(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2020 年 11 月 26 日 (26.11.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/232870 A1

(51) 国际专利分类号:
G06F 9/52 (2006.01)

(21) 国际申请号: PCT/CN2019/102705

(22) 国际申请日: 2019 年 8 月 27 日 (27.08.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910428815.2 2019年5月20日 (20.05.2019) CN

(71) 申请人: 平安科技（深圳）有限公司 (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN]; 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

(72) 发明人: 袁康 (YUAN, Kang); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

(74) 代理人: 广州三环专利商标代理有限公司 (SCIHEAD IP LAW FIRM); 中国广东省广州市越秀区先烈中路80号汇华商贸大厦1508室, Guangdong 510070 (CN)。

(81) 指定国(除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,

(54) **Title:** ACTOR MODEL BASED DATA COLLECTION METHOD AND APPARATUS, AND STORAGE MEDIUM

(54) **发明名称:** 一种基于Actor模型的数据采集方法、装置及存储介质

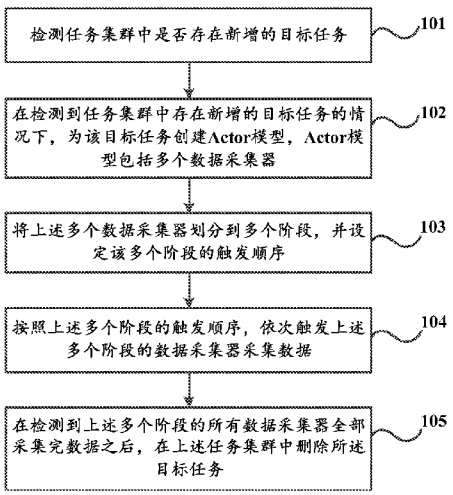


图 1

101 Detect whether a newly added target task exists in a task cluster
102 Where it is detected that a newly added target task exists in the task cluster, create an Actor model for the target task, wherein the Actor model comprises a plurality of data collectors
103 Divide the above-mentioned plurality of data collectors into a plurality of stages, and set a triggering sequence of the plurality of stages
104 According to the triggering sequence of the above-mentioned plurality of stages, sequentially trigger the data collectors at the above-mentioned plurality of stages to collect data
105 Delete the target task from the above-mentioned task cluster after it is detected that all the data collectors at the above-mentioned plurality of stages have collected the data

(57) **Abstract:** Disclosed are an Actor model based data collection method and apparatus, and a storage medium, which are applied to the field of data processing. The method comprises: detecting whether a newly added target task exists in a task cluster; where it is detected that a newly added target task exists in the task cluster, creating an Actor model for the target task, wherein the Actor model comprises a plurality of data collectors; dividing the plurality of data collectors into a plurality of stages, and setting a triggering sequence of the plurality of stages; according to the triggering sequence of the plurality of stages, sequentially triggering the data collectors at the plurality of stages to collect data; and deleting the target task from the task cluster after it is detected that the data collectors at the plurality of stages have all collected the data. In the present application, the plurality of data collectors are all Actors, and an execution sequence of different data collectors is set, so that the plurality of data collectors can be scheduled, without the need of a lock, to complete data collection.

WO 2020/232870 A1

[见续页]

SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要：本申请公开了一种基于Actor模型的数据采集方法、装置及存储介质，应用于数据处理领域，其中方法包括：检测任务集群中是否存在新增的目标任务；在检测到任务集群中存在新增的目标任务的情况下，为目标任务创建Actor模型，Actor模型包括多个数据采集器；将多个数据采集器划分到多个阶段，并设定多个阶段的触发顺序；按照多个阶段的触发顺序，依次触发多个阶段的数据采集器采集数据；在检测到多个阶段的数据采集器全部采集完数据之后，在任务集群中删除目标任务。本申请中的多个数据采集器都为Actor并且设定了不同数据采集器的执行顺序，使得不需要锁便能调度多个数据采集器完成数据采集。

一种基于 Actor 模型的数据采集方法、装置及存储介质

本申请要求于 2019 年 05 月 20 日提交中国专利局、申请号为 201910428815.2、申请名称为“一种基于 Actor 模型的数据采集方法、装置及存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及数据采集领域，尤其涉及一种基于 Actor 模型的数据采集方法、装置及存储介质。

背景技术

数据采集常常因为多线程或多进程并发访问某个临界资源而陷入死锁，这时可通过对该临界资源进行加锁，来建立线程或进程之间的互斥同步机制，以保障临界资源的完整性和一致性。单机的进程同步和线程同步有很多可用的互斥方案，比如互斥锁、信号量和条件变量等。但这些单机下的互斥方案在多机上是无法使用的，因为分布式系统的组件运行在不同的机器上，使得其不在统一的运行环境中。

分布式数据采集中，常常通过分布式锁解决对临界资源的互斥访问，比如 Redis 分布式锁和 Zookeeper 分布式锁的等。

但是分布式锁的实现比较复杂，且维护比较麻烦，需要较多维护成本，因此分布式锁不是一个解决分布式数据采集的死锁问题的理想方案，还是缺少了一种可以更高效的分布式数据采集方法。

发明内容

本申请实施例提供一种基于 Actor 模型数据采集方法，可以不使用锁便实现分布式数据采集。

第一方面，本申请实施例提供了一种基于 Actor 模型数据采集方法，该方法包括：

检测任务集群中是否存在新增的目标任务；

在检测到任务集群中存在新增的目标任务的情况下 ,为所述目标任务创建 Actor 模型，所述 Actor 模型包括多个数据采集器，所述数据采集器用于采集数据，所述多个数据采集器之间通过消息传递的方式通信；

将所述多个数据采集器划分到多个阶段，并设定所述多个阶段的触发顺序，每个阶段包含至少一个数据采集器，相同阶段的不同数据采集器位于不同的终端中，分别用于采集不同终端中的数据；

按照所述多个阶段的触发顺序，依次触发所述多个阶段的数据采集器采集数据；

在检测到所述多个阶段的所有数据采集器全部采集完数据之后，在所述任务集群中删除所述目标任务。

第二方面，本申请实施例提供了一种数据采集装置，该数据采集装置包括用于执行上述第一方面的数据采集方法的单元，该数据采集装置包括：

检测单元，用于检测任务集群中是否存在新增的目标任务；

创建单元，用于在检测到任务集群中存在新增的目标任务的情况下，为所述目标任务创建 Actor 模型，所述 Actor 模型包括多个数据采集器，所述数据采集器用于采集数据，所述多个数据采集器之间通过消息传递的方式通信；

划分单元，用于将所述多个数据采集器划分到多个阶段，并设定所述多个阶段的触发顺序，每个阶段包含至少一个数据采集器，相同阶段的不同数据采集器位于不同的终端中，分别用于采集不同终端中的数据；

执行单元，用于按照所述多个阶段的触发顺序，依次触发所述多个阶段的数据采集器采集数据；

删除单元，用于在检测到所述多个阶段的数据采集器全部采集完数据之后，在所述任务集群中删除所述目标任务。

第三方面，本申请实施例提供了另一种数据采集装置，包括处理器和存储器，所述处理器和存储器相互连接，其中，所述存储器用于存储支持数据采集装置执行上述数据采集方法的计算机程序，所述计算机程序包括程序指令，所述处理器被配置用于调用所述程序指令，用以执行：

检测任务集群中是否存在新增的目标任务；

在检测到任务集群中存在新增的目标任务的情况下，为所述目标任务创建 Actor 模型，所述 Actor 模型包括多个数据采集器，所述数据采集器用于采集数据，所述多个数据采集器之间通过消息传递的方式通信；

将所述多个数据采集器划分到多个阶段，并设定所述多个阶段的触发顺序，每个阶段包含至少一个数据采集器，相同阶段的不同数据采集器位于不同的终端中，分别用于采集不同终端中的数据；

按照所述多个阶段的触发顺序，依次触发所述多个阶段的数据采集器采集数据；

在检测到所述多个阶段的所有数据采集器全部采集完数据之后，在所述任务集群中删除所述目标任务。

第四方面，本申请实施例提供了一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述计算机程序包括程序指令，所述程序指令当被处理器执行，用以执行上述第一方面至第一方面的任意一种实现方式的数据采集方法。

本申请所提供的数据采集方法保证了数据采集过程中不会产生数据竞争和死锁，本申请不需要借助于锁便可以实现复杂的分布式数据采集，并且大大提高分布式数据采集的并发性能。

附图说明

为了更清楚地说明本申请实施例技术方案，下面将对实施例描述中所需要使用的附图

作简单地介绍。

图 1 是本申请实施例提供的一种数据采集方法的示意流程图；

图 2 是本申请另一实施例提供的一种数据采集方法的示意流程图；

图 3 是本申请实施例提供的一种数据采集装置的示意性框图；

图 4 是本申请实施例提供的一种数据采集装置的结构性框图。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述。

本申请主要应用于数据采集装置，该数据采集装置可以是传统数据采集装置、大型存储系统、台式电脑、笔记本电脑、平板电脑、掌上电脑、智能手机、便携式数字播放器、智能手表以及智能手环等等，本申请对此不做限制。

