(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114503109 B
(45)授权公告日 2022.09.23

(21)申请号 202080065635.3
(22)申请日 2020.10.14
(65)同一申请的已公布的文献号
申请公布号 CN 114503109 A
(43)申请公布日 2022.05.13
(30)优先权数据
16/655,372 2019.10.17 US
(85)PCT国际申请进入国家阶段日
2022.03.18
(86)PCT国际申请的申请数据
PCT/IB2020/059628 2020.10.14
(87)PCT国际申请的公布数据
WO2021/074802 EN 2021.04.22

(73)专利权人 国际商业机器公司
地址 美国纽约
(72)发明人 P·巴斯蒂德 R·兰察尔
S·M·坎 S·巴克塔瓦查拉姆
(74)专利代理机构 中国贸促会专利商标事务所有限公司 11038
专利代理师 邹丹

(51)Int.Cl.
G06F 21/56(2006.01)
(56)对比文件
US 2015370845 A1,2015.12.24
CN 101002190 A,2007.07.18
审查员 刘琪

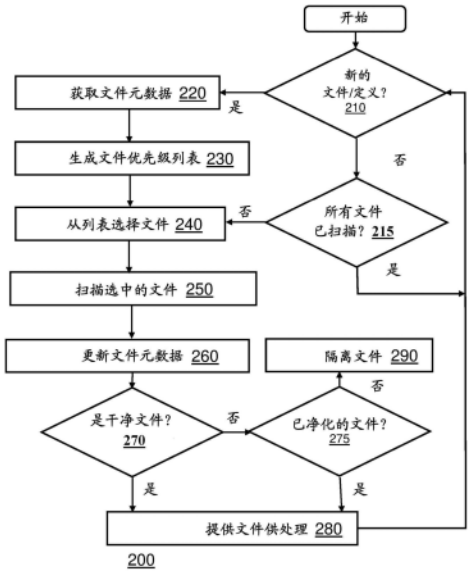
权利要求书3页 说明书12页 附图4页

(54)发明名称

用于维护数据处理系统安全性的方法和系统

(57)摘要

通过以下方式维护系统安全性:从元数据存储单元接收与一个数据文件的至少一部分相关联的元数据,根据元数据为一个数据文件的该至少一部分生成优先级,以及进行数据文件的该部分的扫描。元数据包括一个或多个病毒指示符。



CN 114503109 B

1.一种用于维护数据处理系统安全性的计算机实现的方法,所述方法由一个或多个计算机处理器实现,所述方法包括:

将多个流接收到数据储存区中,每个流对应于相应数据文件;

对于对应的相应数据文件中的每个数据文件:

使用所述多个流中的对应的相应流的至少一部分计算相应校验和,

通过将所计算的相应校验和与和元数据对应地存储的校验和匹配,从元数据存储单元中选择对应的相应元数据,以及

从所选择的对应的相应元数据中检索相应一个或多个病毒指示符;

通过按照所检索的相应一个或多个病毒指示符对每个对应的相应数据文件的优先级进行排序,生成包括每个对应的相应数据文件的按优先级排序的列表;

在进行对按优先级排序的列表中的另一数据文件的扫描之前,进行对按优先级排序的列表中具有最高优先级的第一数据文件的扫描;

在进行对按优先级排序的列表中的另一数据文件的扫描或对按优先级排序的列表中具有最高优先级的第一数据文件的扫描之后,接收与新数据文件对应的新流;

从来自所述元数据存储单元的新选择的对应的元数据中检索与所述新数据文件对应的新的一个或多个病毒指示符;

通过按照所检索的相应一个或多个病毒指示符和所述新的一个或多个病毒指示符对未扫描的每个对应的相应数据文件和所述新数据文件的优先级进行排序,重新生成所述按优先级排序的列表,包括未扫描的每个对应的相应数据文件和所述新数据文件;

依次扫描重新生成的按优先级排序的列表中的每个文件,从重新生成的按优先级排序的列表中的最高优先级文件开始;以及

阻止用户访问重新生成的按优先级排序的列表中的每个文件,直到每个文件被扫描完为止。

2.根据权利要求1所述的计算机实现的方法,其中,与每个对应的相应数据文件流式进入数据储存区并行地连续计算每个相应校验和。

3.根据权利要求1所述的计算机实现的方法,还包括:由所述一个或多个计算机处理器根据对每个对应的相应数据文件的扫描来更新每个对应的相应数据文件的每个对应的相应元数据;以及由所述一个或多个计算机处理器提供每个对应的相应数据文件用于进一步处理。

4.根据权利要求1所述的计算机实现的方法,其中,根据相关性生成所述按优先级排序的列表。

5.根据权利要求1所述的计算机实现的方法,还包括:由所述一个或多个计算机处理器将所述相应数据文件中的至少一个数据文件呈现给净化例程。

6.根据权利要求1所述的计算机实现的方法,还包括通过所述一个或多个计算机处理器更新与所述元数据对应地存储的校验和之一。

7.一种用于维护系统安全性的计算机可读存储设备,所述计算机可读存储设备包括非瞬态计算机可读存储器,所述非瞬态计算机可读存储器包括存储在其上的程序指令,所存储的程序指令在被执行时,使得执行包括以下的步骤:

将多个流接收到数据储存区中,每个流对应于相应数据文件;

对于对应的相应数据文件中的每个数据文件：

使用所述多个流中的对应的相应流的至少一部分计算相应校验和，

通过将所计算的相应校验和与和元数据对应地存储的校验和匹配，从元数据存储单元中选择对应的相应元数据，以及

从所选择的对应的相应元数据中检索相应一个或多个病毒指示符；

通过按照所检索的相应一个或多个病毒指示符对每个对应的相应数据文件的优先级进行排序，生成包括每个对应的相应数据文件的按优先级排序的列表；

在进行对按优先级排序的列表中的另一数据文件的扫描之前，进行对按优先级排序的列表中具有最高优先级的第一数据文件的扫描；

在进行对按优先级排序的列表中的另一数据文件的扫描或对按优先级排序的列表中具有最高优先级的第一数据文件的扫描之后，接收与新数据文件对应的新流；

从来自所述元数据存储单元的新选择的对应的元数据中检索与所述新数据文件对应的新的一个或多个病毒指示符；

通过按照所检索的相应一个或多个病毒指示符和所述新的一个或多个病毒指示符对未扫描的每个对应的相应数据文件和所述新数据文件的优先级进行排序，重新生成所述按优先级排序的列表，包括未扫描的每个对应的相应数据文件和所述新数据文件；

依次扫描重新生成的按优先级排序的列表中的每个文件，从重新生成的按优先级排序的列表中的最高优先级文件开始；以及

阻止用户访问重新生成的按优先级排序的列表中的每个文件，直到每个文件被扫描完为止。

8.根据权利要求7所述的计算机可读存储设备，其中，与每个对应的相应数据文件流式进入数据储存区并行地连续计算每个相应校验和。

9.根据权利要求7所述的计算机可读存储设备，还包括：用于以下的程序指令：根据对每个对应的相应数据文件的扫描来更新每个对应的相应数据文件的每个对应的相应元数据；以及提供每个对应的相应数据文件用于进一步处理。

