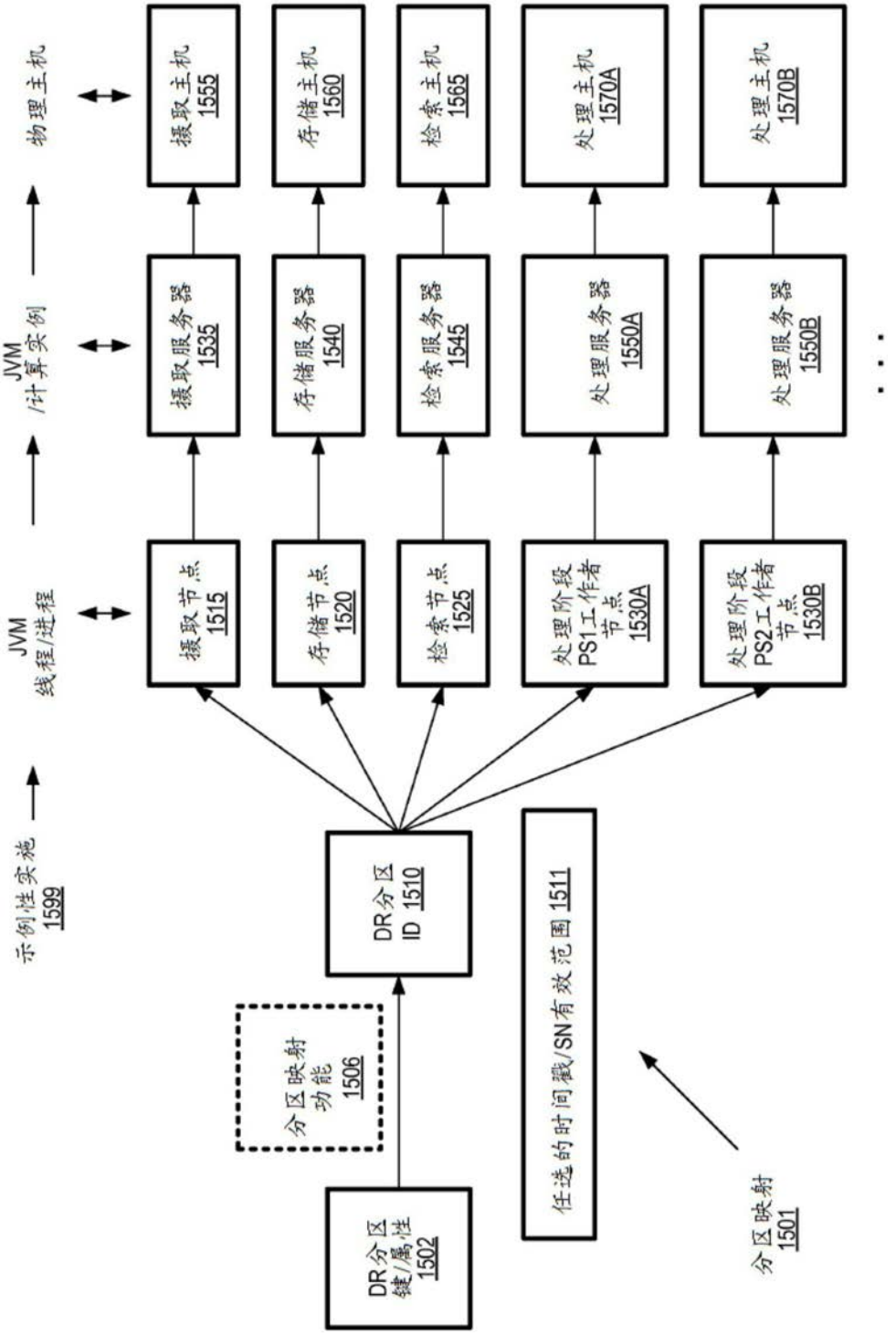

示例性实施 1599
JVM 线程/进程
JVM /计算实例
物理主机
摄取节点 1515
存储节点 1520
检索节点 1525
处理阶段 PS1工作者 节点 1530A
处理阶段 PS2工作者 节点 1530B
摄取服务器 1535
存储服务器 1540
检索服务器 1545
处理服务器 1550A
处理服务器 1550B
摄取主机 1555
存储主机 1560
检索主机 1565
处理主机 1570A
处理主机 1570B
DR分区 ID 1510
分区映射 功能 1506
DR分区 键/属性 1502
任选的时间戳/SN有效范围 1511
分区映射 1501

图15

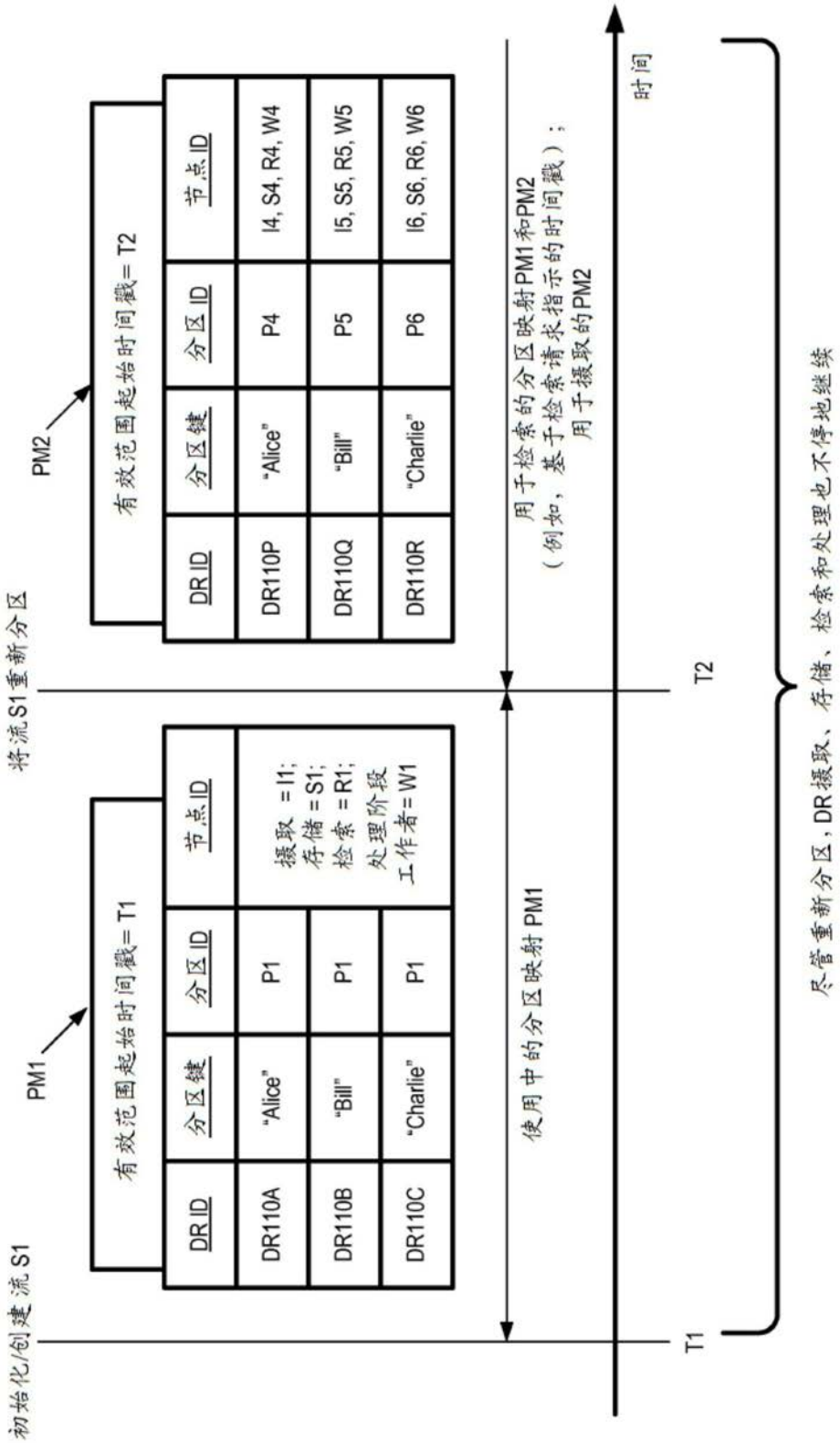

初始化/创建流 S1
PM1
有效范围起始时间戳= T1
DR ID
分区键
分区 ID
节点 ID
DR110A
"Alice"
P1
DR110B
"Bill"
P1
DR110C
"Charlie"
P1
摄取 = I1;
存储 = S1;
检索 = R1;
处理阶段
工作者= W1
将流 S1重新分区
PM2
有效范围起始时间戳= T2
DR ID
分区键
分区 ID
节点 ID
DR110P
"Alice"
P4
I4, S4, R4, W4
DR110Q
"Bill"
P5
I5, S5, R5, W5
DR110R
"Charlie"
P6
I6, S6, R6, W6
使用中的分区映射 PM1
用于检索的分区映射PM1和PM2
（例如，基于检索请求指示的时间戳）；
用于摄取的PM2
时间
T1
T2
尽管重新分区，DR摄取、存储、检索和处理也不停地继续