本申请实施例中描述的终端包括但不限于带通讯功能的设备、智能手机、平板电脑、笔记本电脑、台式电脑、便携式数字播放器、智能手环以及智能手表等。当终端向数据采集装置发送数据的时候，按照预设格式进行对数据的特性进行记录并传送，其中，数据的特性包括时间、地点、类型等。

参见图 1，是本申请实施例提供一种数据采集方法的示意流程图，如图 1 所示数据采集方法可包括：

101：检测任务集群中是否存在新增的目标任务。

在本申请实施例中，检测任务集群中是否存在新增的任务，若存在新增的任务，则将该新增的任务作为目标任务。其中，任务集群指的是待处理的任务的集合，且该任务集群中的不同任务可能由不同的终端所提出，而任务集群中的任务可能是由多个终端共同处理的分布式任务，也可能是仅由一个终端处理的单机任务，检测的过程指的是，本端的数据采集装置定时检测任务集群中是否存在新增的任务，或者任务集群中新增一个任务，本端装置便能接收到用于指示任务集群中新增任务的消息。

在另一种可实施的方式中，上述目标任务不是任务集群中新增的任意任务，而是任务集群中新增的分布式任务，具体的，上述检测任务集群中是否存在新增的目标任务指的是，检测任务集群中的新增的任务，若确定该新增的任务为分布式任务，则确定检测到任务集群中存在新增的目标任务。

在本申请实施例中，上述检测任务集群中新增的任务为分布式任务的过程指的是，获取上述任务集群中新增的任务的类型，若该新增的任务的类型为分布式任务，则表示该新增的任务的类型为分布式任务，或者，获取上述新增的任务的难度等级，若确定该新增的任务的难度等级超过预设级别，则说明该新增的任务需要占用较多的计算资源，并确定该新增的任务为分布式任务，也即是将该难度级别超过预设级别的任务作为目标任务，并接下来采用分布式任务的处理方法对其进行处理。

可见，本申请实施例只有在确定任务集群新增了分布式任务的时候，才确定任务集群

中有新增的目标任务，上述检测到的任务集群中新增的分布式任务即目标任务。

102：在检测到任务集群中存在新增的目标任务的情况下，为该目标任务创建 Actor 模型，Actor 模型包括多个数据采集器。

在本申请实施例中，Actor 模型为一个用于任务处理的模型，具体来说 Actor 模型是一个概念模型，用于处理并发计算，它定义了一系列系统组件应该如何动作和交互的通用规则，Actor 模型可以用于完成任意类型的任务，但在本申请实施例中构建的 Actor 模型用于执行数据采集。Actor 模型中包含了多个 Actor，Actor 在本申请实施例中被称为数据采集器，每个数据采集器是 Actor 模型中进行数据采集的最基本的单元，它能接收一个消息并且基于其执行数据采集。于是上述根据该目标任务的执行内容创建包含多个数据采集器的 Actor 模型指的是，对目标任务进行分解，然后基于分解之后的目标任务的采集任务，创建多个数据采集器来执行对应的采集任务，需要注意的是，数据采集器在采集数据时，多个数据采集器可能是并行执行的，但也不会造成资源竞争，因为数据采集器之间是通过消息传递的方式进行通信，于是不会造成资源竞争。

需要说明的是，由于本申请实施例采用的所有数据采集器都是 Actor，而 Actor 之间是通过消息传递的方式进行通信，因此数据采集器进行数据采集的时候，不会产生资源竞争的问题。其中，消息传递的方式指的是在消息传递系统中，进程间的数据交换是以格式化的消息(Message)为单位的，Actors 的一大重要特征在于 Actors 之间相互隔离，它们并不互相共享内存，一个 Actor 能维持一个私有的状态，并且这个状态不可能被另一个 Actor 所改变，而消息传递也保证了 Actor 之间的异步处理，因此 Actor 之间也不会存在数据竞争。若通信的进程之间不存在可直接访问的共享空间，则必须利用操作系统提供的消息传递方法实现进程通信，进程通过系统提供的发送消息和接收消息两个原语进行数据交换。简单来说，Actor 之间通过消息传递的方式来进行资源共享，当 Actor 需要另一个 Actor 的资源的时候，只需要将消息发送给另一个 Actor，但不关心另一个 Actor 什么时候处理，因此 Actor 之间的消息是异步传输的，类似邮件通信。具体的，在 Actor 模型里每个 Actor 都有地址，所以它们才能够相互发送消息，尽管许多 Actor 同时运行，但是一个 Actor 只能顺序地处理消息。也就是说其它三个 Actor 发送了给一个 Actor 分别发送了一条消息，这个 Actor 只能一次处理一条，如果需要并行处理三条消息，则需要三个 Actor 分别对这三条消息进行处理，并且如果 Actor 在处理消息的过程中又接收到一条新的消息，则该新的消息会被存储到 Actor 的邮箱 Mailbox 中，每个 Actor 对应有一个 Mailbox。

可见，Actor 通过消息传递的方式将不同 Actor 的可访问的资源进行了分隔，并保障了 Actor 之间的消息的异步处理，从而避免了使用锁来处理数据条件竞争(data race)的问题。消息传递机制最大的优点就是不会产生数据竞争状态 (data race)。

103：将上述多个数据采集器划分到多个阶段，并设定该多个阶段的触发顺序。

在本申请实施例中，将上述多个数据采集器划分到多个阶段，并设定该多个阶段的触发顺序，每个阶段包含至少一个数据采集器，相同阶段的不同数据采集器位于不同的终端中，分别用于采集不同终端中的数据。

需要说明的是，由于相同阶段的不同数据采集器分别分布于不同的终端上，于是本申请实施例应用于分布式数据采集，可见，由于位于不同的终端上的数据采集器在采集数据的时候，只会占用其所在终端的计算资源，于是在一个终端集群中的多个终端共同处理一个分布式采集任务，相比单机处理，分布式采集可以大大提升了计算能力和计算速度。需要注意的是，由于本申请实施例采用了 Actor 模型中的 Actor 来作为数据采集器，才可以不需要分布式锁便能实现分布式数据采集，这是因为 Actor 只关注接收到的消息的内容，而不关注消息的发送方是本地的 Actor 还是终端上的 Actor，于是只要保证多个 Actor 之间的消息能够顺利达到即可，也也允许了本申请实施例可以不及住锁便能在包含多个终端的终端集群中构建 Actor 模型，并实现分布式数据采集。

104：按照上述多个阶段的触发顺序，依次触发上述多个阶段的数据采集器采集数据。

在本申请实施例中，按照上述多个阶段的触发顺序，依次触发上述多个阶段的数据采集器采集数据。具体的，按照上述多个阶段的触发顺序，确定当前待触发阶段；若确定待触发阶段位于上述触发顺序的首位，或者确定接收到待触发阶段的前一阶段的所有数据采集器的完成采集的信息，则触发待触发阶段的所有数据采集器采集数据；在接收到待触发阶段的所有数据采集器的完成采集的信息的情况下，触发待触发阶段的下一阶段的所有数据采集器采集数据，直至接收到所述多个阶段的所有数据采集器的完成采集的信息。

举例来说，上述 Actor 模型包含三个阶段的数据采集器，按照执行顺序排序，该三个阶段分别为第一阶段、第二阶段和第三阶段，于是首先想第一阶段的所有数据采集器发送数据采集指令，使得该阶段的所有数据采集器并发执行数据采集，当第一阶段的所有数据采集器采集完数据之后，本端装置将会收到第一阶段的每个数据采集器发送的完成采集的信息，当接受到第一阶段的所有数据采集器的完成采集的信息之后，表示第一阶段的所有数据采集器完成数据采集，可以触发执行第二阶段的数据采集器，于是触发第二阶段的所有数据采集器并发执行数据采集，当接收到第二阶段的所有数据采集器的完成采集的信息之后，触发执行第三阶段的数据采集器执行数据采集，当接收到第三阶段的所有数据采集器的完成采集的信息之后，表示数据全部采集完成，即目标任务完成。

需要说明的是，本申请实施例虽然没有用 Actor 来作为调度节点来调度数据采集器进行数据采集，但是本申请通过将数据采集器划分到多个阶段，使得会产生数据竞争的数据采集器划分到不同的阶段，于是本申请即使不使用 Actor 作为调度节点，也可以实现分布式数据采集，并且也不会用到锁。

在另一种可实施的方式中，上一阶段的数据采集器采集到的数据可以传输给下一阶段的数据采集器，以供下一阶段的数据采集器使用。具体的，在上述触发待触发阶段的所有数据采集器采集数据之前，获取待触发阶段的前一阶段的所有数据采集器采集到的数据，并将待触发阶段的前一阶段的所有数据采集器采集到的数据保存到本地数据库中，然后接收到待触发阶段的目标数据采集器的数据请求指令，该数据请求指令用于请求所述本地数据库中的目标数据，于是在上述本地数据库中调取目标数据，并将该目标数据发送给目标数据采集器，最后在接收到待触发阶段的所有数据采集器的采集完成消息的情况下，将待