10.根据权利要求7所述的计算机可读存储设备，其中，根据相关性生成所述按优先级排序的列表。

11.根据权利要求7所述的计算机可读存储设备，还包括用于将所述相应数据文件中的至少一个数据文件呈现给净化例程的程序指令。

12.根据权利要求7所述的计算机可读存储设备，还包括用于更新与所述元数据对应地存储的校验和之一的程序指令。

13.一种用于维护系统安全性的计算机系统，所述计算机系统包括：一个或多个计算机处理器；和一个或多个计算机可读存储设备，所述一个或多个计算机可读存储设备包括非瞬态计算机可读存储器，所述非瞬态计算机可读存储器包括存储在其上的程序指令以供由所述一个或多个计算机处理器执行，所存储的程序指令包括用于以下的程序指令：

将多个流接收到数据储存区中，每个流对应于相应数据文件；

对于对应的相应数据文件中的每个数据文件：

使用所述多个流中的对应的相应流的至少一部分计算相应校验和，

通过将所计算的相应校验和与和元数据对应地存储的校验和匹配，从元数据存储单元

中选择对应的相应元数据,以及

从所选择的对应的相应元数据中检索相应一个或多个病毒指示符;

通过按照所检索的相应一个或多个病毒指示符对每个对应的相应数据文件的优先级进行排序,生成包括每个对应的相应数据文件的按优先级排序的列表;

在进行对按优先级排序的列表中的另一数据文件的扫描之前,进行对按优先级排序的列表中具有最高优先级的第一数据文件的扫描;

在进行对按优先级排序的列表中的另一数据文件的扫描或对按优先级排序的列表中具有最高优先级的第一数据文件的扫描之后之后,接收与新数据文件对应的新流;

从来自所述元数据存储单元的新选择的对应的元数据中检索与所述新数据文件对应的新的一个或多个病毒指示符;

通过按照所检索的相应一个或多个病毒指示符和所述新的一个或多个病毒指示符对未扫描的每个对应的相应数据文件和所述新数据文件的优先级进行排序,重新生成所述按优先级排序的列表,包括未扫描的每个对应的相应数据文件和所述新数据文件;

依次扫描重新生成的按优先级排序的列表中的每个文件,从重新生成的按优先级排序的列表中的最高优先级文件开始;以及

阻止用户访问重新生成的按优先级排序的列表中的每个文件,直到每个文件被扫描完为止。

14.根据权利要求13所述的计算机系统,其中,与每个对应的相应数据文件流式进入数据储存区并行地连续计算每个相应校验和。

15.根据权利要求13所述的计算机系统,还包括:用于以下的程序指令:根据对每个对应的相应数据文件的扫描来更新每个对应的相应数据文件的每个对应的相应元数据;以及提供每个对应的相应数据文件用于进一步处理。

16.根据权利要求13所述的计算机系统,其中,根据相关性生成所述按优先级排序的列表。

17.根据权利要求13所述的计算机系统,还包括用于将所述相应数据文件中的至少一个数据文件呈现给净化例程的程序指令。

18.根据权利要求13所述的计算机系统,还包括用于更新与所述元数据对应地存储的校验和之一的程序指令。

用于维护数据处理系统安全性的方法和系统

技术领域

[0001] 本公开总体上涉及维护数据处理系统安全性和大型数据存储系统中的数据的完整性。本公开尤其涉及通过扫描恶意代码和数据来维护系统安全性和数据完整性。

背景技术

[0002] 当前的多租户系统(诸如医疗保健、零售、电信和金融服务系统)可以具有连接到单个逻辑数据处理系统的多个供应商、消费者和服务组织。逻辑数据处理系统可以是基于云的或者是本地的,并且可以接收连续的数据流以及大块数据批次,在数据储存区中累积成千兆字节和太字节的数据。

[0003] 输入数据必须扫描病毒、恶意软件和其他恶意数据,以保持系统和其用户的安全性。当随着识别到新的零日漏洞而更新病毒定义时,必须重新扫描现有数据。入侵预防系统(IPS)用于监视输入数据。IPS可以是内联的,当其从外部源流传输到数据湖以用于存储时监控实时数据,或者可以是带外的,处理传递到IPS的数据的副本。传入数据还可包括冗余数据,这通过需要扫描副本内容而增加IPS的工作负荷。

[0004] IPS的扫描活动必须在不中断消费者对数据的访问并且不中断输入数据的连续流的情况下执行。流必须在不中断的情况下着陆在系统中以继续正常操作、降低进入/输出成本、减少或消除系统资源的任何浪费,并且维持系统平台的用户和其他部分对系统服务的完全访问。

[0005] 因此,在本领域中需要解决上述问题。

发明内容

[0006] 从另一方面来看,本发明提供了一种用于维护数据处理系统安全性的计算机实现的方法,该方法包括:由一个或多个计算机处理器从元数据存储单元接收与数据文件的至少一部分相关联的元数据;由该一个或多个计算机处理器根据元数据生成用于数据文件的该至少一部分的优先级;以及由该一个或多个计算机处理器根据优先级来执行对数据文件的扫描;其中,元数据包括一个或多个病毒指示符。

[0007] 从另一个方面来看,本发明提供了用于维护系统安全性的计算机实现的方法,该方法包括:由一个或多个计算机处理器接收更新的病毒定义;由该一个或多个计算机处理器从元数据存储单元中获取与至少一个数据文件相关联的元数据;由该一个或多个计算机处理器根据元数据和更新的病毒定义生成该至少一个数据文件的优先级;由该一个或多个计算机处理器根据优先级来选择数据文件;由该一个或多个计算机处理器进行对数据文件的扫描;由该一个或多个计算机处理器根据扫描来更新数据文件的元数据;以及由该一个或多个计算机处理器提供数据文件用于进一步处理。

[0008] 从另一方面来看,本发明提供了一种用于维护系统安全性的计算机实现的方法,该方法包括:由一个或多个计算机处理器从元数据存储单元获得与至少一个数据文件相关联的元数据;由该一个或多个计算机处理器根据元数据为该述至少一个数据文件生成优先

级;由该一个或多个计算机处理器根据优先级来选择数据文件;由该一个或多个计算机处理器进行对数据文件的扫描;由该一个或多个计算机处理器确定数据文件的常见漏洞和暴露评分;由该一个或多个计算机处理器根据扫描更新数据文件的元数据;以及由该一个或多个计算机处理器将评分提供给系统用户用于评估。

[0009] 从另一方面来看,本发明提供了用于维护系统安全性的计算机系统,该计算机系统包括:一个或多个计算机处理器;一个或多个计算机可读存储设备;以及在该一个或多个计算机可读存储设备上存储的用于由该一个或多个计算机处理器执行的程序指令,该存储的程序指令包括:用于从元数据存储单元接收与至少一个数据文件相关联的元数据的程序指令;用于根据元数据生成用于该至少一个数据文件的优先级的程序指令;以及用于根据优先级进行对数据文件的扫描的程序指令;其中,元数据包括一个或多个病毒指示符。