图16

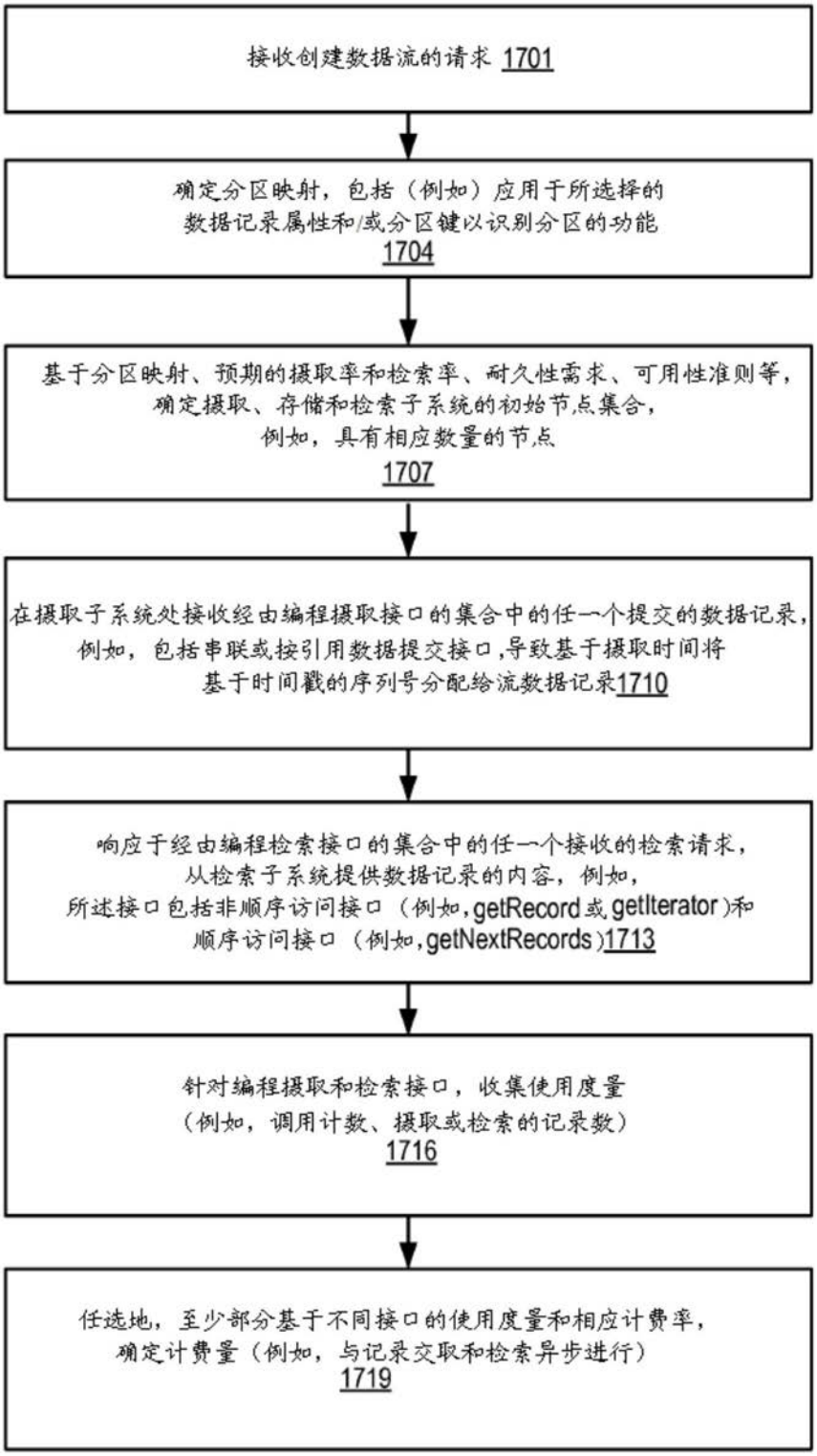

接收创建数据流的请求 1701
确定分区映射，包括（例如）应用于所选择的
数据记录属性和/或分区键以识别分区的功能
1704
基于分区映射、预期的摄取率和检索率、耐久性需求、可用性准则等，
确定摄取、存储和检索子系统的初始节点集合，
例如，具有相应数量的节点
1707
在摄取子系统处接收经由编程摄取接口的集合中的任一个提交的数据记录，
例如，包括串联或按引用数据提交接口，导致基于摄取时间将
基于时间戳的序列号分配给流数据记录1710
响应于经由编程检索接口的集合中的任一个接收的检索请求，
从检索子系统提供数据记录的内容，例如，
所述接口包括非顺序访问接口（例如，getRecord 或 getIterator）和
顺序访问接口（例如，getNextRecords）1713
针对编程摄取和检索接口，收集使用度量
（例如，调用计数、摄取或检索的记录数）
1716
任选地，至少部分基于不同接口的使用度量和相应计费率，
确定计费量（例如，与记录交取和检索异步进行）
1719

图17

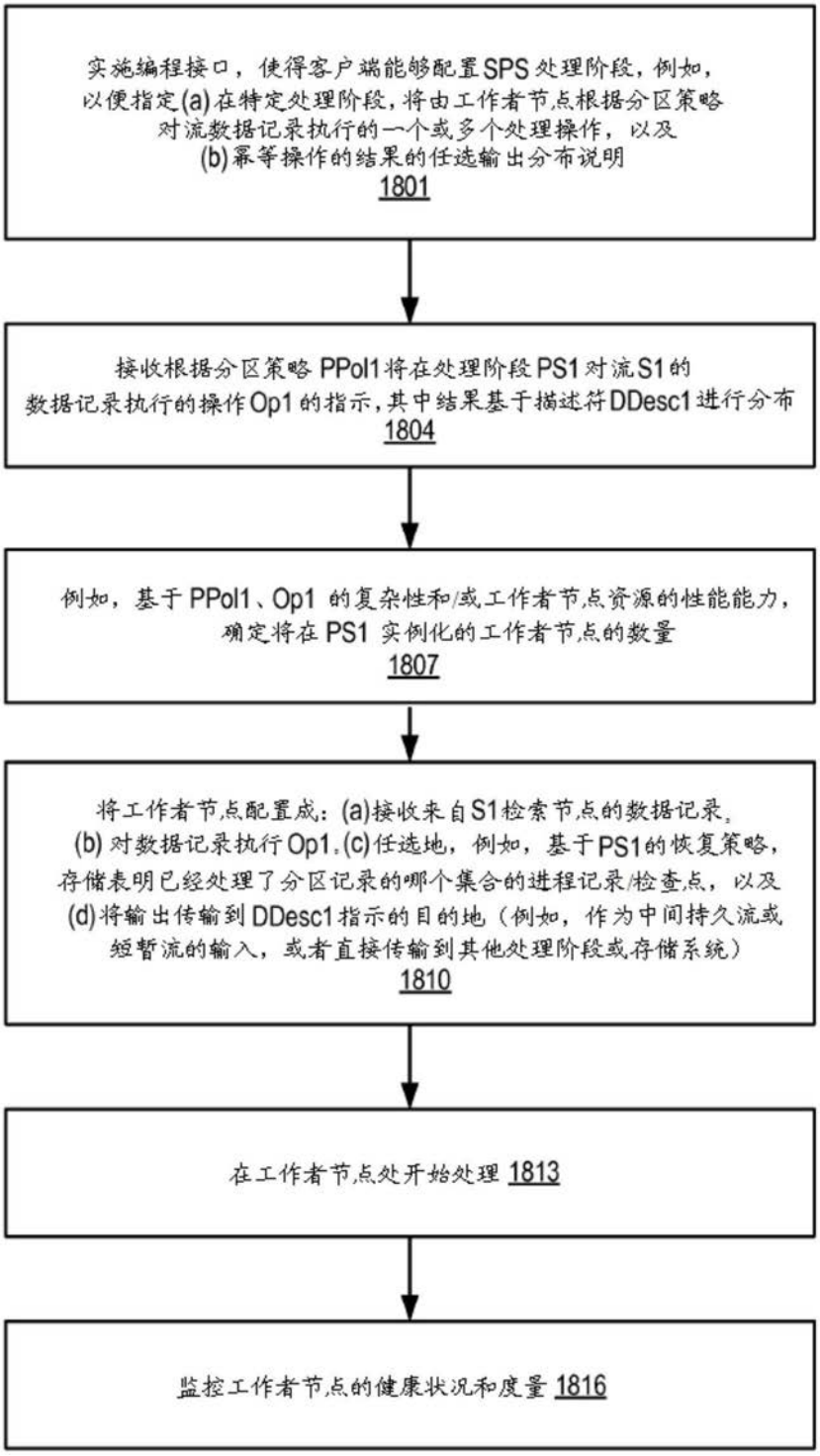

实施编程接口，使得客户端能够配置SPS处理阶段，例如，
以便指定(a)在特定处理阶段，将由工作者节点根据分区策略
对流数据记录执行的一个或多个处理操作，以及
(b)幂等操作的结果的任选输出分布说明
1801
接收根据分区策略PPol1将在处理阶段PS1对流S1的
数据记录执行的操作Op1的指示，其中结果基于描述符DDesc1进行分布
1804
例如，基于PPol1、Op1的复杂性和/或工作者节点资源的性能能力，
确定将在PS1实例化的工作者节点的数量
1807
将工作者节点配置成：(a)接收来自S1检索节点的数据记录，
(b)对数据记录执行Op1，(c)任选地，例如，基于PS1的恢复策略，
存储表明已经处理了分区记录的哪个集合的进程记录/检查点，以及
(d)将输出传输到DDesc1指示的目的地（例如，作为中间持久流或
短暂流的输入，或者直接传输到其他处理阶段或存储系统）
1810
在工作者节点处开始处理 1813
监控工作者节点的健康状况和度量 1816