触发阶段的所有数据采集器采集到的数据保存到上述本地数据库中。简单来说，就是每当一个数据采集器完成了数据采集之后，便将该数据采集器采集到的 A 数据保存到本地数据库中，然后在下一阶段的数据采集器请求该 A 数据的时候，便从本地数据库中提取该 A 数据发送给上述下一阶段的数据采集器。

可见，不同阶段的数据采集器之间是可以进行数据交流的，上一阶段采集到的数据可以供下一阶段的数据采集器使用，使得下一阶段的数据采集器更好完成数据采集。于是不同阶段之间的数据交流有利于提高数据采集的效率。

在另一种可实施的方式中，在检测到宕机的数据采集器的情况下，删除该宕机的数据采集器，并获取该宕机的数据采集器的初始化信息，然后利用该初始化信息创建新的数据采集器，以替代所述宕机的数据数据采集器。

在本申请实施例中，在检测到数据采集器宕机的情况下，按照该宕机的数据采集器的初始化信息创建的新的数据采集器，然后将该新的数据采集器替换上述宕机的数据采集器，然后新的数据采集器可以从头开始采集数据，或者按照宕机的数据采集器之前采集到的信息，以及采集的进度，新的数据采集器按照该采集的进度，继续执行采集。

在另一种可实施的方式中，上述 Actor 模型还包括至少一个监控节点，监控节点用于初始化宕机的数据采集器，于是在检测到宕机的数据采集器的时候，利用监控节点对宕机的数据采集器进行初始化，以控制宕机的数据采集器恢复正常。

本申请实施例描述了另一种解决数据采集器宕机问题的方法，即如果数据采集器宕机，则使用该监控设备对宕机的数据采集器进行初始化。具体的，在检测到数据采集器宕机之后，使用监控节点根据数据采集器的初始化信息对宕机的数据采集器进行初始化，以控制宕机的数据采集器恢复正常。或者，监控节点用于将宕机的数据采集器恢复到宕机前一个时刻的正常状态，具体的，在检测到上述数据采集器宕机之后，获取记录在数据库中的该宕机的数据采集器在宕机前一个时刻的状态信息，并使用监控节点根据上述数据采集器在宕机前一个时刻的状态信息对数据采集器进行恢复，使得数据采集器恢复到上述状态信息对应的正常状态。

在另一种可实施的方式中，在为目标任务创建 Actor 模型的时候，根据目标任务的重要程度决定是否在上述 Actor 模型中创建监控节点。具体的，在任务集群中检测到目标任务之后，获取该目标任务的标签，该标签用于指示目标任务的重要等级，若目标任务的重要等级超过了预设等级，则在创建 Actor 模型的时候，创建上述监控节点，若目标任务的重要等级未超过预设等级，则不创建上述监控节点。

105：在检测到上述多个阶段的所有数据采集器全部采集完数据之后，在上述任务集群中删除所述目标任务。

在本申请实施例中，在检测到上述目标任务完成之后，便在上述任务集群中删除上述目标任务，并检测和执行任务集群中的下一个目标任务。

在另一种可实施的方式中，上述在任务集群中删除目标任务之后，检测已创建的数据采集器的个数，并获取所述已创建的数据采集器的创建时间；若已创建的数据采集器的个

数大于预设个数，则按照创建时间从早到晚的排序，依次删除已创建的数据采集器，直到剩下的已创建的数据采集器的个数小于或等于上述预设个数。

在本申请实施例中，系统中的数据采集器可能太多而导致本端装置存储空间不足或者系统运行缓慢，于是在完成目标任务之后，检测已创建的数据采集器的个数，如果已创建的数据采集器的个数超过预设个数，就对数据采集器进行删减，直到剩下的数据采集器的个数小于等于预设个数。其中，按照数据采集器的创建时间由早到晚顺序进行删除，这是因为早创建的数据采集器，相较于晚创建的数据采集器而言，被再次使用的几率更小。

上述删除已创建的数据采集器的详细过程为，检测已创建的数据采集器的个数，并获取该已创建的数据采集器的创建时间，若该已创建的数据采集器的个数大于预设个数，则确定该已创建的数据采集器中创建时间最早的时间节点，并删除该已创建的数据采集器中创建时间最早的数据采集器，然后若在删除该已创建的数据采集器中创建时间最晚的数据采集器之后，已创建的数据采集器的个数仍大于预设个数，则触发执行上述确定已创建的数据采集器中创建时间最早的数据采集器，至删除已创建的数据采集器中创建时间最早的数据采集器的步骤，直到已创建的数据采集器的个数在删除了创建时间最晚的数据采集器之后小于或等于预设个数。

本申请实施例每当检测到任务集群中存在新的目标任务，就为该目标任务建立包括多个数据采集器的 Actor 模型，并利用多个数据采集器来完成目标任务所指示的数据采集任务，具体的，将上述多个数据采集器划分到多个阶段，并设置该多个阶段的执行顺序，然后再按照该执行顺序，依次触发该多个阶段的数据采集器采集数据，最后在检测到所有阶段的所有数据采集器都采集完数据之后，就在任务集群中删除该目标任务，可见，本申请实施例所提供的数据采集方法中，由于多个数据采集器都是 Actor，而 Actor 是通过信息传递的方式进行通信的，信息传递的好处就是不会产生数据竞争的问题，在处理任务的时候无需同步阻塞等待，并且本申请实施例虽然不是通过 Actor 来调度上述多个数据采集器进行数据采集，但是本申请实施例通过将上述多个数据采集器划分到不同的阶段，从而设定了上述数据采集器的执行顺序，因此进一步保证了数据采集过程中不会产生数据竞争和死锁，于是本申请实施例不需要借助于锁便可以实现复杂的分布式数据采集，并且大大提高分布式数据采集的并发性能，可见，本申请实施例提供了一种高效的数据采集方法。

参见图 2，是本申请实施例提供另一种数据采集方法的示意流程图，如图 2 所示数据采集方法可包括：

201：检测任务集群中新增的任务。

在本申请实施例中，检测任务集群中是否存在新增的任务，具体的检测过程指的是，本端的数据采集装置定时检测任务集群中是否存在新增的任务，或者任务集群中新增一个任务，本端装置便能接收到用于指示任务集群中新增任务的消息。其中，任务集群指的是待处理的任务的集合，且该任务集群中的不同任务可能由不同的终端所提出，而任务集群中的任务可能是由多个终端共同处理的分布式任务，也可能是仅由一个终端处理的单机任

务。

202：若确定上述新增的任务为分布式任务，则确定检测到上述任务集群中存在新增的目标任务。

在本申请实施例中，在检测到任务集群中的新增的任务之后，进一步确定该新增的任务是否为目标任务。具体的，确定该新增的任务是否为分布式任务，若是则确定检测到任务集群中存在新增的目标任务。

在本申请实施例中，上述检测任务集群中新增的任务为分布式任务的过程指的是，获取上述任务集群中新增的任务的类型，若该新增的任务的类型为分布式任务，则表示该新增的任务的类型为分布式任务，或者，获取上述新增的任务的难度等级，若确定该新增的任务的难度等级超过预设级别，则说明该新增的任务需要占用较多的计算资源，并确定该新增的任务为分布式任务，也即是将该难度级别超过预设级别的任务作为目标任务，并接下来采用分布式任务的处理方法对其进行处理。

可见，本申请实施例只有在确定任务集群新增了分布式任务的时候，才确定任务集群中有新增的目标任务，上述检测到的任务集群中新增的分布式任务即目标任务。

203：在检测到任务集群中存在新增的目标任务的情况下，为该目标任务创建 Actor 模型，Actor 模型包括多个数据采集器。

在本申请实施例中，Actor 模型为一个用于任务处理的模型，具体来说 Actor 模型是一个概念模型，用于处理并发计算，它定义了一系列系统组件应该如何动作和交互的通用规则，Actor 模型可以用于完成任意类型的任务，但在本申请实施例中构建的 Actor 模型用于执行数据采集。Actor 模型中包含了多个 Actor，Actor 在本申请实施例中被称为数据采集器，每个数据采集器是 Actor 模型中进行数据采集的最基本的单元，它能接收一个消息并且基于其执行数据采集。于是上述根据该目标任务的执行内容创建包含多个数据采集器的 Actor 模型指的是，对目标任务进行分解，然后基于分解之后的目标任务的采集任务，创建多个数据采集器来执行对应的采集任务，需要注意的是，数据采集器在采集数据时，多个数据采集器可能是并行执行的，但也不会造成资源竞争，因为数据采集器之间是通过消息传递的方式进行通信，于是不会造成资源竞争。

需要说明的是，由于本申请实施例采用的所有数据采集器都是 Actor，而 Actor 之间是通过消息传递的方式进行通信，因此数据采集器进行数据采集的时候，不会产生资源竞争的问题。其中，消息传递的方式指的是在消息传递系统中，进程间的数据交换是以格式化的消息(Message)为单位的，Actors 的一大重要特征在于 Actors 之间相互隔离，它们并不互相共享内存，一个 Actor 能维持一个私有的状态，并且这个状态不可能被另一个 Actor 所改变，而消息传递也保证了 Actor 之间的异步处理，因此 Actor 之间也不会存在数据竞争。若通信的进程之间不存在可直接访问的共享空间，则必须利用操作系统提供的消息传递方法实现进程通信，进程通过系统提供的发送消息和接收消息两个原语进行数据交换。简单来说，Actor 之间通过消息传递的方式来进行资源共享，当 Actor 需要另一个 Actor 的资源的时候，只需要将消息发送给另一个 Actor，但不关心另一个 Actor 什么时候处理，因此 Actor