[0010] 从另一方面来看,本发明提供了一种用于维护系统安全性的计算机程序产品,该计算机程序产品包括计算机可读存储介质,该计算机可读存储介质可由处理电路读取并且存储用于由该处理电路执行以便执行用于执行本发明的步骤的方法的指令。

[0011] 从另一方面来看,本发明提供一种存储在计算机可读介质上并且能够加载到数字计算机的内部存储器中的计算机程序,该计算机程序包括当该程序在计算机上运行时用于执行本发明的步骤的软件代码部分。

[0012] 从另一方面来看,本发明提供了一种用于维护系统安全性的计算机程序产品,该计算机程序产品包括一个或多个计算机可读存储设备以及在该一个或多个计算机可读存储设备上的所存储的程序指令,所存储的程序指令包括:用于从元数据存储单元接收与至少一个数据文件相关联的元数据的程序指令;用于根据元数据生成用于该至少一个数据文件的优先级的程序指令;以及用于根据优先级进行对数据文件的扫描的程序指令;其中,该元数据包括一个或多个病毒指示符。

[0013] 以下呈现概述以提供对本公开的一个或多个实施例的基本理解。本概述并不旨在识别关键或重要元素或者划定特定实施例的任何范围或权利要求的任何范围。其唯一的目的是以简化的形式呈现概念,作为稍后呈现的更详细描述的序言。在本文所描述的一个或多个实施例中,设备、系统、计算机实现的方法、装置和/或计算机程序产品使得能够在允许用户访问数据之前针对恶意代码对数据文件进行优先级排序和扫描。

[0014] 在一个方面,本发明包括通过在数据文件被用户处理或以其他方式访问之前,随着数据文件被成批地流式传输或接收而连续地监视这些数据文件而与维护系统安全性相关联的方法、系统和计算机可读介质。在此方面,该系统、方法和计算机程序产品通过以下方式进行:从元数据存储单元接收与一个数据文件的至少一部分相关联的元数据;根据元数据为一个数据文件的该至少一部分生成优先级;以及进行数据文件的该部分的扫描。元数据包括一个或多个病毒指示符。一旦被接收,数据就被提供给用户而不中断。

[0015] 在一个方面,本发明包括在病毒定义随时间改变时与维护系统安全性相关联的方法、系统和计算机可读介质。在此方面,这些系统、方法和计算机程序产品通过以下方式进行:接收更新的病毒定义,从元数据存储单元获得与至少一个数据文件相关联的元数据,根据该元数据和更新的病毒定义为该至少一个数据文件生成优先级,根据所生成的优先级选择数据文件,对该数据文件进行扫描,根据该扫描更新该数据文件的元数据,并且提供该数据文件用于进一步处理。

[0016] 在一个方面，本发明包括与维护系统安全性相关联的方法、系统和计算机可读介质，包括向用户或自动接收系统提供用于每个扫描文件的常见漏洞和暴露评分的附加优点。用户或接收系统然后确定扫描文件是否应被加载到数据处理系统中以供进一步处理和使用。此方面的发明是通过以下方式进行的：从元数据存储单元中获得与至少一个数据文件相关联的元数据；根据该元数据为该至少一个数据文件生成优先级；根据该优先级选择数据文件；对该数据文件进行扫描；为该数据文件确定常见漏洞和暴露评分；根据该扫描对该数据文件的元数据进行更新；并且将该评分提供给系统用户用于评估。

附图说明

[0017] 通过在附图中对本公开的一些实施例进行更详细的描述，本公开的上述和其他目的、特征和优点将变得更加明显，其中，在本公开的实施例中相同的参考标记通常指代相同的部件。

[0018] 图1提供了根据本发明的实施例的系统的示意图。

[0019] 图2提供了描绘根据本发明的实施例的操作序列的流程图。

[0020] 图3描绘了根据本发明的实施例的云计算环境。

[0021] 图4示出了根据本发明的实施例的抽象模型层。

具体实施方式

[0022] 将参考附图更详细地描述一些实施例，在附图中示出了本公开的实施例。然而，本公开可以以各种方式实施，并且因此不应被解释为局限于本文中公开的实施例。

[0023] 在实施例中，系统的一个或多个部件可以采用硬件和/或软件来解决本质上高度技术性的问题(例如，根据元数据对数据文件进行优先级排序和扫描按优先级排序的文件等)。这些解决方案不是抽象的，并且由于需要处理能力来促进例如代码扫描和纠正，这些解决方案不能作为人类的一组智力活动来执行。进一步，所执行的一些过程可由用于执行与存储器操作相关的已定义任务的专用计算机执行。例如，可以采用专用计算机来执行与代码扫描和纠正等相关的任务。

[0024] 分布式大数据处理系统面临独特的系统安全性问题。这些系统接收和分发以千兆字节和太字节计量的非常大的数据文件。所接收的数据来自于多个源的主机，一些源比其他源更可靠，并且其中一些源可能不检查或确保所提供的数据没有病毒、恶意软件和其他与数据相关的嵌入问题。系统用户需要跨分布式环境的对系统数据的不间断访问，其中停止服务或数据访问以审阅或补救数据问题不是可行的选项。系统用户可以下载数据块以用于进一步处理和使用。对数据的安全性简档的指示帮助用户确定哪些块对于添加到数据处理系统以供使用是安全的。这样的系统还可以从单个源或从多个源接收副本文件。针对安全性问题处理副本记录浪费了稀有的系统资源。

[0025] 在实施例中，该系统包括用于存储大(千兆字节和太字节大小)文件的数据储存区。该系统还包括元数据存储单元，该元数据存储单元保持一个或多个元数据表。元数据存储单元可以是本地的或基于云的。这些表包括用于被保持在数据储存区中的大文件的元数据记录。在本实施例中，这些表保持元数据字段，包括以下一些或全部：与文件标识相关的数据、文件处理状态——文件是否可用于下游使用等、文件处理时间——通过该方法处理

数据的最后时间的记录、文件的常见漏洞和暴露(CVE)评分、在文件内识别的病毒、详细的病毒描述数据、文件格式信息、以及与文件的任何下游处理需求相关的数据。不同的文件格式与数据传播问题更密切相关。与长期存储相反，与下游处理相关联的文件需要更立即的关注和扫描。

[0026] 所公开的发明可应用于固定地部署在单个服务器上或跨多个云租户分布的单个或多个应用系统。本发明可应用于流数据的近实时数据处理以及批次和大块(bulk)数据处理。

[0027] 在一些实施例中，元数据还包括与数据文件相关联的校验和数据。对于带外接收的大块或批次数据文件，该方法计算该文件的校验和，如散列值(SHA256)。即使对文件的微小改变也将导致不同的校验和值，从而指示文件已经被改变。校验和被存储为与文件相关联的元数据。在一些实施例中，可以向文件提供校验和值。在这些实施例中，该方法比较计算的校验和与所提供的值以确定是否已经对文件做出改变。校验和可以使用校验和技术，例如SHA256、MD5或使用数据文件字节表示来计算校验和的任何其他方法。