图18a

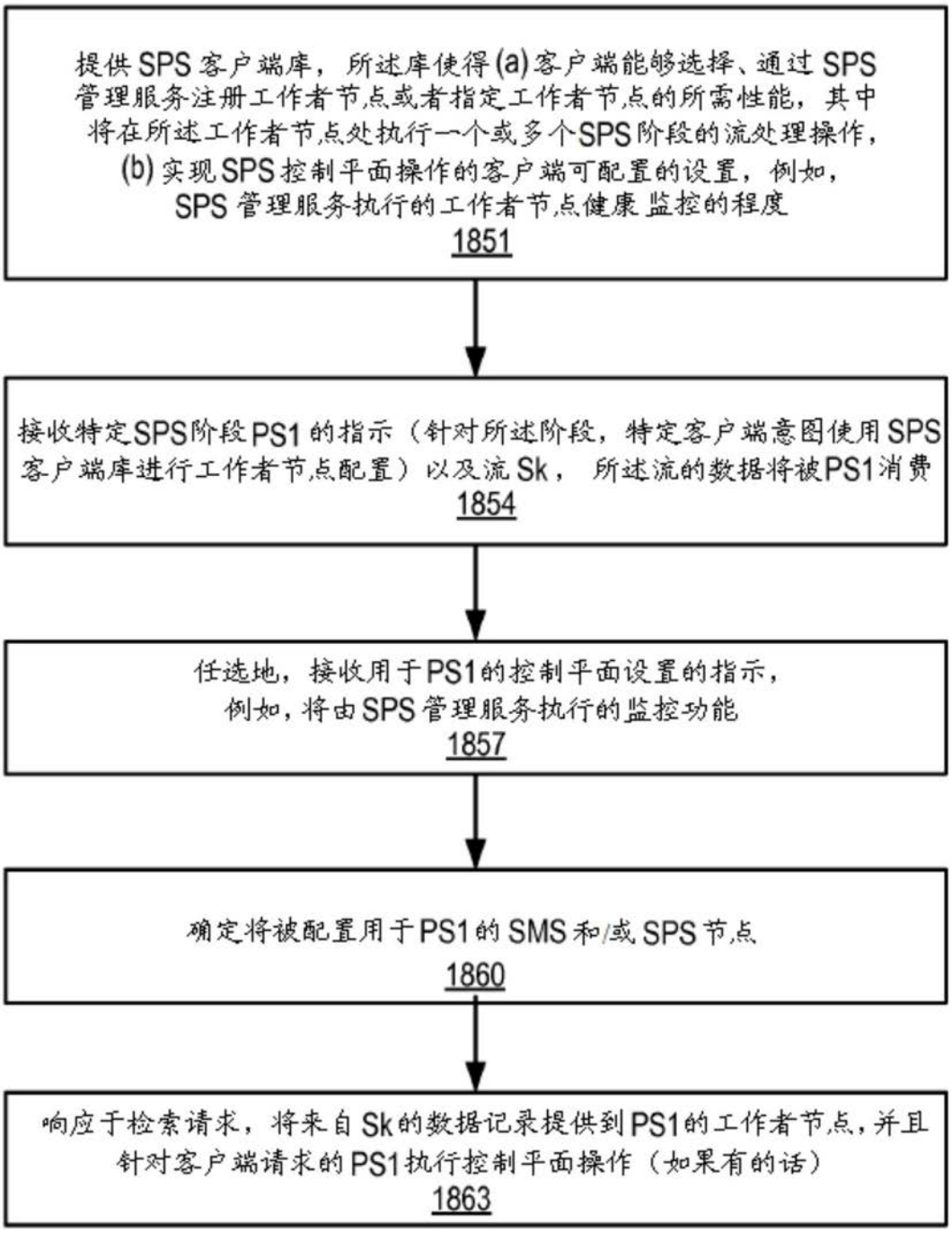

提供SPS客户端库，所述库使得(a)客户端能够选择、通过SPS管理服务注册工作者节点或者指定工作者节点的所需性能，其中将在所述工作者节点处执行一个或多个SPS阶段的流处理操作，(b)实现SPS控制平面操作的客户端可配置的设置，例如，SPS管理服务执行的工作者节点健康监控的程度
1851
接收特定SPS阶段PS1的指示（针对所述阶段，特定客户端意图使用SPS客户端库进行工作者节点配置）以及流Sk，所述流的数据将被PS1消费
1854
任选地，接收用于PS1的控制平面设置的指示，例如，将由SPS管理服务执行的监控功能
1857
确定将被配置用于PS1的SMS和/或SPS节点
1860
响应于检索请求，将来自Sk的数据记录提供到PS1的工作者节点，并且针对客户端请求的PS1执行控制平面操作（如果有的话）
1863

图18b

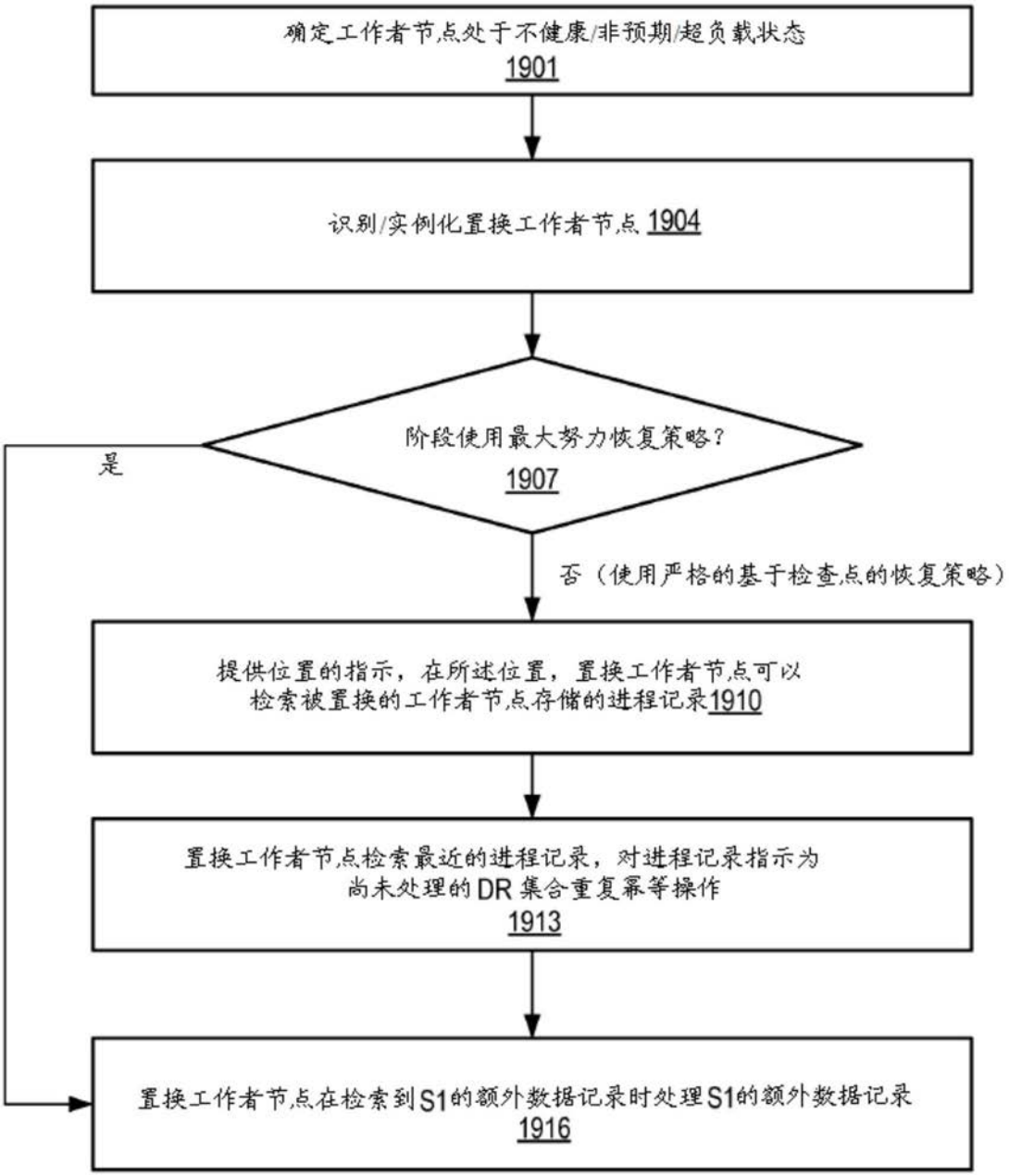

确定工作者节点处于不健康/非预期/超负载状态
1901
识别/实例化置换工作者节点 1904
阶段使用最大努力恢复策略?
1907
是
否（使用严格的基于检查点的恢复策略）
提供位置的指示，在所述位置，置换工作者节点可以检索被置换的工作者节点存储的进程记录1910
置换工作者节点检索最近的进程记录，对进程记录指示为尚未处理的 DR 集合重复幂等操作
1913
置换工作者节点在检索到S1的额外数据记录时处理S1的额外数据记录
1916