之间的消息是异步传输的，类似邮件通信。具体的，在 Actor 模型里每个 Actor 都有地址，所以它们才能够相互发送消息，尽管许多 Actor 同时运行，但是一个 Actor 只能顺序地处理消息。也就是说其它三个 Actor 发送了给一个 Actor 分别发送了一条消息，这个 Actor 只能一次处理一条，如果需要并行处理三条消息，则需要三个 Actor 分别对这三条消息进行处理，并且如果 Actor 在处理消息的过程中又接收到一条新的消息，则该新的消息会被存储到 Actor 的邮箱 Mailbox 中，每个 Actor 对应有一个 Mailbox。

可见，Actor 通过消息传递的方式将不同 Actor 的可访问的资源进行了分隔，并保障了 Actor 之间的消息的异步处理，从而避免了使用锁来处理数据条件竞争(data race)的问题。消息传递机制最大的优点就是不会产生数据竞争状态（data race）。

204：将上述多个数据采集器划分到多个阶段，并设定该多个阶段的触发顺序。

在本申请实施例中，将上述多个数据采集器划分到多个阶段，并设定该多个阶段的触发顺序，每个阶段包含至少一个数据采集器，相同阶段的不同数据采集器位于不同的终端中，分别用于采集不同终端中的数据。

需要说明的是，由于相同阶段的不同数据采集器分别分布于不同的终端上，于是本申请实施例应用于分布式数据采集，可见，由于位于不同的终端上的数据采集器在采集数据的时候，只会占用其所在终端的计算资源，于是在一个终端集群中的多个终端共同处理一个分布式采集任务，相比单机处理，分布式采集可以大大提升了计算能力和计算速度。需要注意的是，由于本申请实施例采用了 Actor 模型中的 Actor 来作为数据采集器，才可以不需要分布式锁便能实现分布式数据采集，这是因为 Actor 只关注接收到的消息的内容，而不关注消息的发送方是本地的 Actor 还是终端上的 Actor，于是只要保证多个 Actor 之间的消息能够顺利达到即可，也也允许了本申请实施例可以不及住锁便能在包含多个终端的终端集群中构建 Actor 模型，并实现分布式数据采集。

205：按照上述多个阶段的触发顺序，依次触发上述多个阶段的数据采集器采集数据。

在本申请实施例中，按照上述多个阶段的触发顺序，依次触发上述多个阶段的数据采集器采集数据。具体的，按照上述多个阶段的触发顺序，确定当前待触发阶段；若确定待触发阶段位于上述触发顺序的首位，或者确定接收到待触发阶段的前一阶段的所有数据采集器的完成采集的信息，则触发待触发阶段的所有数据采集器采集数据；在接收到待触发阶段的所有数据采集器的完成采集的信息的情况下，触发待触发阶段的下一阶段的所有数据采集器采集数据，直至接收到所述多个阶段的所有数据采集器的完成采集的信息。

举例来说，上述 Actor 模型包含三个阶段的数据采集器，按照执行顺序排序，该三个阶段分别为第一阶段、第二阶段和第三阶段，于是首先想第一阶段的所有数据采集器发送数据采集指令，使得该阶段的所有数据采集器并发执行数据采集，当第一阶段的所有数据采集器采集完数据之后，本端装置将会收到第一阶段的每个数据采集器发送的完成采集的信息，当接受到第一阶段的所有数据采集器的完成采集的信息之后，表示第一阶段的所有数据采集器完成数据采集，可以触发执行第二阶段的数据采集器，于是触发第二阶段的所有数据采集器并发执行数据采集，当接收到第二阶段的所有数据采集器的完成采集的信息

之后，触发执行第三阶段的数据采集器执行数据采集，当接收到第三阶段的所有数据采集器的完成采集的信息之后，表示数据全部采集完成，即目标任务完成。

需要说明的是，本申请实施例虽然没有用 Actor 来作为调度节点来调度数据采集器进行数据采集，但是本申请通过将数据采集器划分到多个阶段，使得会产生数据竞争的数据采集器划分到不同的阶段，于是本申请即使不使用 Actor 作为调度节点，也可以实现分布式数据采集，并且也不会用到锁。

在另一种可实施的方式中，上一阶段的数据采集器采集到的数据可以传输给下一阶段的数据采集器，以供下一阶段的数据采集器使用。具体的，在上述触发待触发阶段的所有数据采集器采集数据之前，获取待触发阶段的前一阶段的所有数据采集器采集到的数据，并将待触发阶段的前一阶段的所有数据采集器采集到的数据保存到本地数据库中，然后接收到待触发阶段的目标数据采集器的数据请求指令，该数据请求指令用于请求所述本地数据库中的目标数据，于是在上述本地数据库中调取目标数据，并将该目标数据发送给目标数据采集器，最后在接收到待触发阶段的所有数据采集器的采集完成消息的情况下，将待触发阶段的所有数据采集器采集到的数据保存到上述本地数据库中。简单来说，就是每当一个数据采集器完成了数据采集之后，便将该数据采集器采集到的 A 数据保存到本地数据库中，然后在下一阶段的数据采集器请求该 A 数据的时候，便从本地数据库中提取该 A 数据发送给上述下一阶段的数据采集器。

可见，不同阶段的数据采集器之间是可以进行数据交流的，上一阶段采集到的数据可以供下一阶段的数据采集器使用，使得下一阶段的数据采集器更好完成数据采集。于是不同阶段之间的数据交流有利于提高数据采集的效率。

在另一种可实施的方式中，在检测到宕机的数据采集器的情况下，删除该宕机的数据采集器，并获取该宕机的数据采集器的初始化信息，然后利用该初始化信息创建新的数据采集器，以替代所述宕机的数据数据采集器。

在本申请实施例中，在检测到数据采集器宕机的情况下，按照该宕机的数据采集器的初始化信息创建的新的数据采集器，然后将该新的数据采集器替换上述宕机的数据采集器，然后新的数据采集器可以从头开始采集数据，或者按照宕机的数据采集器之前采集到的信息，以及采集的进度，新的数据采集器按照该采集的进度，继续执行采集。

在另一种可实施的方式中，上述 Actor 模型还包括至少一个监控节点，监控节点用于初始化宕机的数据采集器，于是在检测到宕机的数据采集器的时候，利用监控节点对宕机的数据采集器进行初始化，以控制宕机的数据采集器恢复正常。

本申请实施例描述了另一种解决数据采集器宕机问题的方法，即如果数据采集器宕机，则使用该监控设备对宕机的数据采集器进行初始化。具体的，在检测到数据采集器宕机之后，使用监控节点根据数据采集器的初始化信息对宕机的数据采集器进行初始化，以控制宕机的数据采集器恢复正常。或者，监控节点用于将宕机的数据采集器恢复到宕机前一个时刻的正常状态，具体的，在检测到上述数据采集器宕机之后，获取记录在数据库中的该宕机的数据采集器在宕机前一个时刻的状态信息，并使用监控节点根据上述数据采集器在

宕机前一个时刻的状态信息对数据采集器进行恢复，使得数据采集器恢复到上述状态信息对应的正常状态。

在另一种可实施的方式中，在为目标任务创建 Actor 模型的时候，根据目标任务的重要程度决定是否在上述 Actor 模型中创建监控节点。具体的，在任务集群中检测到目标任务之后，获取该目标任务的标签，该标签用于指示目标任务的重要等级，若目标任务的重要等级超过了预设等级，则在创建 Actor 模型的时候，创建上述监控节点，若目标任务的重要等级超过了预设等级，则不创建上述监控节点。

206：在检测到上述多个阶段的所有数据采集器全部采集完数据之后，在上述任务集群中删除所述目标任务。

在本申请实施例中，由于本端通过检测所有数据采集器是否都执行完数据采集而可以感知到目标任务是否完成，于是可以检测到上述目标任务是否已完成，若在检测到上述目标任务完成，则在上述任务集群中删除上述目标任务，并检测和执行任务集群中的下一个目标任务。

207：检测已创建的数据采集器的个数，并获取该已创建的数据采集器的创建时间。

在本申请实施例中，在上述目标任务完成，检测上述已创建的数据采集器，统计数据采集器的个数，以及每个已创建的数据采集器的创建时间。基于创建时间从早到晚的顺序，对已创建的数据采集器进行排序。

208：若上述已创建的数据采集器的个数大于预设个数，则按照创建时间从早到晚的排序，依次删除上述已创建的数据采集器，直到剩下的已创建的数据采集器的个数小于或等于预设个数。