[0028] 在实施例中，该方法计算用于使带外接收的数据文件以及大块文件精简的校验和值。对于流式数据，该方法与流式数据在系统的数据储存区中的着陆并行地连续地计算校验和值。该方法将在流传输期间计算的校验和值作为元数据存储在查找表中。在数据文件已经完全着陆之后，文件的最终校验和值与流式传输的文件相关联地存储在元数据存储单元中。在该实施例中，该方法将数据文件细分成作为文件的块或段的部分。在该实施例中，该方法计算数据文件的每个部分的校验和。当每个部分被完全接收时，数据文件的每个部分的校验和被存储为与该文件的该部分相关联的元数据。在实施例中，部分重叠以确保病毒不在相邻的块之间被拆分且被忽略。可以填充段，使得每个段包括先前段的一部分，例如，可以填充每个100兆字节的段以包括先前段的最后100千字节。在一个实施例中，100兆字节的段被填充+/-500千字节，以确保在段扫描中不存在由于分割数据文件而产生的间隙。在该实施例中，段在0-1.5MB、1-2.5MB、2-3.5MB…的范围内，从而减少恶意代码的一部分通过在两个相邻段之间被拆分而逃避检测的可能性。

[0029] 在实施例中，该方法将校验和值与所存储的与现有数据文件和段相关联的元数据值进行比较。该比较识别当前批次或流式文件/段的副本。该方法在文件被流式传输到数据储存区时比较连续计算的校验和，由此识别潜在相似或相同的文件。在该实施例中，该方法使用最终校验和来将任何潜在相同的文件验证为实际相同的。

[0030] 在实施例中，响应于匹配元数据中的校验和，该方法基于匹配的文件/段的CVE评分和病毒详情触发文件或文件段的净化例程。在该实施例中，匹配的元数据与具有已知病毒的文件相关联。在该实施例中，该方法净化文件或段并更新指示已经发生清理的元数据。清理文件可包括从文件中移除与检测到的病毒相关联的已知恶意代码部分，或使用文件的备份副本来恢复文件。已清理的文件被添加到列表中以用于优先级排序。在该实施例中，该方法为已清理的文件计算新的校验和，并将该新的校验和添加到元数据中。

[0031] 在实施例中，该方法检查元数据存储单元以确定是否已经加载了新的数据储存区批次文件，或确定是否正在加载新的流式数据文件。在该实施例中，该方法检查与任何新文件相关联的元数据。然后，该方法针对新数据文件的指示连续地监测元数据存储单元。在实施例中，当没有检测到新的文件或病毒定义时，本发明的方法和系统停止。

[0032] 在实施例中，该方法获得用于任何最近加载的或当前流传输的文件的元数据。该方法使用元数据生成新的数据文件的按优先级排序的列表。鉴于病毒指示符或文件具有恶意软件的可能性对文件进行优先级排序。该方法关联所有新文件的元数据以生成按优先级排序的列表。就先前发现的病毒而言，该方法考虑相对于先前扫描的文件的校验和值。该方法还考虑在文件或副本的任何先前扫描中使用的病毒定义，以及该病毒定义是否是最新的。使用现今过期的病毒定义扫描进行的扫描对于新扫描被列为更高优先级。需要下游处理的数据文件比缺乏这种需要的文件被列为更高优先级。与已提供的校验和或先前的扫描校验和相比具有校验和偏离的文件对于新扫描被列为更高的优先级。在该实施例中，该方法鉴于对病毒定义的任何改变来考虑该组新的和现有的文件的元数据。该方法对数据文件集合进行排名，并将文件从最高(最需要扫描)到最低(最不需要扫描)按优先级排列。在该实施例中，该方法检查新的未处理文件——不具有匹配的文件标识或文件校验和的文件，将这些文件列为高优先级，并且在用于文件的元数据存储单元中创建新的元数据列表。该方法还阻止用户对未处理文件的访问，至少直到已经完成了对文件的扫描之后。在该实施例中，该方法将文件标识和文件校验和与现有文件相同的副本文件视为低优先级。

[0033] 在实施例中，该方法从所生成的按优先级排序的列表中选择文件并且扫描该文件以查找病毒、恶意软件和其他数据传播问题。在本实施例中，该方法从最高优先级开始，根据其优先级选择数据文件。按照降序的优先级依次扫描列表中的每个文件。在该实施例中，该方法选择列表上的最高优先级文件。将文件与它们的优先级一起添加到列表，使得可以在具有低优先级的较旧文件之前扫描新的、高优先级的文件。

[0034] 在实施例中，该方法利用详细的当前病毒定义针对病毒扫描文件。在该实施例中，该方法确定每个文件的CVE评分以及为每个文件发现的病毒列表。将扫描的时间、所使用的病毒定义、发现的病毒、发现的病毒数量、文件校验和、文件格式以及与该文件相关联的下游处理需求全部更新为元数据存储单元中该文件的元数据。

[0035] 在实施例中，将文件细分成块或段以促进更快的扫描和优化文件存储。在该实施例中，该方法确定并存储用于每个段以及整个文件的元数据。在该实施例中，该方法包括相关段数据作为元数据的一部分。扫描一组相关段中的一个段触发对附近相关段的扫描。在该实施例中，相关段的扫描重叠，从而增加了发现病毒和恶意软件的可能性。在该实施例中，文件可以被划分、存储和扫描为一系列逻辑上相关的段——segment0、segment1、segment2等。在实施例中，该方法扫描文件的二进制版本。在该实施例中，该方法将文件维持为二进制的，直到扫描完成并且该文件适于作为可用数据储存区的一部分而用于下游处理和用户访问之后。在该实施例中，该方法采用提取-变换-加载或其他数据处理代理操作来从文件检索数据以用作数据储存区的一部分。

[0036] 在实施例中，该方法在扫描文件或文件段之后评估元数据。然后，该方法提供用于下游处理的文件，提供用于病毒/恶意软件净化例程的文件，或者当净化例程不可用时隔离该文件。在实施例中，该方法向用户提供对每个文件的CVE评分的访问。在该实施例中，用户确定与CVE评分相关联的可接受风险的水平并且建立CVE评分阈值。具有低于阈值的CVE评分的文件被接受到用户的数据处理系统中以供数据消费者使用。具有高于阈值的评分的那些文件被拒绝。在实施例中，该方法上传未发现病毒或具有可接受的低CVE评分的文件以用于下游使用。该方法对所有扫描文件更新元数据存储单元中的元数据。

[0037] 例如：

[0038] 客户端应用使用传输层安全性相互认证与方法平台连接。

[0039] 客户端应用开始健康护理数据的上传。

[0040] 客户端尚未声明内容的大小，并且平台最大支持是针对多个太字节文件的。

[0041] 平台将数据流式传输到数据储存区中并且使用SHA256开始校验和。

[0042] 平台尚未标记准备好下载的数据并且下游消费者尚未能够消费新数据。

[0043] 平台开始建立用于文件的各部分的校验和。{部分1：0-1M，部分2：1-2M，部分3：2-3M}

[0044] 客户端应用完成文件的上传。

[0045] 平台更新元数据支持列中的校验和——01234567A以及为文件的部分计算的校验和。

[0046] 并行运行，平台的病毒扫描服务检查元数据列。

[0047] 元数据包含上传文件的以下信息：

[0048]