图19

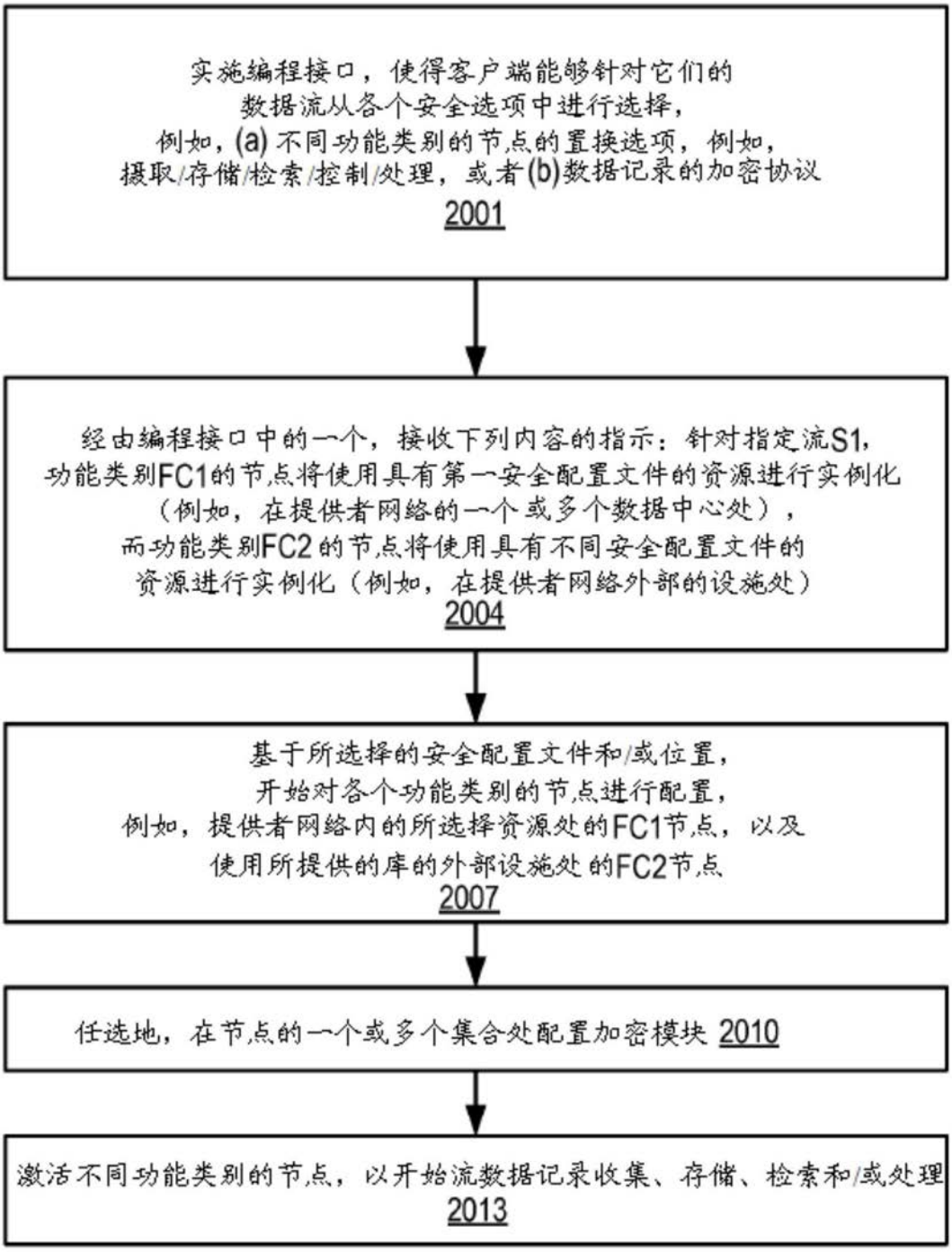

实施编程接口，使得客户端能够针对它们的
数据流从各个安全选项中进行选择，
例如，(a)不同功能类别的节点的置换选项，例如，
摄取/存储/检索/控制/处理，或者(b)数据记录的加密协议
2001
经由编程接口中的一个，接收下列内容的指示：针对指定流S1，
功能类别FC1的节点将使用具有第一安全配置文件的资源进行实例化
（例如，在提供者网络的一个或多个数据中心处），
而功能类别FC2的节点将使用具有不同安全配置文件的
资源进行实例化（例如，在提供者网络外部的设施处）
2004
基于所选择的安全配置文件和/或位置，
开始对各个功能类别的节点进行配置，
例如，提供者网络内的所选择资源处的FC1节点，以及
使用所提供的库的外部设施处的FC2节点
2007
任选地，在节点的一个或多个集合处配置加密模块 2010
激活不同功能类别的节点，以开始流数据记录收集、存储、检索和/或处理
2013

图20

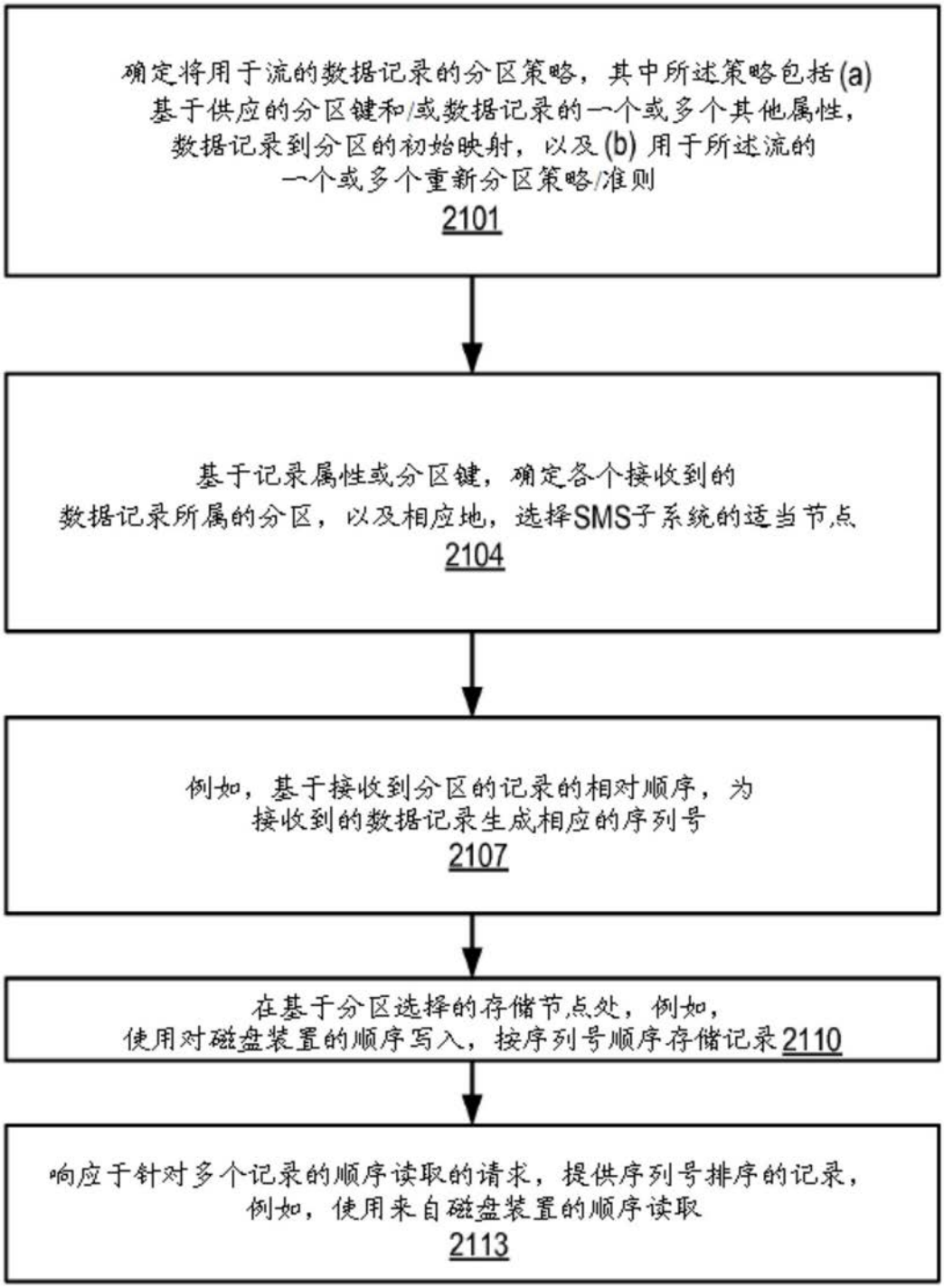

确定将用于流的数据记录的分区策略，其中所述策略包括(a)
基于供应的分区键和/或数据记录的一个或多个其他属性，
数据记录到分区的初始映射，以及(b)用于所述流的
一个或多个重新分区策略/准则
2101
基于记录属性或分区键，确定各个接收到的
数据记录所属的分区，以及相应地，选择SMS子系统的适当节点
2104
例如，基于接收到分区的记录的相对顺序，为
接收到的数据记录生成相应的序列号
2107
在基于分区选择的存储节点处，例如，
使用对磁盘装置的顺序写入，按序列号顺序存储记录2110
响应于针对多个记录的顺序读取的请求，提供序列号排序的记录，
例如，使用来自磁盘装置的顺序读取
2113