在本申请实施例中，系统中的数据采集器可能太多而导致本端装置存储空间不足或者系统运行缓慢，于是在完成目标任务之后，检测已创建的数据采集器的个数，如果已创建的数据采集器的个数超过预设个数，就对数据采集器进行删减，直到剩下的数据采集器的个数小于等于预设个数。其中，按照数据采集器的创建时间由早到晚顺序进行删除，这是因为早创建的数据采集器，相较于晚创建的数据采集器而言，被再次使用的几率更小。

上述删除已创建的数据采集器的详细过程为，检测已创建的数据采集器的个数，并获取该已创建的数据采集器的创建时间，若该已创建的数据采集器的个数大于预设个数，则确定该已创建的数据采集器中创建时间最早的时间节点，并删除该已创建的数据采集器中创建时间最早的数据采集器，然后若在删除该已创建的数据采集器中创建时间最晚的数据采集器之后，已创建的数据采集器的个数仍大于预设个数，则触发执行上述确定已创建的数据采集器中创建时间最早的数据采集器，至删除已创建的数据采集器中创建时间最早的数据采集器的步骤，直到已创建的数据采集器的个数在删除了创建时间最晚的数据采集器之后小于或等于预设个数。

本申请实施例相对于上一申请实施例来说更加的详细，且详细描述了在只有当在任务集群中新增的任务是分布式任务才能执行上述分布式数据采集方法，此外，还详细描述了在目标任务完成之后，若已创建的数据采集器超过预设个数，就对多余的数据采集器进行

清除，清除创建时间较早的数据采集器，使得删除之后已创建的数据采集器的个数小于等于预设个数。因此本申请实施例一方面只适用于分布式任务，另一方面，本申请实施例通过在目标任务完成之后及时得对多余的数据采集器进行清除，来保障了本装置的系统的顺畅运行，并且保障了下一个目标任务的处理。总的来说数据采集的效率进一步被提高了。

需要说明的是，上文对各个实施例的描述倾向于强调各个实施例之间的不同之处，其相同或相似之处可以互相参考，为了简洁，本文不再赘述。

本申请实施例还提供一种数据采集装置，该数据采集装置用于执行前述任一项的数据采集方法的单元。具体地，参见图 3，是本申请实施例提供的一种数据采集装置的示意框图。本实施例的数据采集装置包括：检测单元 301、创建单元 302、划分单元 303、执行单元 304 以及删除单元 305。具体的：

检测单元 301，用于检测任务集群中是否存在新增的目标任务；

创建单元 302，用于在检测到任务集群中存在新增的目标任务的情况下，为上述目标任务创建 Actor 模型，上述 Actor 模型包括多个数据采集器，上述数据采集器用于采集数据，上述多个数据采集器之间通过消息传递的方式通信；

划分单元 303，用于将上述多个数据采集器划分到多个阶段，并设定上述多个阶段的触发顺序，每个阶段包含至少一个数据采集器，相同阶段的不同数据采集器位于不同的终端中，分别用于采集不同终端中的数据；

执行单元 304，用于按照上述多个阶段的触发顺序，依次触发上述多个阶段的数据采集器采集数据；

删除单元 305，用于在检测到上述多个阶段的数据采集器全部采集完数据之后，在上述任务集群中删除上述目标任务。

在另一种可实施的方式中，上述数据采集装置还包括确定单元 306，用于按照上述多个阶段的触发顺序，确定当前待触发阶段；上述执行单元 304，具体用于若确定上述待触发阶段位于上述触发顺序的首位，或者确定接收到上述待触发阶段的前一阶段的所有数据采集器的完成采集的信息，则触发上述待触发阶段的所有数据采集器采集数据；上述执行单元 304，还用于在接收到上述待触发阶段的所有数据采集器的完成采集的信息的情况下，触发上述待触发阶段的下一阶段的所有数据采集器采集数据，直至接收到上述多个阶段的所有数据采集器的完成采集的信息。

在另一种可实施的方式中，上述数据采集装置还包括获取单元 307，用于在上述触发上述待触发阶段的所有数据采集器采集数据之前，获取上述待触发阶段的前一阶段的所有数据采集器采集到的数据；上述数据采集装置还包括保存单元 308，用于将上述待触发阶段的前一阶段的所有数据采集器采集到的数据保存到本地数据库中；上述数据采集装置还包括接收单元 309，用于接收到上述待触发阶段的目标数据采集器的数据请求指令，上述数据请求指令用于请求上述本地数据库中的目标数据；上述获取单元 307，还用于上述在上述本地数据库中调取上述目标数据；上述数据采集装置还包括发送单元 310，用于将上

述目标数据发送给上述目标数据采集器；上述保存单元 308 还用于在接收到上述待触发阶段的所有数据采集器的采集完成消息的情况下，将上述待触发阶段的所有数据采集器采集到的数据保存到上述本地数据库中。

在另一种可实施的方式中，上述检测单元 301，还用于检测上述任务集群中新增的任务；上述确定单元 306，还用于确定上述新增的任务为分布式任务，以及确定检测到上述任务集群中存在新增的目标任务。

在另一种可实施的方式中，上述检测单元 301，还用于检测已创建的数据采集器的个数；上述获取单元 307，还用于获取上述已创建的数据采集器的创建时间；上述删除单元 305，还用于若上述已创建的数据采集器的个数大于预设个数，则按照上述创建时间从早到晚的排序，依次删除上述已创建的数据采集器，直到剩下的已创建的数据采集器的个数小于或等于上述预设个数。

在另一种可实施的方式中，上述检测单元 301 还用于检测宕机的数据采集器；上述删除单元 305，还用于删除上述宕机的数据采集器；上述获取单元 307，还用于获取上述宕机的数据采集器的初始化信息；上述数据采集装置还包括替换单元 311，利用上述初始化信息创建新的数据采集器，以替代上述宕机的数据数据采集器。

在另一种可实施的方式中，上述 Actor 模型还包括至少一个监控节点，上述监控节点用于初始化宕机的数据采集器，上述检测单元 301，还用于检测宕机的数据采集器；上述数据采集装置还包括恢复单元 312，用于利用上述监控节点对上述宕机的数据采集器进行初始化，以控制上述宕机的数据采集器恢复正常。

本申请实施例每当检测单元检测到任务集群中存在新的目标任务，创建单元就为该目标任务建立包括多个数据采集器的 Actor 模型，并执行单元利用多个数据采集器来完成目标任务所指示的数据采集任务，具体的，划分单元将上述多个数据采集器划分到多个阶段，并设置该多个阶段的执行顺序，然后执行单元再按照该执行顺序，依次触发该多个阶段的数据采集器采集数据，最后在检测单元检测到所有阶段的所有数据采集器都采集完数据之后，删除单元就在任务集群中删除该目标任务，可见，本申请实施例所提供的数据采集方法中，由于多个数据采集器都是 Actor，而 Actor 是通过信息传递的方式进行通信的，信息传递的好处就是不会产生数据竞争的问题，在处理任务的时候无需同步阻塞等待，并且本申请实施例虽然不是通过 Actor 来调度上述多个数据采集器进行数据采集，但是本申请实施例通过将上述多个数据采集器划分到不同的阶段，从而设定了上述数据采集器的执行顺序，因此进一步保证了数据采集过程中不会产生数据竞争和死锁，于是本申请实施例不需要借助于锁便可以实现复杂的分布式数据采集，并且大大提高分布式数据采集的并发性能，可见，本申请实施例提供了一种高效的数据采集方法。

参见图 4，是本申请另一实施例提供的一种数据采集装置示意框图。如图所示的本实施例中的数据采集装置可以包括：一个或多个处理器 410 和存储器 420。上述处理器 410 和存储器 420 通过总线 430 连接。存储器 420 用于存储计算机程序，计算机程序包括程序

指令，处理器 410 用于执行存储器 420 存储的程序指令。具体的：

处理器 410，用于执行检测单元 301 的功能，用于检测任务集群中是否存在新增的目标任务；还用于执行创建单元 302 的功能，用于在检测到任务集群中存在新增的目标任务的情况下，为上述目标任务创建 Actor 模型，上述 Actor 模型包括多个数据采集器，上述数据采集器用于采集数据，上述多个数据采集器之间通过消息传递的方式通信；还用执行划分单元 303 的功能，用于将上述多个数据采集器划分到多个阶段，并设定上述多个阶段的触发顺序，每个阶段包含至少一个数据采集器，相同阶段的不同数据采集器位于不同的终端中，分别用于采集不同终端中的数据；还用于执行执行单元 304 的功能，用于按照上述多个阶段的触发顺序，依次触发上述多个阶段的数据采集器采集数据；还用于执行删除单元 305 的功能，用于在检测到上述多个阶段的数据采集器全部采集完数据之后，在上述任务集群中删除上述目标任务。

在另一种可实施的方式中，上述处理器 410，还用于执行确定单元 306 的功能，用于按照上述多个阶段的触发顺序，确定当前待触发阶段；还用于若确定上述待触发阶段位于上述触发顺序的首位，或者确定接收到上述待触发阶段的前一阶段的所有数据采集器的完成采集的信息，则触发上述待触发阶段的所有数据采集器采集数据；还用于在接收到上述待触发阶段的所有数据采集器的完成采集的信息的情况下，触发上述待触发阶段的下一阶段的所有数据采集器采集数据，直至接收到上述多个阶段的所有数据采集器的完成采集的信息。