文 件	1-2-1-1

[0049]

ID	
已处理	N
校验和	01234567A
处理时间	NULL
CVE 评分	NULL
病毒和病毒细节	NULL
相关段（前一个/下一个）	NULL

[0050] 方法协调器进行检查以查看校验和是否已经存在，如果存在，则本发明对没有相关或相似校验和的校验和进行优先级排序。

[0051] 未找到任何相关校验和的协调器基于时间提取文件。

[0052] 协调器扫描文件，并且拉回结果——发现的病毒、病毒元数据，并且用病毒内容和CVE评分将文件标记为已处理。

[0053]

文件ID	1-2-1-1
已处理	N
校验和	01234567A
处理时间	2019-02-01 07:56:00AM UTC
CVE评分	CVSS v2.0 严重性和度量: 基础评分: 5.0 中等 向量: (AV:N/AC:L/Au:N/C:N/I:P/A:N) (V2 legend) 影响子评分: 2.9 可发现性子评分: 10.0
病毒评分和内容	Eicar
相关段(前一个/下一个)	无

[0054]

[0055] 图1提供了与实践所公开的发明相关联的示例性网络资源的示意图。本发明可以在处理指令流的任何公开的元件的处理器中实践。如图所示,联网的客户端设备110无线地连接到服务器子系统102。客户端设备104经由网络114无线地连接到服务器子系统102。客户端设备104和110包括病毒扫描程序客户端(未示出)连同用于执行程序的足够的计算资源(处理器、存储器、网络通信硬件)。如图1所示,服务器子系统102包括服务器计算机150。图1示出了根据本发明实施例的联网计算机系统1000内的服务器计算机150的部件的框图。应当理解,图1仅提供一个实现方式的图示并且不暗示关于其中可以实现不同实施例的环境的任何限制。可以对所描绘的环境做出许多修改。

[0056] 服务器计算机150可以包括(一个或多个)处理器154、存储器158、持久性存储装置170、通信单元152、输入/输出(I/O)接口156和通信结构140。通信结构140提供高速缓存162、存储器158、持久性存储装置170、通信单元152和输入/输出(I/O)接口156之间的通信。通信结构140可用设计用于在处理器(诸如微处理器、通信和网络处理器等)、系统存储器、外围设备和系统内的任何其他硬件部件之间传递数据和/或控制信息的任何架构来实现。例如,通信结构140可用一个或多个总线实现。

[0057] 存储器158和持久性存储装置170是计算机可读存储介质。在该实施例中,存储器158包括随机存取存储器(RAM)160。通常,存储器158可包括任何合适的易失性或非易失性计算机可读存储介质。高速缓存162是快速存储器,其通过保持来自存储器158的最近访问的数据和最近访问的数据附近的数据来增强(一个或多个)处理器154的性能。

[0058] 用于实践本发明的实施例的程序指令和数据,例如病毒扫描程序175,被存储在持久性存储装置170中以供服务器计算机150的对应(一个或多个)处理器154中的一个或多个

经由高速缓存162执行和/或访问。在该实施例中，持久性存储装置170包括磁性硬盘驱动器。作为磁硬盘驱动器的替代或补充，持久性存储装置170可包括固态硬盘驱动器、半导体存储设备、只读存储器(ROM)、可擦可编程只读存储器(EPROM)、闪存、或能够存储程序指令或数字信息的任何其他计算机可读存储介质。

[0059] 持久性存储装置170所使用的介质也可以是可移除的。例如，可移除硬盘驱动器可以用于持久性存储装置170。其他示例包括光盘和磁盘、拇指驱动器和智能卡，它们被插入到驱动器中以便转移到其他的计算机可读存储介质(其也是持久性存储装置170的一部分)上。

[0060] 在这些示例中，通信单元152提供与其他数据处理系统或设备的通信，其他数据处理系统或设备包括客户端计算设备104和110的资源。在这些示例中，通信单元152包括一个或多个网络接口卡。通信单元152可通过使用物理和无线通信链路中的任一者或两者提供通信。软件分发程序以及用于实现本发明的其他程序和数据可以通过通信单元152下载到服务器计算机150的持久性存储装置170中。

[0061] (一个或多个)I/O接口156允许与可连接到服务器计算机150的其他设备进行数据的输入和输出。例如，(一个或多个)I/O接口156可提供至(一个或多个)外部设备190(诸如键盘、小键盘、触摸屏、麦克风、数码相机和/或一些其他合适的输入设备)的连接。(一个或多个)外部设备190还可包括便携式计算机可读存储介质，诸如例如拇指驱动器、便携式光盘或磁盘、以及存储卡。用于实践本发明的实施例的软件和数据，例如服务器计算机150上的病毒扫描程序175，可存储在这种便携式计算机可读存储介质上并且可经由(一个或多个)I/O接口156加载到持久性存储装置170上。(一个或多个)I/O接口156还连接到显示器180。

[0062] 显示器180提供向用户显示数据的机制，并且可以是例如计算机监视器。显示器180还可用作触摸屏，诸如平板计算机的显示器。

[0063] 图2提供了示出与本公开的实践相关联的示例性活动的流程图200。在程序启动之后，在210，程序175的方法检查新的批次文件或流式文件和/或新的病毒定义。批次文件可能已经存在于数据储存区中，并且流式文件可能正着陆在数据储存区中。可以通过网络通信单元152向系统入侵预防系统提供新的病毒定义。在220处，该方法检查持久性存储装置170、系统存储区域网络或其他系统存储器的元数据存储单元中的新文件元数据。对于没有已存储的元数据的新文件，程序175的方法创建新的元数据记录并且开始收集文件元数据。从文件本身以及从与文件相关联的方法活动收集元数据。对于新病毒定义，该方法检查尚未用新病毒定义扫描的文件。

[0064] 在230，程序175的方法根据它们对病毒扫描的需要生成所有新文件的按优先级排序的列表。该方法使用所计算的文件校验和值中的一个或多个以及来自新文件元数据的病毒相关性指示符来确定文件优先级。较高的优先级被分配给具有用于扫描的较高紧急性的文件。(需要下游处理访问的文件、不具有历史元数据的文件、具有校验和问题的文件等)。

[0065] 在240，程序175的方法从所生成的列表中选择最高优先级数据文件部分用于扫描。在250，该方法使用当前病毒定义扫描文件。在260，该方法更新文件的元数据，包括扫描的时间和日期、使用的病毒定义、发现的任何病毒的数量和细节、文件的新CVE评分、任何相关段信息等。在270，该方法在扫描时将文件标记为没有已知病毒。在该实施例中，该方法向

下游数据消费者、数据用户和管理员提供CVE评分。CVE评分的接收者确定文件是否足够干净以对消费者可用。

[0066] 该方法将包含病毒的文件传递到文件净化过程。在275，该方法确定该文件是否已经被该过程净化。干净文件被传递至步骤280。包含尚未被净化的文件的病毒被传递至步骤290，并且与数据消费者使用隔离，而且不被添加到数据消费者访问。