图21

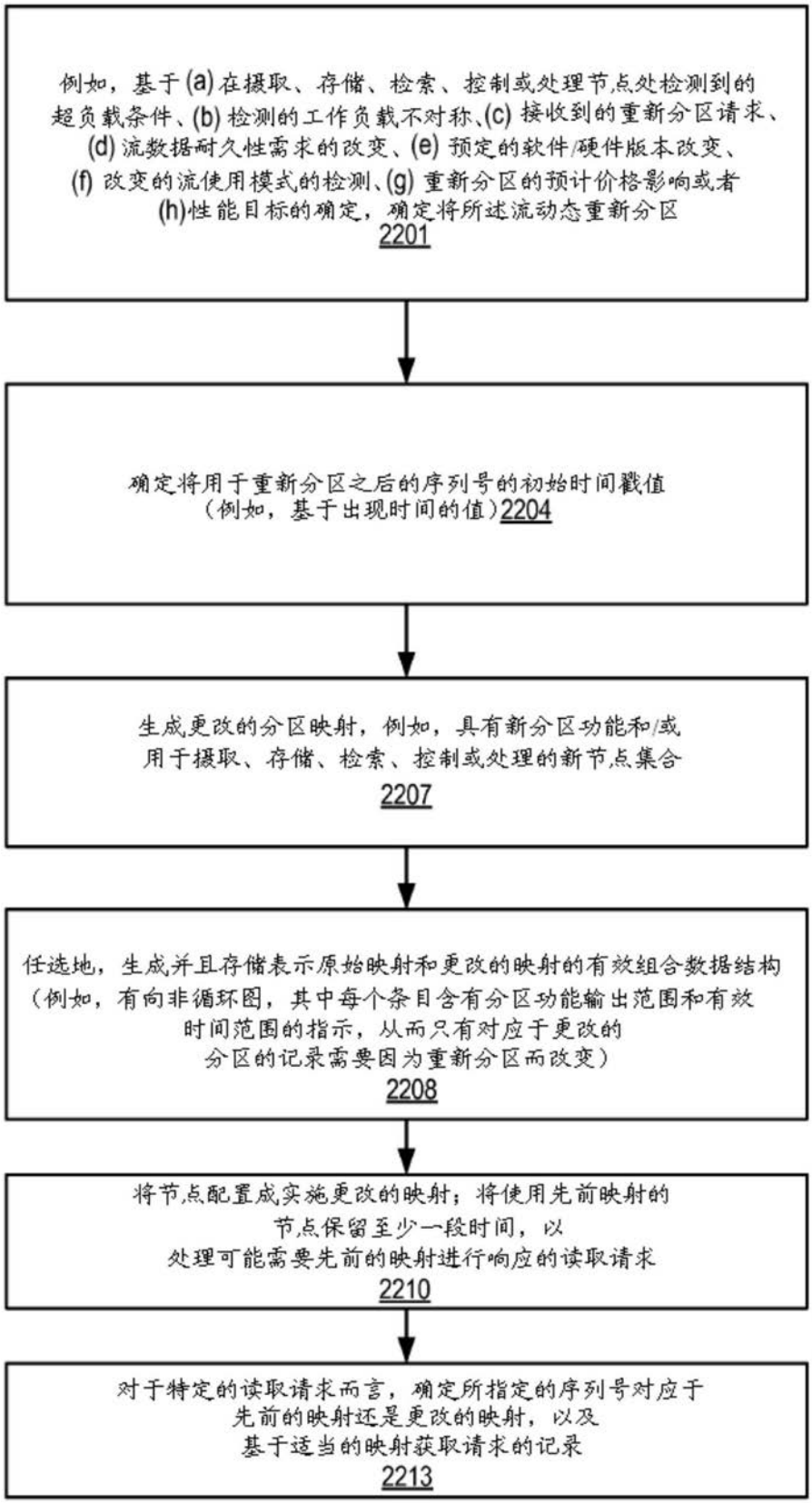

例如，基于(a)在摄取、存储、检索、控制或处理节点处检测到的超负载条件、(b)检测的工作负载不对称、(c)接收到的重新分区请求、(d)流数据耐久性需求的改变、(e)预定的软件/硬件版本改变、(f)改变的流使用模式的检测、(g)重新分区的预计价格影响或者(h)性能目标的确定，确定将所述流动态重新分区
2201
确定将用于重新分区之后的序列号的初始时间戳值（例如，基于出现时间的值）2204
生成更改的分区映射，例如，具有新分区功能和/或用于摄取、存储、检索、控制或处理的新节点集合
2207
任选地，生成并且存储表示原始映射和更改的映射的有效组合数据结构（例如，有向非循环图，其中每个条目含有分区功能输出范围和有效时间范围的指示，从而只有对应于更改的分区的记录需要因为重新分区而改变）
2208
将节点配置成实施更改的映射；将使用先前映射的节点保留至少一段时间，以处理可能需要先前的映射进行响应的读取请求
2210
对于特定的读取请求而言，确定所指定的序列号对应于先前的映射还是更改的映射，以及基于适当的映射获取请求的记录
2213

图22

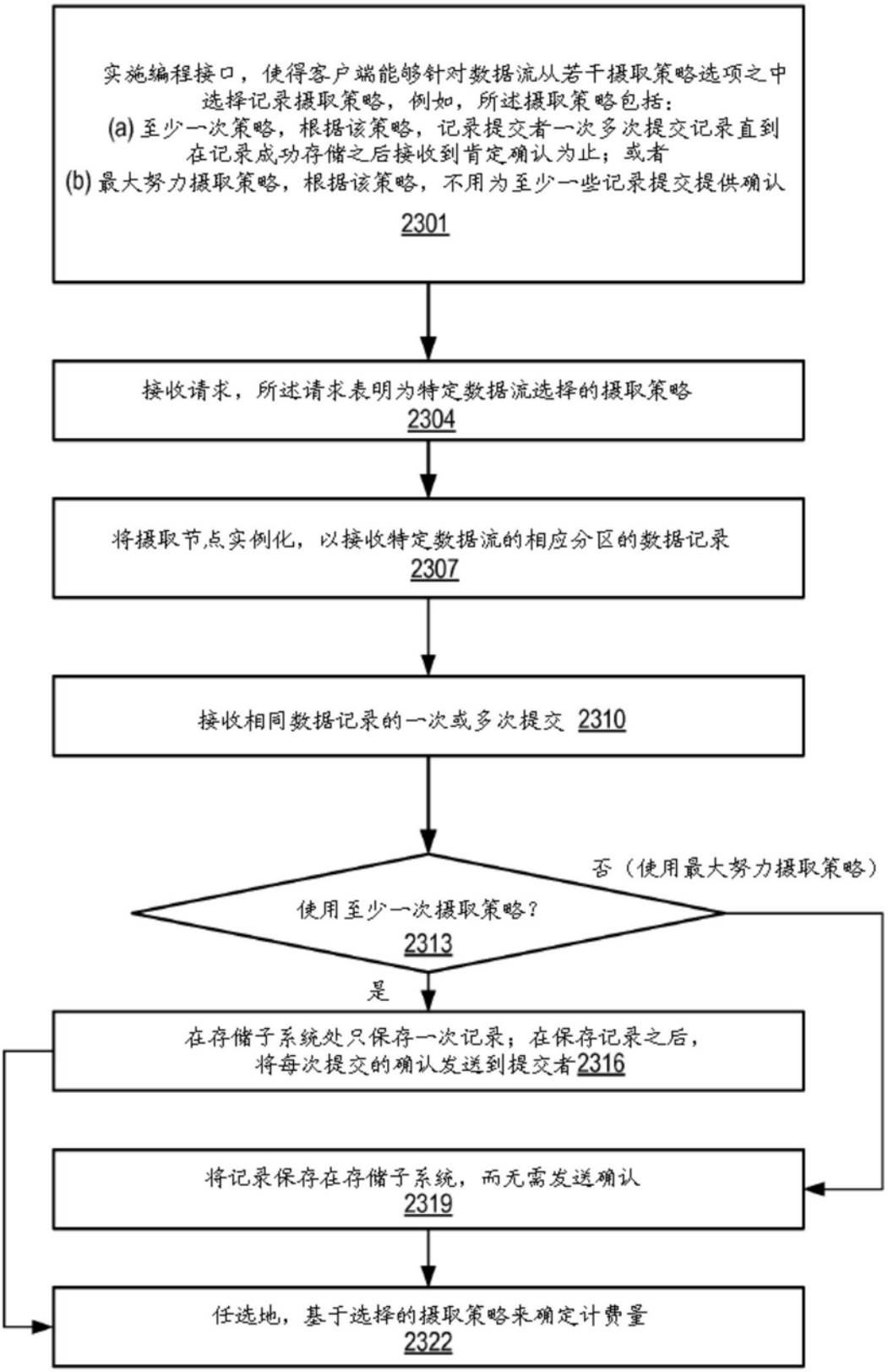

实施编程接口，使得客户端能够针对数据流从若干摄取策略选项之中选择记录摄取策略，例如，所述摄取策略包括：
(a) 至少一次策略，根据该策略，记录提交者一次多次提交记录直到在记录成功存储之后接收到肯定确认为止；或者
(b) 最大努力摄取策略，根据该策略，不用为至少一些记录提交提供确认
2301
接收请求，所述请求表明为特定数据流选择的摄取策略
2304
将摄取节点实例化，以接收特定数据流的相应分区的数据记录
2307
接收相同数据记录的一次或多次提交 2310
使用至少一次摄取策略？
2313
否（使用最大努力摄取策略）
是
在存储子系统处只保存一次记录；在保存记录之后，将每次提交的确认发送到提交者2316
将记录保存在存储子系统，而无需发送确认
2319
任选地，基于选择的摄取策略来确定计费量
2322