在另一种可实施的方式中，上述处理器 410 还用于执行获取单元 307 的功能，用于在上述触发上述待触发阶段的所有数据采集器采集数据之前，获取上述待触发阶段的前一阶段的所有数据采集器采集到的数据；上述处理器 410 还用于保存单元 308，用于将上述待触发阶段的前一阶段的所有数据采集器采集到的数据保存到本地数据库中；还用于接收到上述待触发阶段的目标数据采集器的数据请求指令，上述数据请求指令用于请求上述本地数据库中的目标数据；还用于上述在上述本地数据库中调取上述目标数据；上述数据采集装置还包括通信单元 430，用于执行发送单元 310 的功能，用于将上述目标数据发送给上述目标数据采集器；上述上述处理器 410 还用于在接收到上述待触发阶段的所有数据采集器的采集完成消息的情况下，将上述待触发阶段的所有数据采集器采集到的数据保存到上述本地数据库中。

在另一种可实施的方式中，上述处理器 410，还用于检测上述任务集群中新增的任务；上述处理器 410，还用于确定上述新增的任务为分布式任务，以及确定检测到上述任务集群中存在新增的目标任务。

在另一种可实施的方式中，上述处理器 410，还用于检测已创建的数据采集器的个数；上述处理器 410，还用于获取上述已创建的数据采集器的创建时间；上述处理器 410，还用于若上述已创建的数据采集器的个数大于预设个数，则按照上述创建时间从早到晚的排序，依次删除上述已创建的数据采集器，直到剩下的已创建的数据采集器的个数小于或等于上述预设个数。

在另一种可实施的方式中，上述处理器 410 还用于检测宕机的数据采集器；上述处理器 410，还用于删除上述宕机的数据采集器；上述处理器 410，还用于获取上述宕机的数据采集器的初始化信息；上述处理器 410 还用于执行替换单元 311 的功能，利用上述初始化信息创建新的数据采集器，以替代上述宕机的数据数据采集器。

在另一种可实施的方式中，上述 Actor 模型还包括至少一个监控节点，上述监控节点用于初始化宕机的数据采集器，上述处理器 410 还用于检测宕机的数据采集器；上述处理器 410 还用于执行恢复单元 312 的功能，用于利用上述监控节点对上述宕机的数据采集器进行初始化，以控制上述宕机的数据采集器恢复正常。

应当理解，在本申请实施例中，所称处理器 410 可以是中央处理单元 (Central Processing Unit，CPU)，该处理器还可以是其他通用处理器、数字信号处理器 (Digital Signal Processor，DSP)、专用集成电路 (Application Specific Integrated Circuit，ASIC)、现成可编程门阵列 (Field-Programmable Gate Array，FPGA) 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

该存储器 420 可以包括只读存储器和随机存取存储器，并向处理器 410 提供指令和数据。存储器 420 的一部分还可以包括非易失性随机存取存储器。例如，存储器 420 还可以存储设备类型的信息。

具体实现中，本申请实施例中所描述的处理器 410 可执行本申请实施例提供的数据采集方法的第一实施例和第二实施例中所描述的实现方式，也可执行本申请实施例所描述的数据采集装置的实现方式，在此不再赘述。

在本申请的另一实施例中提供一种计算机可读存储介质，计算机可读存储介质存储有计算机程序，计算机程序包括程序指令，程序指令被处理器执行。

计算机可读存储介质可以是前述任一实施例的数据采集装置的内部存储单元，例如数据采集装置的硬盘或内存。计算机可读存储介质也可以是数据采集装置的外部存储设备，例如数据采集装置上配备的插接式硬盘，智能存储卡（Smart Media Card, SMC），安全数字（Secure Digital, SD）卡，闪存卡（Flash Card）等。进一步地，计算机可读存储介质还可以既包括数据采集装置的内部存储单元也包括外部存储设备。计算机可读存储介质用于存储计算机程序以及数据采集装置所需的其他程序和数据。计算机可读存储介质还可以用于暂时地存储已经输出或者将要输出的数据。

权 利 要 求

1、一种基于 Actor 模型的数据采集方法，其特征在于，包括：

检测任务集群中是否存在新增的目标任务；

在检测到任务集群中存在新增的目标任务的情况下 ,为所述目标任务创建 Actor 模型，所述 Actor 模型包括多个数据采集器，所述数据采集器用于采集数据，所述多个数据采集器之间通过消息传递的方式通信；

将所述多个数据采集器划分到多个阶段，并设定所述多个阶段的触发顺序，每个阶段包含至少一个数据采集器，相同阶段的不同数据采集器位于不同的终端中，分别用于采集不同终端中的数据；

按照所述多个阶段的触发顺序，依次触发所述多个阶段的数据采集器采集数据；

在检测到所述多个阶段的所有数据采集器全部采集完数据之后，在所述任务集群中删除所述目标任务。

2、根据权利要求 1 所述的方法，其特征在于，所述按照所述多个阶段的触发顺序，依次触发所述多个阶段的数据采集器采集数据，包括：

按照所述多个阶段的触发顺序，确定当前待触发阶段；

若确定所述待触发阶段位于所述触发顺序的首位，或者确定接收到所述待触发阶段的前一阶段的所有数据采集器的完成采集的信息，则触发所述待触发阶段的所有数据采集器采集数据，所述完成采集的信息用于指示所述数据采集器完成了数据采集；

在接收到所述待触发阶段的所有数据采集器的完成采集的信息的情况下，触发所述待触发阶段的下一阶段的所有数据采集器采集数据，直至接收到所述多个阶段的所有数据采集器的完成采集的信息。

3、根据权利要求 2 所述的方法，其特征在于，在所述触发所述待触发阶段的所有数据采集器采集数据之前，所述方法还包括：

获取所述待触发阶段的前一阶段的所有数据采集器采集到的数据，并将所述待触发阶段的前一阶段的所有数据采集器采集到的数据保存到本地数据库中；

接收到所述待触发阶段的目标数据采集器的数据请求指令，所述数据请求指令用于请求所述本地数据库中的目标数据；

在所述本地数据库中调取所述目标数据，并将所述目标数据发送给所述目标数据采集器；

在接收到所述待触发阶段的所有数据采集器的采集完成消息的情况下，将所述待触发阶段的所有数据采集器采集到的数据保存到所述本地数据库中。

4、根据权利要求 1 所述的方法，其特征在于，所述检测任务集群中是否存在新增的目标任务，包括：

检测所述任务集群中新增的任务；

若确定所述新增的任务为分布式任务，则确定检测到所述任务集群中存在新增的目标

任务。

5、根据权利要求 1 所述的方法，其特征在于，所述在所述任务集群中删除所述目标任务之后，还包括：

检测已创建的数据采集器的个数，并获取所述已创建的数据采集器的创建时间；

若所述已创建的数据采集器的个数大于预设个数，则按照所述创建时间从早到晚的排序，依次删除所述已创建的数据采集器，直到剩下的已创建的数据采集器的个数小于或等于所述预设个数。

6、根据权利要求 1 所述的方法，其特征在于，所述方法还包括：

在检测到宕机的数据采集器的情况下，删除所述宕机的数据采集器；

获取所述宕机的数据采集器的初始化信息；

利用所述初始化信息创建新的数据采集器，以替代所述宕机的数据数据采集器。

7、根据权利要求 1 所述的方法，其特征在于，所述 Actor 模型还包括至少一个监控节点，所述监控节点用于初始化宕机的数据采集器，所述方法还包括：

若检测到宕机的数据采集器，则利用所述监控节点对所述宕机的数据采集器进行初始化，以控制所述宕机的数据采集器恢复正常。

8、根据权利要求 1 所述的方法，其特征在于，所述 Actor 模型还包括至少一个监控节点，所述监控节点用于初始化宕机的数据采集器，所述方法还包括：

若检测到宕机的数据采集器，则获取所述宕机的数据采集器在宕机前一个时刻的状态信息；

利用所述监控节点根据所述宕机的数据采集器在宕机前一个时刻的状态信息，对所述宕机的数据采集器进行恢复，使得所述宕机的数据采集器恢复到所述状态信息对应的正常状态。

9、根据权利要求 1 所述的方法，其特征在于，所述在检测到任务集群中存在新增的目标任务的情况下，所述方法还包括：

获取所述目标任务的标签，所述标签用于指示所述目标任务的重要等级；

根据所述目标任务的标签确定所述目标任务的重要等级；

若所述目标任务的重要等级超过了预设等级，则创建包括有至少一个监控节点的 Actor 模型，所述监控节点用于初始化宕机的数据采集器。

10、一种数据采集装置，其特征在于，包括：

检测单元，用于检测任务集群中是否存在新增的目标任务；

创建单元，用于在检测到任务集群中存在新增的目标任务的情况下，为所述目标任务创建 Actor 模型，所述 Actor 模型包括多个数据采集器，所述数据采集器用于采集数据，所述多个数据采集器之间通过消息传递的方式通信；