[0067] 在280处，将干净文件添加到数据储存区作为可用于数据消费者。在步骤280之后，该方法返回至步骤210以检查新的文件/病毒定义。当没有找到新文件或定义时，该方法前进到步骤215，并且确定是否用当前病毒定义扫描了所生成的优先级列表上的所有文件。该方法返回步骤240，当尚未扫描所有文件时，选择下一个按优先级排序的文件。当已经扫描了所有文件时，该方法返回至步骤210并且继续检查新的文件和病毒定义。

[0068] 应当理解，虽然本公开包括关于云计算的详细描述，但是本文所记载的教导的实现不限于云计算环境。相反，本发明的实施例能够结合现在已知的或以后开发的任何其他类型的计算环境来实现。

[0069] 云计算是服务交付的模型，用于使得能够对可配置计算资源(例如，网络、网络带宽、服务器、处理、存储器、存储装置、应用、虚拟机和服务)的共享池进行方便地、按需地网络访问，该可配置计算资源可以以最小的管理成本或与服务提供者的交互来快速供应和释放。该云模型可以包括至少五个特性、至少三个服务模型和至少四个部署模型。

[0070] 特性如下：

[0071] 按需自助服务：云消费者可以单方面地根据需要自动地提供计算能力，诸如服务器时间和网络存储，而不需要与服务提供者进行人类交互。

[0072] 广泛的网络接入：能力可通过网络获得并且通过标准机制接入，该标准机制促进异构瘦客户机平台或厚客户机平台(例如，移动电话、膝上型计算机和PDA)的使用。

[0073] 资源池化：提供者的计算资源被池化以使用多租户模型来服务于多个消费者，其中不同的物理和虚拟资源根据需要动态地指派和重新指派。存在位置独立性的感觉，因为消费者通常不具有对所提供的资源的确切位置的控制或了解，但可能能够以较高抽象级别(例如，国家、州或数据中心)指定位置。

[0074] 快速弹性：能够快速和弹性地提供能力，在一些情况下自动地快速缩小和快速释放以快速放大。对于消费者而言，可用于供应的能力通常显得不受限制并且可以在任何时间以任何数量购买。

[0075] 经测量的服务：云系统通过在适合于服务类型(例如，存储、处理、带宽和活动用户账户)的某个抽象级别处利用计量能力来自动控制和优化资源使用。可以监视、控制和报告资源使用，从而为所利用的服务的提供者和消费者提供透明度。

[0076] 服务模型如下：

[0077] 软件即服务(SaaS)：提供给消费者的能力是使用在云基础设施上运行的提供者的应用。可通过诸如web浏览器(例如，基于web的电子邮件)之类的瘦客户端接口从不同客户端设备访问应用。消费者不管理或控制包括网络、服务器、操作系统、存储或甚至单独的应用能力的底层云基础设施，可能的例外是有限的用户特定应用配置设置。

[0078] 平台即服务(PaaS)：提供给消费者的能力是将消费者创建的或获取的使用由提供商支持的编程语言和工具创建的应用部署到云基础设施上。消费者不管理或控制包括网

络、服务器、操作系统或存储的底层云基础设施，但是对所部署的应用和可能的应用托管环境配置具有控制。

[0079] 基础设施即服务(IaaS)：提供给消费者的能力是提供处理、存储、网络和消费者能够部署和运行任意软件的其他基本计算资源，该软件可以包括操作系统和应用。消费者不管理或控制底层云基础设施，而是具有对操作系统、存储、所部署的应用的控制以及对所选联网组件(例如，主机防火墙)的可能受限的控制。

[0080] 部署模型如下：

[0081] 私有云：云基础架构仅为组织运作。它可以由组织或第三方管理，并且可以存在于场所内或场所外。

[0082] 社区云：云基础架构被若干组织共享并支持具有共同关切(例如，任务、安全性要求、策略、和合规性考虑)的特定社区。它可以由组织或第三方管理，并且可以存在于场所内或场所外。

[0083] 公共云：使云基础架构对公众或大型行业组可用，并且由出售云服务的组织拥有。

[0084] 混合云：云基础架构是两个或更多个云(私有、社区或公共)的组合，这些云保持独特的实体但通过使数据和应用能够移植的标准化或专有技术(例如，云突发以用于云之间的负载平衡)绑定在一起。

[0085] 云计算环境是面向服务的，聚焦于无状态、低耦合、模块化和语义互操作性。云计算的核心是包括互连节点网络的基础设施。

[0086] 现在参见图3，描绘了说明性云计算环境50。如图所示，云计算环境50包括云消费者使用的本地计算设备可以与其通信的一个或多个云计算节点10，本地计算设备诸如例如个人数字助理(PDA)或蜂窝电话54A、台式计算机54B、膝上型计算机54C和/或汽车计算机系统54N。节点10可彼此通信。它们可以物理地或虚拟地分组(未示出)在一个或多个网络中，诸如如上所述的私有云、社区云、公共云或混合云、或其组合。这允许云计算环境50提供基础设施、平台和/或软件作为服务，云消费者不需要为该服务维护本地计算设备上的资源。应当理解，图3中所示的计算设备54A-N的类型仅旨在是说明性的，并且计算节点10和云计算环境50可通过任何类型的网络和/或网络可寻址连接(例如，使用网络浏览器)与任何类型的计算机化设备通信。

[0087] 现在参见图4，示出了由云计算环境50(图3)提供的一组功能抽象层。应事先理解，图4中所示的组件、层和功能仅旨在是说明性的，并且本发明的实施例不限于此。如所描述，提供以下层和对应功能：

[0088] 硬件和软件层60包括硬件和软件组件。硬件组件的示例包括：大型机61；基于RISC(精简指令集计算机)架构的服务器62；服务器63；刀片服务器64；存储设备65；以及网络和联网组件66。在一些实施例中，软件组件包括网络应用服务器软件67和数据库软件68。

[0089] 虚拟层70提供抽象层，从该抽象层可以提供虚拟实体的以下示例：虚拟服务器71；虚拟存储器72；虚拟网络73，包括虚拟专用网络；虚拟应用和操作系统74；以及虚拟客户端75。

[0090] 在一个示例中，管理层80可以提供以下描述的功能。资源供应81提供用于在云计算环境内执行任务的计算资源和其他资源的动态采购。计量和定价82在云计算环境内利用资源时提供成本跟踪，并为这些资源的消费开账单或发票。在一个示例中，这些资源可以包

括应用软件许可证。安全性为云消费者和任务提供身份验证，以及为数据和其他资源提供保护。用户门户83为消费者和系统管理员提供对云计算环境的访问。服务级别管理84提供云计算资源分配和管理，使得满足所需的服务水平。服务级别协议(SLA)规划和履行85提供根据SLA预期未来需求的云计算资源的预安排和采购。

[0091] 工作负载层90提供可以利用云计算环境的功能的示例。可以从该层提供的工作负载和功能的示例包括：地图绘制和导航91；软件开发和生命周期管理92；虚拟教室教育交付93；数据分析处理94；事务处理95；以及病毒扫描程序175。