图23

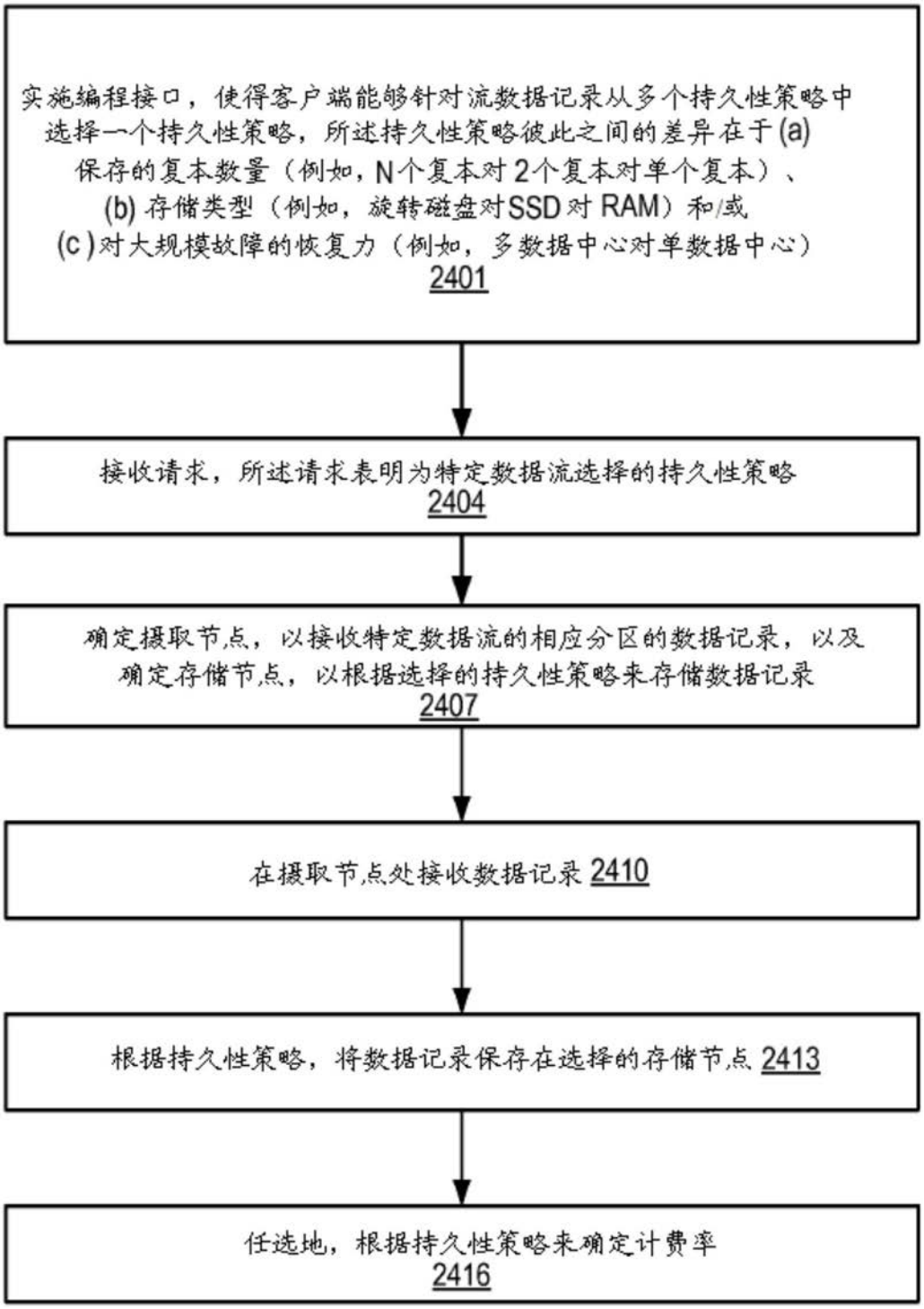

实施编程接口，使得客户端能够针对流数据记录从多个持久性策略中选择一个持久性策略，所述持久性策略彼此之间的差异在于(a)保存的复本数量（例如，N个复本对2个复本对单个复本）、(b)存储类型（例如，旋转磁盘对SSD对RAM）和/或(c)对大规模故障的恢复力（例如，多数据中心对单数据中心）
2401
接收请求，所述请求表明为特定数据流选择的持久性策略
2404
确定摄取节点，以接收特定数据流的相应分区的数据记录，以及确定存储节点，以根据选择的持久性策略来存储数据记录
2407
在摄取节点处接收数据记录 2410
根据持久性策略，将数据记录保存在选择的存储节点 2413
任选地，根据持久性策略来确定计费率
2416

图24

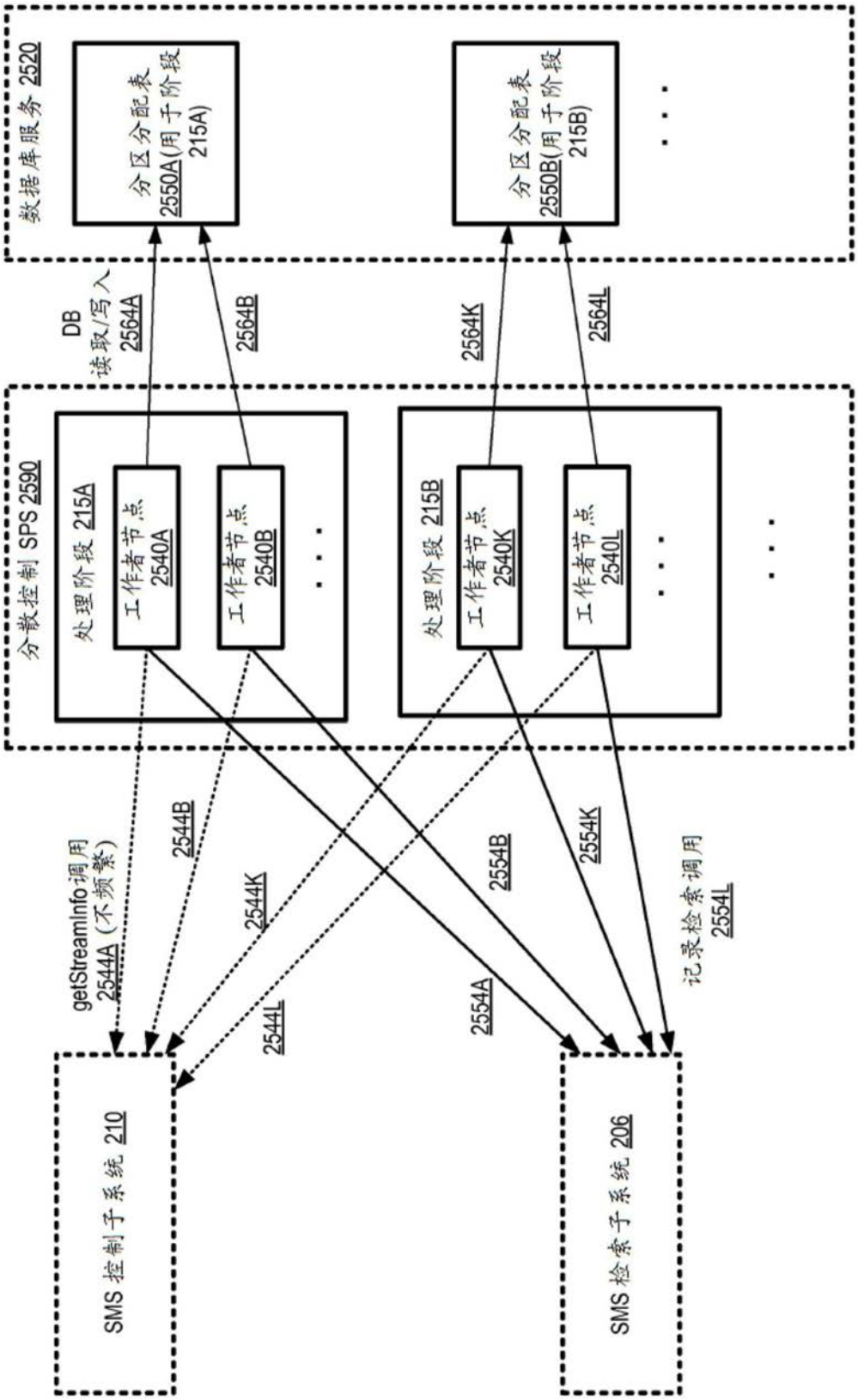

数据库服务 2520
分区分配表 2550A(用于阶段 215A)
分区分配表 2550B(用于阶段 215B)
DB 读取/写入 2564A
2564B
2564K
2564L
分散控制 SPS 2590
处理阶段 215A
工作者节点 2540A
工作者节点 2540B
处理阶段 215B
工作者节点 2540K
工作者节点 2540L
getStreamInfo调用 2544A (不频繁)
2544B
2544K
2544L
2554A
2554B
2554K
记录检索调用 2554L
SMS 控制子系统 210
SMS 检索子系统 206