划分单元，用于将所述多个数据采集器划分到多个阶段，并设定所述多个阶段的触发顺序，每个阶段包含至少一个数据采集器，相同阶段的不同数据采集器位于不同的终端中，分别用于采集不同终端中的数据；

执行单元，用于按照所述多个阶段的触发顺序，依次触发所述多个阶段的数据采集器采集数据；

删除单元，用于在检测到所述多个阶段的数据采集器全部采集完数据之后，在所述任务集群中删除所述目标任务。

11、一种数据采集装置，其特征在于，包括处理器和存储器，所述处理器和存储器相互连接；所述存储器用于存储计算机程序，所述计算机程序包括程序指令；所述处理器被配置用于调用所述程序指令，用以执行：

检测任务集群中是否存在新增的目标任务；

在检测到任务集群中存在新增的目标任务的情况下，为所述目标任务创建 Actor 模型，所述 Actor 模型包括多个数据采集器，所述数据采集器用于采集数据，所述多个数据采集器之间通过消息传递的方式通信；

将所述多个数据采集器划分到多个阶段，并设定所述多个阶段的触发顺序，每个阶段包含至少一个数据采集器，相同阶段的不同数据采集器位于不同的终端中，分别用于采集不同终端中的数据；

按照所述多个阶段的触发顺序，依次触发所述多个阶段的数据采集器采集数据；

在检测到所述多个阶段的所有数据采集器全部采集完数据之后，在所述任务集群中删除所述目标任务。

12、根据权利要求 11 所述的装置，其特征在于，所述处理器具体用于：

按照所述多个阶段的触发顺序，确定当前待触发阶段；

若确定所述待触发阶段位于所述触发顺序的首位，或者确定接收到所述待触发阶段的前一阶段的所有数据采集器的完成采集的信息，则触发所述待触发阶段的所有数据采集器采集数据，所述完成采集的信息用于指示所述数据采集器完成了数据采集；

在接收到所述待触发阶段的所有数据采集器的完成采集的信息的情况下，触发所述待触发阶段的下一阶段的所有数据采集器采集数据，直至接收到所述多个阶段的所有数据采集器的完成采集的信息。

13、根据权利要求 12 所述的装置，其特征在于，所述处理器，还用于：

获取所述待触发阶段的前一阶段的所有数据采集器采集到的数据，并将所述待触发阶段的前一阶段的所有数据采集器采集到的数据保存到本地数据库中；

接收到所述待触发阶段的目标数据采集器的数据请求指令，所述数据请求指令用于请求所述本地数据库中的目标数据；

在所述本地数据库中调取所述目标数据，并将所述目标数据发送给所述目标数据采集器；

在接收到所述待触发阶段的所有数据采集器的采集完成消息的情况下，将所述待触发阶段的所有数据采集器采集到的数据保存到所述本地数据库中。

14、根据权利要求 11 所述的装置，其特征在于，所述处理器具体用于：

检测所述任务集群中新增的任务；

若确定所述新增的任务为分布式任务，则确定检测到所述任务集群中存在新增的目标任务。

15、根据权利要求 11 所述的装置，其特征在于，所述处理器还用于：

检测已创建的数据采集器的个数，并获取所述已创建的数据采集器的创建时间；

若所述已创建的数据采集器的个数大于预设个数，则按照所述创建时间从早到晚的排序，依次删除所述已创建的数据采集器，直到剩下的已创建的数据采集器的个数小于或等于所述预设个数。

16、根据权利要求 11 所述的装置，其特征在于，所述处理器还用于：

在检测到宕机的数据采集器的情况下，删除所述宕机的数据采集器；

获取所述宕机的数据采集器的初始化信息；

利用所述初始化信息创建新的数据采集器，以替代所述宕机的数据数据采集器。

17、根据权利要求 11 所述的装置，其特征在于，所述 Actor 模型还包括至少一个监控节点，所述监控节点用于初始化宕机的数据采集器，所述处理器还用于：

若检测到宕机的数据采集器，则利用所述监控节点对所述宕机的数据采集器进行初始化，以控制所述宕机的数据采集器恢复正常。

18、根据权利要求 11 所述的装置，其特征在于，所述 Actor 模型还包括至少一个监控节点，所述监控节点用于初始化宕机的数据采集器，所述处理器还用于：

若检测到宕机的数据采集器，则获取所述宕机的数据采集器在宕机前一个时刻的状态信息；

利用所述监控节点根据所述宕机的数据采集器在宕机前一个时刻的状态信息，对所述宕机的数据采集器进行恢复，使得所述宕机的数据采集器恢复到所述状态信息对应的正常状态。

19、根据权利要求 11 所述的装置，其特征在于，所述处理器还用于：

获取所述目标任务的标签，所述标签用于指示所述目标任务的重要等级；

根据所述目标任务的标签确定所述目标任务的重要等级；

若所述目标任务的重要等级超过了预设等级，则创建包括有至少一个监控节点的 Actor 模型，所述监控节点用于初始化宕机的数据采集器。

20、一种计算机可读存储介质，其特征在于，所述计算机可读存储介质存储有计算机程序，所述计算机程序包括程序指令，所述程序指令当被处理器执行时使所述处理器执行如权利要求 1-9 任一项所述的方法。

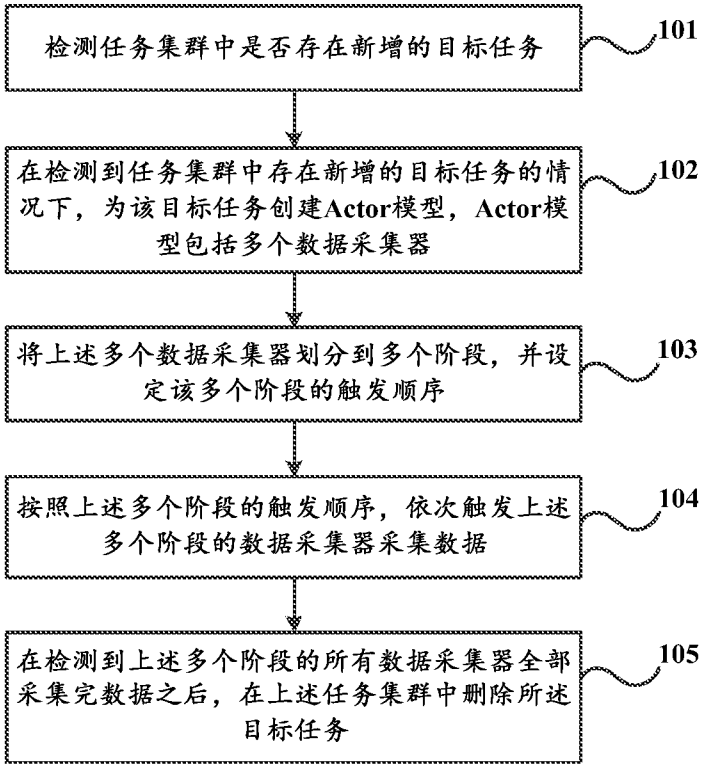

检测任务集群中是否存在新增的目标任务
101
在检测到任务集群中存在新增的目标任务的情况下，为该目标任务创建Actor模型，Actor模型包括多个数据采集器
102
将上述多个数据采集器划分到多个阶段，并设定该多个阶段的触发顺序
103
按照上述多个阶段的触发顺序，依次触发上述多个阶段的数据采集器采集数据
104
在检测到上述多个阶段的所有数据采集器全部采集完数据之后，在上述任务集群中删除所述目标任务
105