[0092] 本发明可以是处于任何可能的技术细节集成度的系统、方法和/或计算机程序产品。本发明可以在处理指令流的任何系统(单个或并行)中有益地实践。计算机程序产品可包括其上具有用于使处理器执行本发明的各方面的计算机可读程序指令的计算机可读存储介质(或多个介质)。

[0093] 计算机可读存储介质可为可保留和存储供指令执行装置使用的指令的有形装置。计算机可读存储介质可以是，例如但不限于，电子存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备、或者上述的任意合适的组合。计算机可读存储介质的更具体示例的非穷尽列表包括以下各项：便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦可编程只读存储器(EPROM或闪存)、静态随机存取存储器(SRAM)、便携式紧凑盘只读存储器(CD-ROM)、数字通用盘(DVD)、记忆棒、软盘、诸如穿孔卡或槽中的凸出结构之类的具有记录在其上的指令的机械编码设备、以及上述各项的任何合适的组合。如本文所使用的计算机可读存储介质不应被解释为暂时性信号本身，诸如无线电波或其他自由传播的电磁波、通过波导或其他传输介质传播的电磁波(例如，穿过光纤光缆的光脉冲)或通过电线传输的电信号。

[0094] 本文中所描述的计算机可读程序指令可以经由网络(例如，互联网、局域网、广域网和/或无线网络)从计算机可读存储介质下载到相应的计算/处理设备，或者下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光传输光纤、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配器卡或网络接口接收来自网络的计算机可读程序指令，并转发计算机可读程序指令以存储在相应计算/处理设备内的计算机可读存储介质中。

[0095] 用于执行本发明的操作的计算机可读程序指令可以是汇编指令、指令集架构(ISA)指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、集成电路的配置数据、或以一种或多种程序设计语言的任何组合编写的源代码或目标代码，这些程序设计语言包括面向对象的程序设计语言(诸如Smalltalk、C++等)和过程性程序设计语言(诸如“C”程序设计语言或类似程序设计语言)。计算机可读程序指令可以作为独立软件包执行完全地在用户计算机上执行或者部分在用户计算机上执行、部分在用户计算机上部分在远程计算机上执行或者完全在远程计算机或服务器上执行。在后一种情况下，远程计算机可通过任何类型的网络(包括局域网(LAN)或广域网(WAN))连接至用户计算机，或者可连接至外部计算机(例如，使用互联网服务提供商通过互联网)。在一些实施例中，包括例如可编程逻辑电路、现场可编程门阵列(FPGA)或可编程逻辑阵列(PLA)的电子电路可以通过利用计算机可读程序指令的状态信息来使电子电路个性化来执行计算机可读程序指令，以便执行本发明的各方面。

[0096] 下面将参照根据本发明实施例的方法、装置(系统)和计算机程序产品的流程图和/或框图描述本发明的各方面。将会理解到,流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合,都可以由计算机可读程序指令实现。

[0097] 这些计算机可读程序指令可被提供给通用计算机、专用计算机或其他可编程数据处理装置的处理器以产生机器,使得经由计算机或其他可编程数据处理装置的处理器执行的指令创建用于实现在流程图和/或框图的一个或多个框中指定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中,这些指令使得计算机、可编程数据处理装置、和/或其他设备以特定方式工作,从而,其中存储有指令的计算机可读存储介质包括包含实现流程图和/或框图中的一个或多个方框中规定的功能/动作的方面的指令的制造品。

[0098] 也可以把计算机可读程序指令加载到计算机、其他可编程数据处理装置、或其他设备上,使得在计算机、其他可编程装置或其他设备上执行一系列操作步骤,以产生计算机实现的处理,使得在计算机、其他可编程装置或其他设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

[0099] 附图中的流程图和框图示出了根据本发明的不同实施例的系统、方法和计算机程序产品的可能实现方式的架构、功能和操作。对此,流程图或框图中的每个方框可表示指令的模块、段或部分,其包括用于实现指定的逻辑功能的一个或多个可执行指令。在一些备选实现方式中,方框中标注的功能可以不按照附图中标注的顺序发生。例如,取决于所涉及的功能,连续示出的两个方框实际上可以基本上同时执行,或者这些方框有时可以以相反的顺序执行。也要注意的是,框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合,可以用执行规定的功能或动作或执行专用硬件与计算机指令的组合的专用的基于硬件的系统来实现。

[0100] 说明书中对“一个实施例”、“实施例”、“示例性实施例”等的引用表示所描述的实施例可以包括特定的特征、结构或特性,但是每个实施例可能不一定包括特定的特征、结构或特性。此外,这样的短语不一定指相同的实施例。进一步,当结合实施例描述特定特征、结构或特性时,认为结合其他实施例(无论是否明确描述)影响这样的特征、结构或特性在本领域技术人员的知识范围内。

[0101] 本文中使用的术语仅用于描述具体实施例的目的,而并非旨在限制本发明。如本文所使用的,除非上下文另有明确指示,否则单数形式“一个”、“一种”和“该”旨在也包括复数形式。还应当理解,当在本说明书中使用术语“包括”和/或“包含”时,其指定所述特征、整体、步骤、操作、元件和/或部件的存在,但不排除一个或多个其他特征、整体、步骤、操作、元件、部件和/或其组合的存在或添加。

[0102] 已经出于说明的目的呈现了本发明的各种实施例的描述,但并不旨在是详尽的或者限于所公开的实施例。在不背离本发明的范围的情况下,许多修改和变化对于本领域普通技术人员来说是显而易见的。在此所使用的术语被选择来最好地解释实施例的原理、实际应用、或优于市场中所发现的技术的技术改进、或使得本领域普通技术人员能够理解在此所披露的实施例。