图25

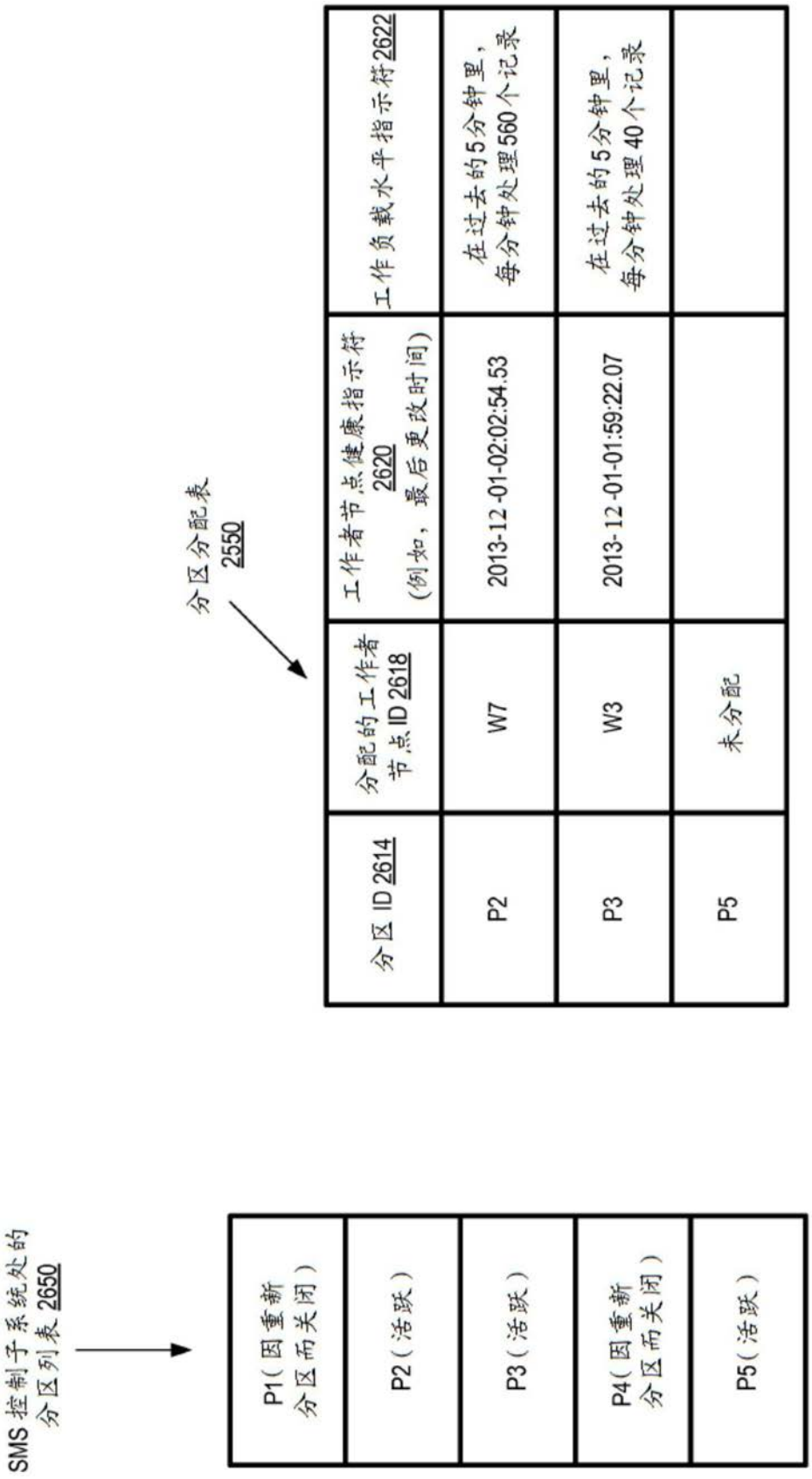

分区分配表
2550
分区 ID 2614
分配的工作者节点 ID 2618
工作者节点健康指示符 2620
(例如，最后更改时间)
工作负载水平指示符 2622
P2
W7
2013-12-01-02:02:54.53
在过去的5分钟里，每分钟处理560个记录
P3
W3
2013-12-01-01:59:22.07
在过去的5分钟里，每分钟处理40个记录
P5
未分配
SMS 控制子系统处的分区列表 2650
P1(因重新分区而关闭)
P2(活跃)
P3(活跃)
P4(因重新分区而关闭)
P5(活跃)

图26

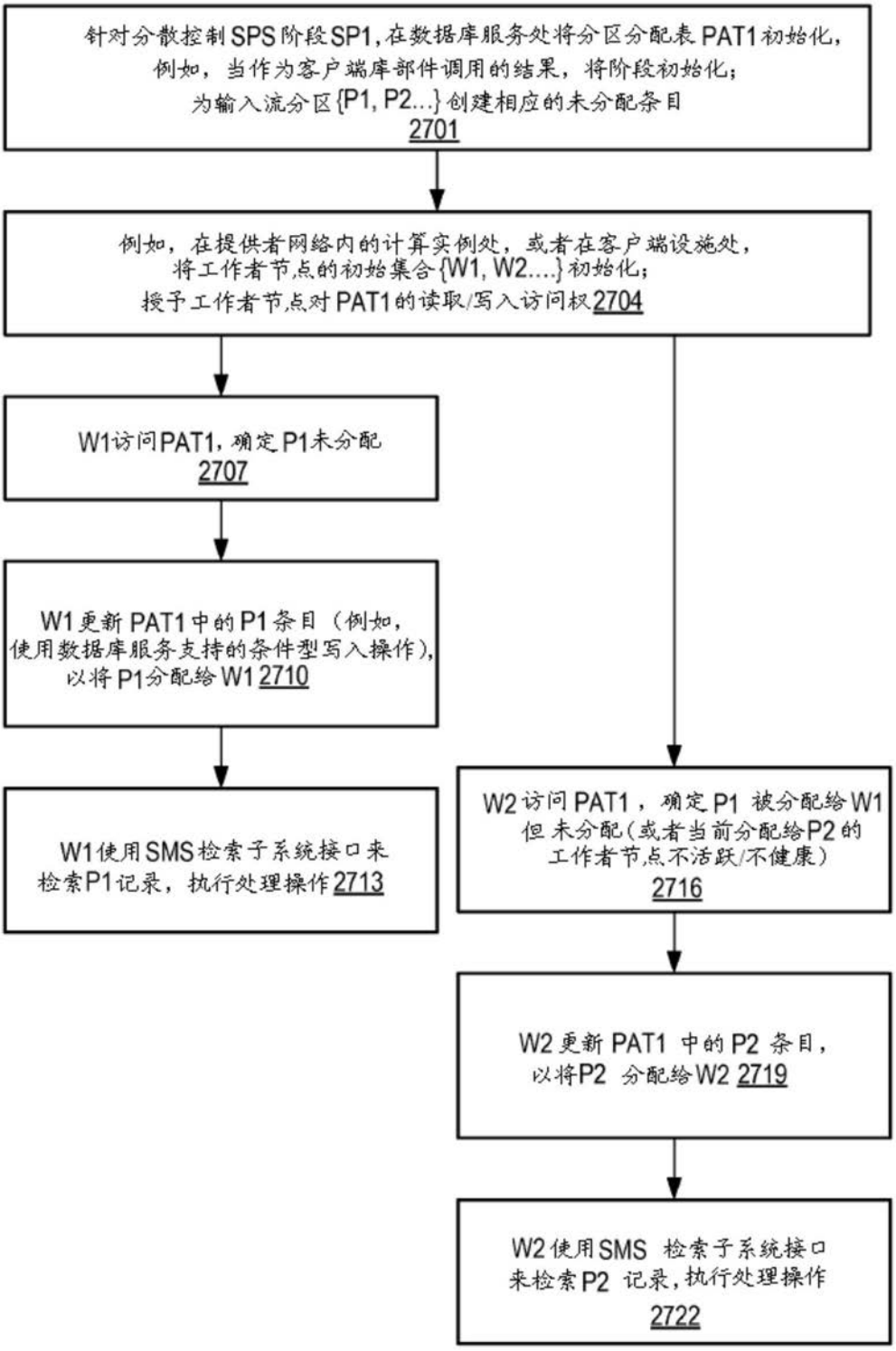

针对分散控制SPS阶段SP1，在数据库服务处将分区分配表PAT1初始化，
例如，当作为客户端库部件调用的结果，将阶段初始化；
为输入流分区{P1, P2...}创建相应的未分配条目
2701
例如，在提供者网络内的计算实例处，或者在客户端设施处，
将工作者节点的初始集合{W1, W2....}初始化；
授予工作者节点对PAT1的读取/写入访问权2704
W1访问PAT1，确定P1未分配
2707
W1更新PAT1中的P1条目（例如，
使用数据库服务支持的条件型写入操作），
以将P1分配给W1 2710
W1使用SMS检索子系统接口来
检索P1记录，执行处理操作2713
W2访问PAT1，确定P1被分配给W1
但未分配（或者当前分配给P2的
工作者节点不活跃/不健康）
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

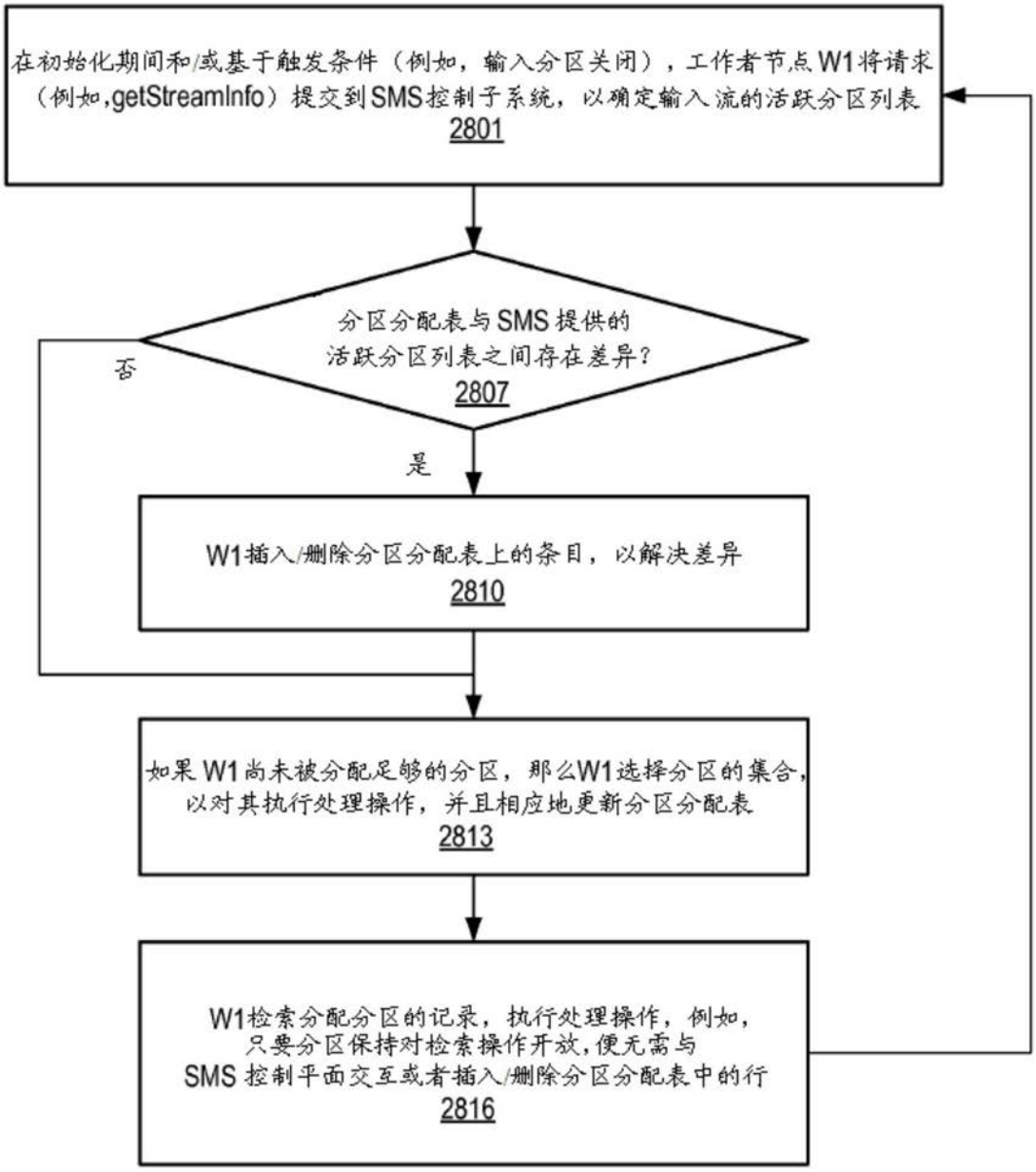

在初始化期间和/或基于触发条件（例如，输入分区关闭），工作者节点W1将请求（例如,getStreamInfo）提交到SMS控制子系统，以确定输入流的活跃分区列表
2801
分区分配表与SMS提供的活跃分区列表之间存在差异？
2807
否
是
W1插入/删除分区分配表上的条目，以解决差异
2810
如果W1尚未被分配足够的分区，那么W1选择分区的集合，以对其执行处理操作，并且相应地更新分区分配表
2813
W1检索分配分区的记录，执行处理操作，例如，只要分区保持对检索操作开放，便无需与SMS控制平面交互或者插入/删除分区分配表中的行
2816