图 1

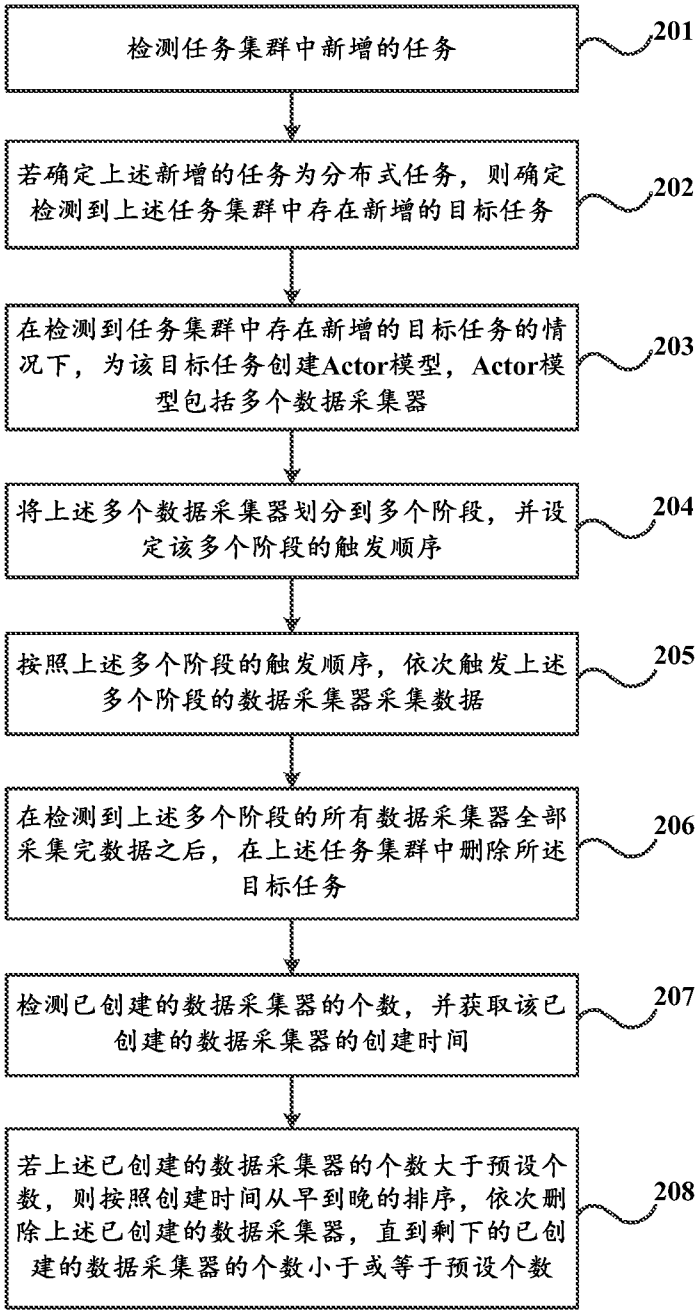

检测任务集群中新增的任务
201
若确定上述新增的任务为分布式任务，则确定检测到上述任务集群中存在新增的目标任务
202
在检测到任务集群中存在新增的目标任务的情况下，为该目标任务创建Actor模型，Actor模型包括多个数据采集器
203
将上述多个数据采集器划分到多个阶段，并设定该多个阶段的触发顺序
204
按照上述多个阶段的触发顺序，依次触发上述多个阶段的数据采集器采集数据
205
在检测到上述多个阶段的所有数据采集器全部采集完数据之后，在上述任务集群中删除所述目标任务
206
检测已创建的数据采集器的个数，并获取该已创建的数据采集器的创建时间
207
若上述已创建的数据采集器的个数大于预设个数，则按照创建时间从早到晚的排序，依次删除上述已创建的数据采集器，直到剩下的已创建的数据采集器的个数小于或等于预设个数
208

图 2

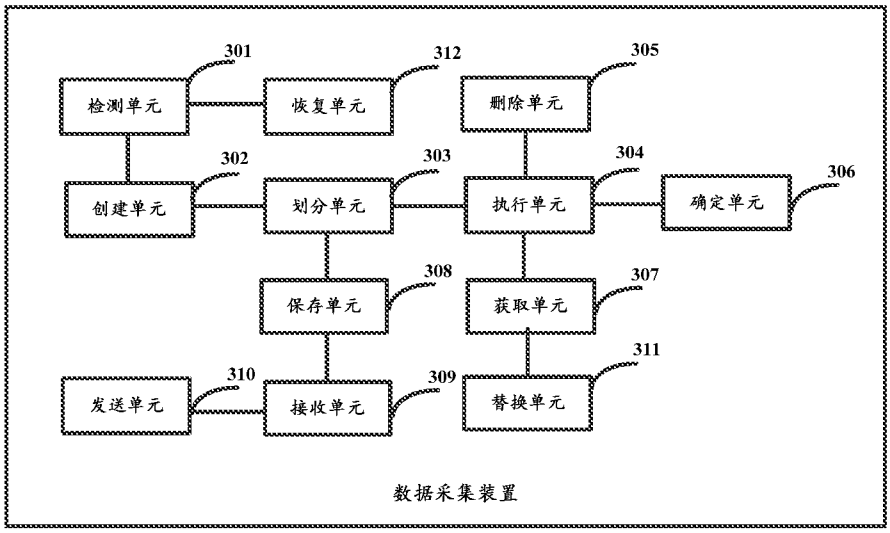

301
检测单元
312
恢复单元
305
删除单元
302
创建单元
303
划分单元
304
执行单元
306
确定单元
308
保存单元
307
获取单元
311
310
发送单元
309
接收单元
替换单元
数据采集装置

图 3

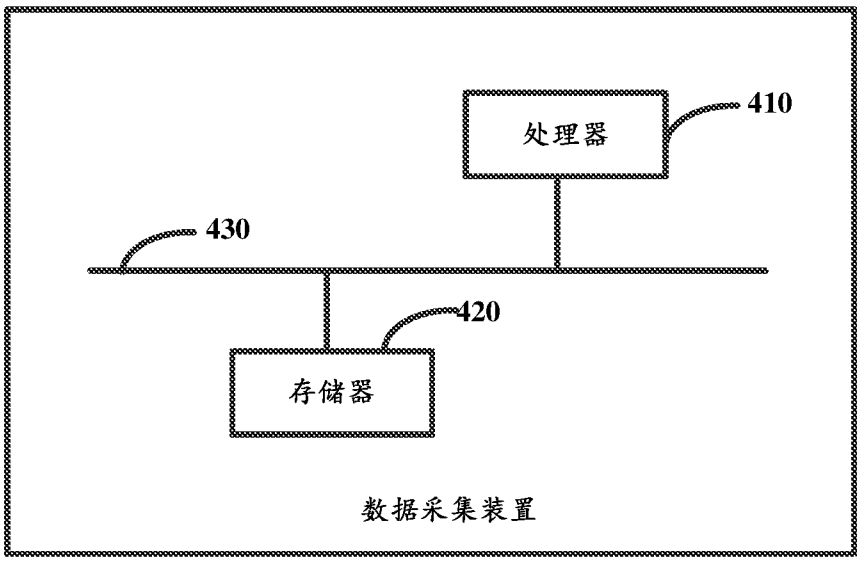

410
处理器
430
420
存储器
数据采集装置

图 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/102705

A. CLASSIFICATION OF SUBJECT MATTER

G06F 9/52(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, CNKI, WPI, EPODOC: Actor, 模型, 任务, 数据, 采集, 阶段, 顺序, 并发, 锁, 分布式, 多机, 多核, 多终端, AKKA, model, task, data, acquisition, phase, concurrency, lock, distributed, multi, machine, core, terminal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	朱雨晴 (ZHU, Yuqing). "基于流式处理的数据采集系统的设计与实现 (Design and implementation of data acquisition system based on streaming processing)" *中国优秀硕士学位论文全文数据库信息科技辑 (China Master's Theses Full-text Database, Information Science and Technology),* No. 11, 15 November 2018 (2018-11-15), abstract, main body, sections 2.2.2, 5.1	1-20
Y	CN 109766326 A (CHONGQING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 17 May 2019 (2019-05-17) description, paragraphs [0029]-[0038]	1-20
A	王世凯 (WANG, Shikai). "灾害监测中海量数据的处理方法究 (The research of massive data processing in disaster monitoring)" *中国优秀硕士学位论文全文数据库 信息科技辑 (Chinese Master's Theses Full-Text Database, Information Science & Technology),* No. 6, 15 June 2016 (2016-06-15), entire document	1-20
A	US 2012173715 A1 (ORACLE INTERNATIONAL CORPORATION) 05 July 2012 (2012-07-05) entire document	1-20

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
12 February 2020	**24 February 2020**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/102705

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 109766326 A	17 May 2019	None	
US 2012173715 A1	05 July 2012	US 2012210012 A1	16 August 2012
		US 2012173614 A1	05 July 2012

国际检索报告

国际申请号
PCT/CN2019/102705

A. 主题的分类

G06F 9/52(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用）)

CNPAT, CNKI, WPI, EPODOC: Actor, 模型, 任务, 数据, 采集, 阶段, 顺序, 并发, 锁, 分布式, 多机, 多核, 多终端, AKKA, model, task, data, acquisition, phase, concurrency, lock, distributed, multi, machine, core, terminal

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
Y	朱雨晴. "基于流式处理的数据采集系统的设计与实现" 中国优秀硕士学位论文全文数据库 信息科技辑, 第11期，2018年 11月 15日 (2018-11-15)， 摘要、正文第2.2.2节，第5.1节	1-20
Y	CN 109766326 A (重庆邮电大学) 2019年 5月 17日 (2019-05-17) 说明书第[0029]-[0038]段	1-20
A	王世凯. "灾害监测中海量数据的处理方法究" 中国优秀硕士学位论文全文数据库 信息科技辑, 第6期，2016年 6月 15日 (2016-06-15)， 全文	1-20
A	US 2012173715 A1 (ORACLE INTERNATIONAL CORPORATION) 2012年 7月 5日 (2012-07-05) 全文	1-20

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

"A" 认为不特别相关的表示了现有技术一般状态的文件

"E" 在国际申请日的当天或之后公布的在先申请或专利

"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

"O" 涉及口头公开、使用、展览或其他方式公开的文件

"P" 公布日先于国际申请日但迟于所要求的优先权日的文件

"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

"&" 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2020年 2月 12日	2020年 2月 24日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 文娟 电话号码 86-10-53961609

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/102705

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	109766326	A	2019年 5月 17日	无			
US	2012173715	A1	2012年 7月 5日	US	2012210012	A1	2012年 8月 16日
				US	2012173614	A1	2012年 7月 5日