图28

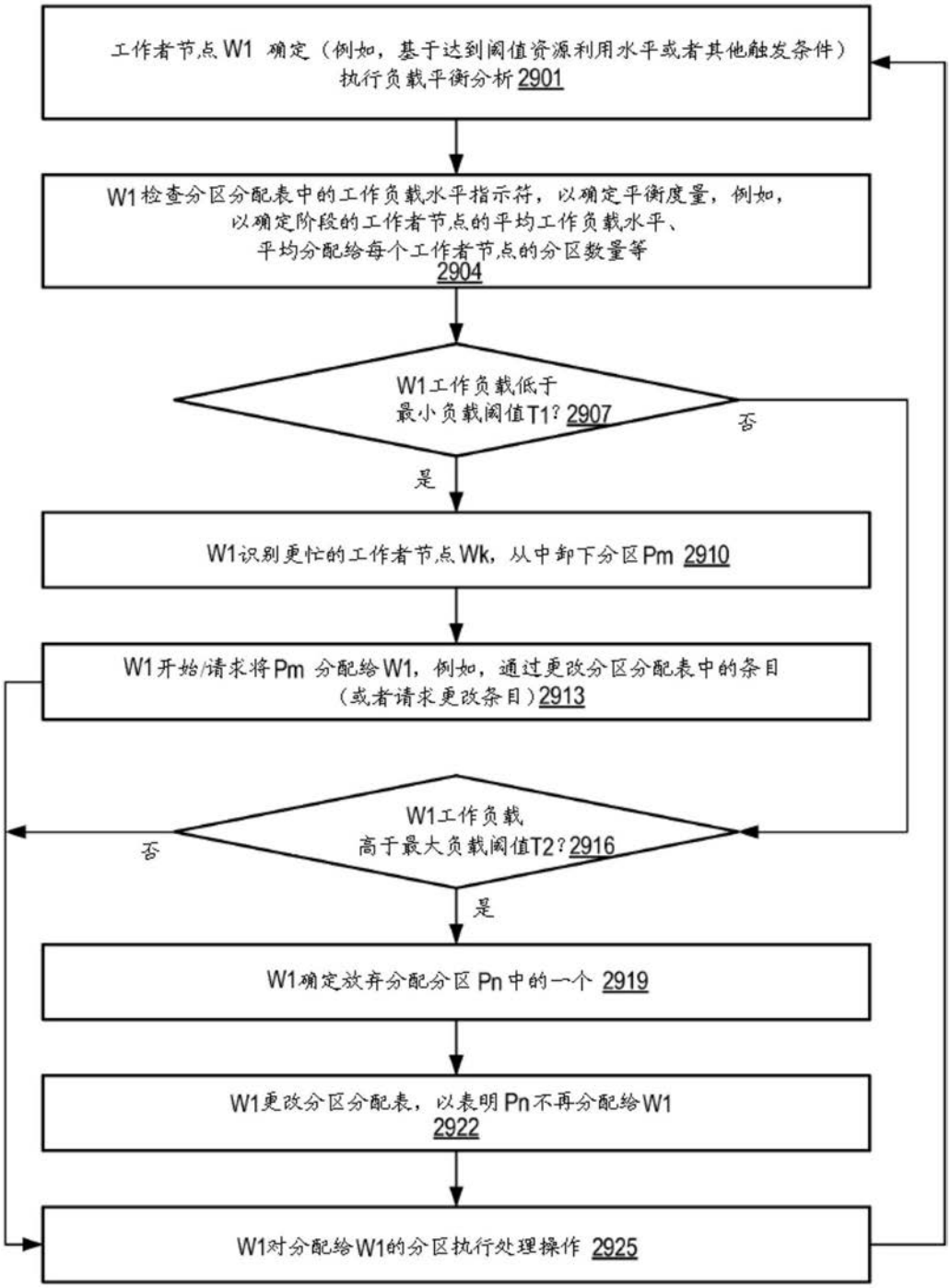

工作者节点W1 确定（例如，基于达到阈值资源利用水平或者其他触发条件）
执行负载平衡分析 2901
W1检查分区分配表中的工作负载水平指示符，以确定平衡度量，例如，
以确定阶段的工作者节点的平均工作负载水平、
平均分配给每个工作者节点的分区数量等
2904
W1工作负载低于
最小负载阈值T1？2907
否
是
W1识别更忙的工作者节点Wk，从中卸下分区Pm 2910
W1开始/请求将Pm 分配给W1，例如，通过更改分区分配表中的条目
（或者请求更改条目）2913
W1工作负载
高于最大负载阈值T2？2916
否
是
W1确定放弃分配分区 Pn 中的一个 2919
W1更改分区分配表，以表明Pn不再分配给W1
2922
W1对分配给W1的分区执行处理操作 2925

图29

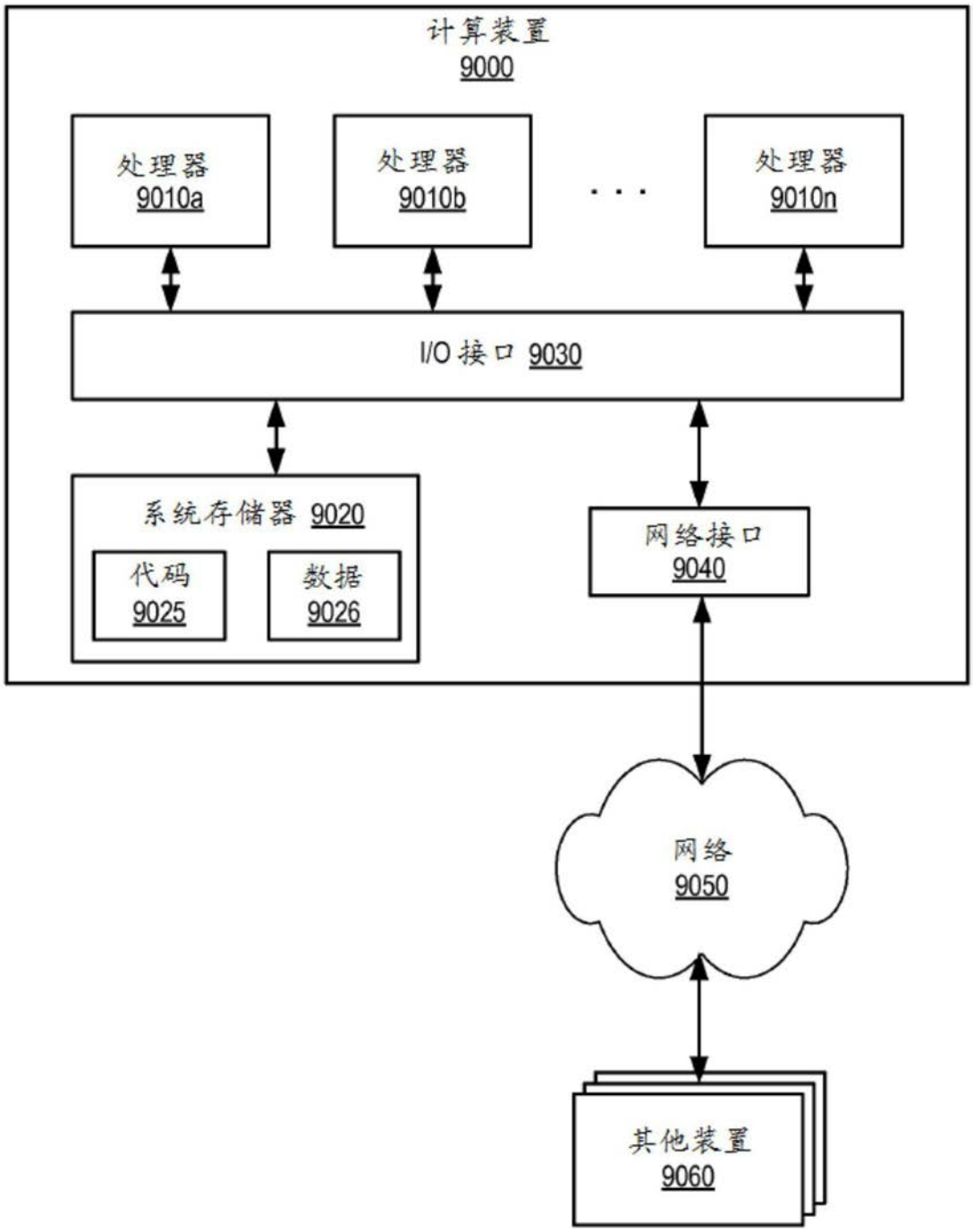

计算装置
9000
处理器
9010a
处理器
9010b
处理器
9010n
I/O 接口 9030
系统存储器 9020
代码
9025
数据
9026
网络接口
9040
网络
9050
其他装置
9060

图30