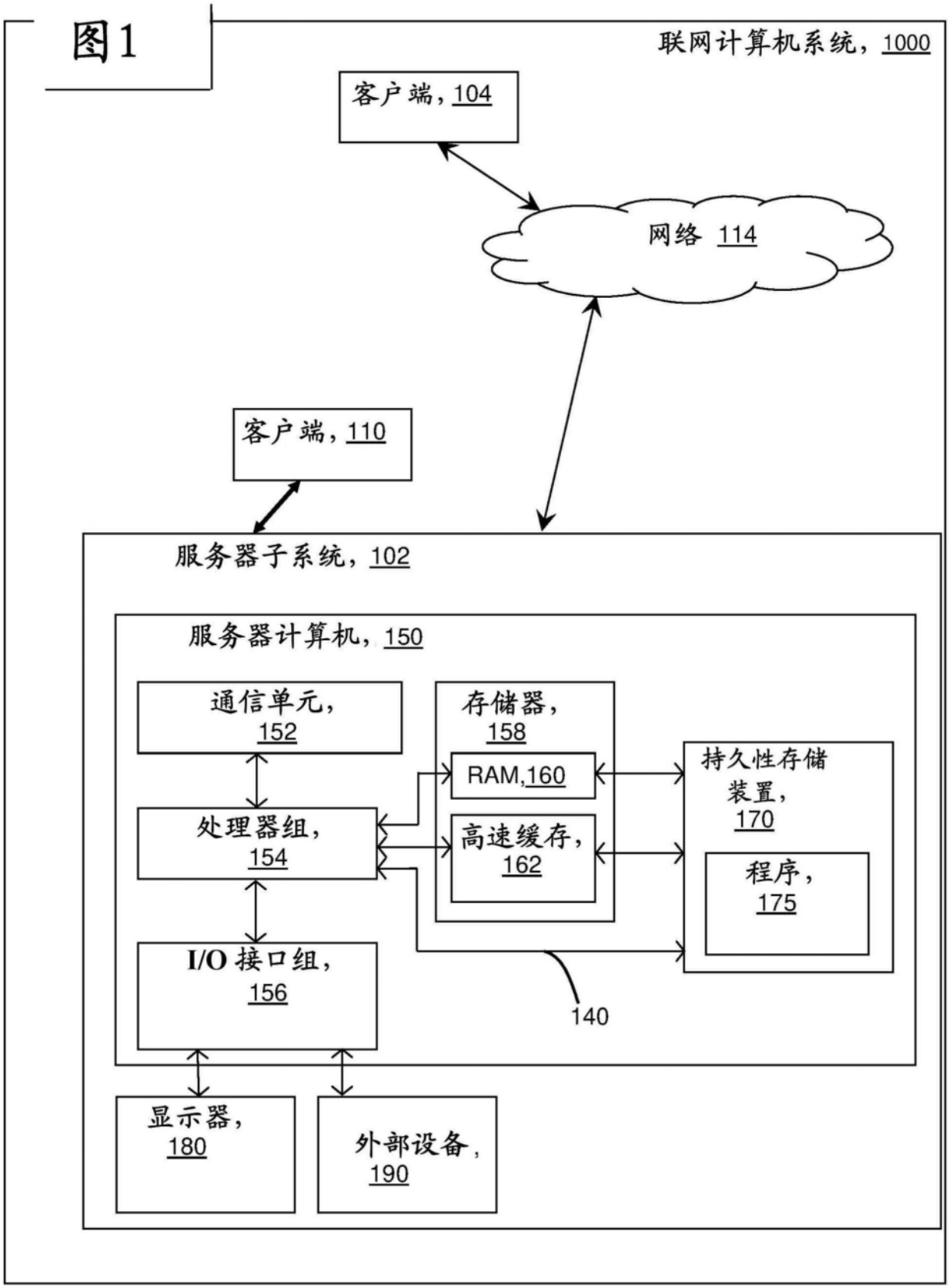

图1
联网计算机系统，1000
客户端，104
网络 114
客户端，110
服务器子系统，102
服务器计算机，150
通信单元，152
存储器，158
RAM,160
高速缓存，162
持久性存储装置，170
程序，175
处理器组，154
I/O 接口组，156
140
显示器，180
外部设备，190

图1

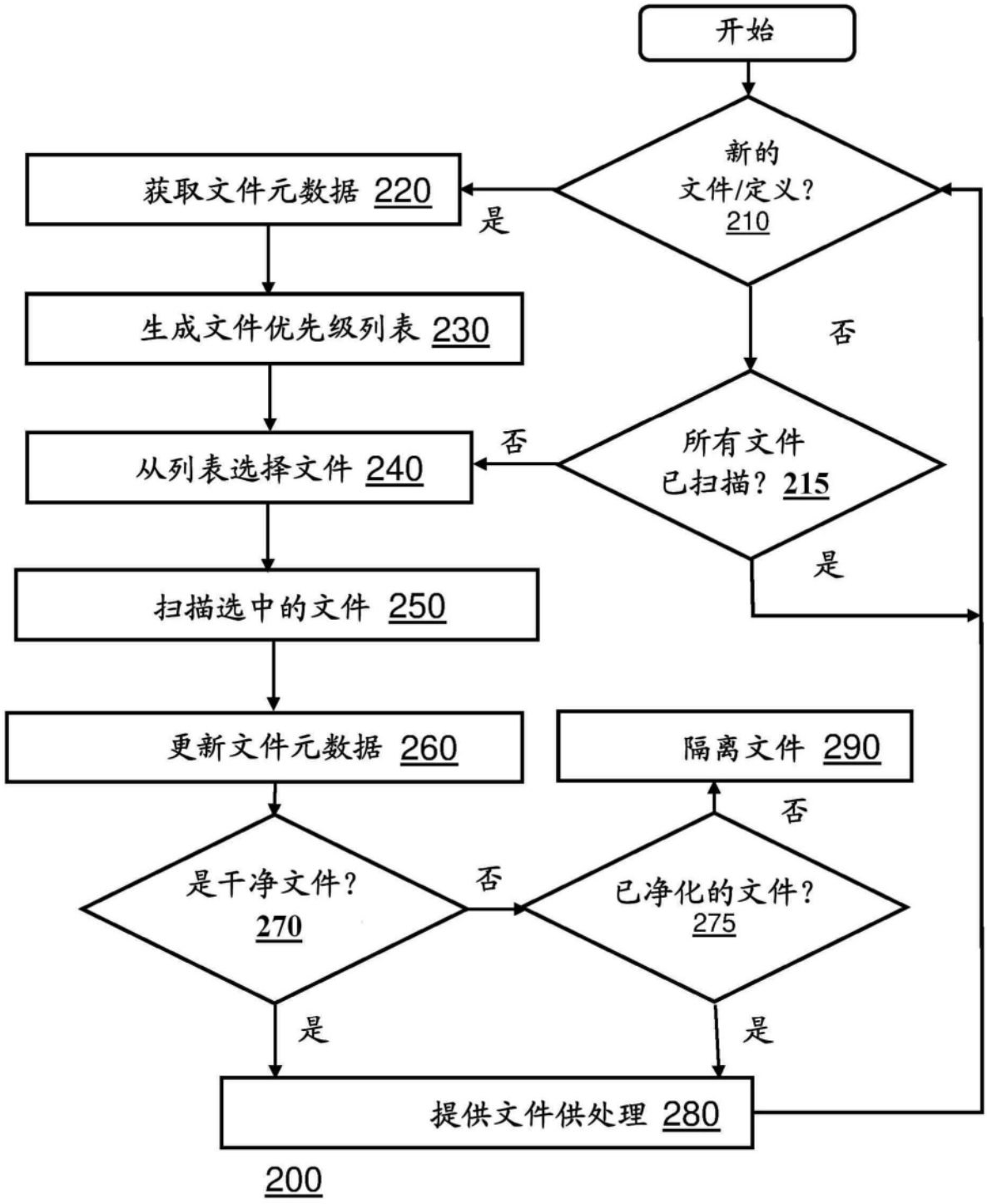

开始
新的
文件/定义?
210
是
获取文件元数据 220
否
生成文件优先级列表 230
所有文件
已扫描? 215
否
从列表选择文件 240
是
扫描选中的文件 250
更新文件元数据 260
隔离文件 290
否
是干净文件?
270
否
已净化的文件?
275
是
是
提供文件供处理 280
200

图2

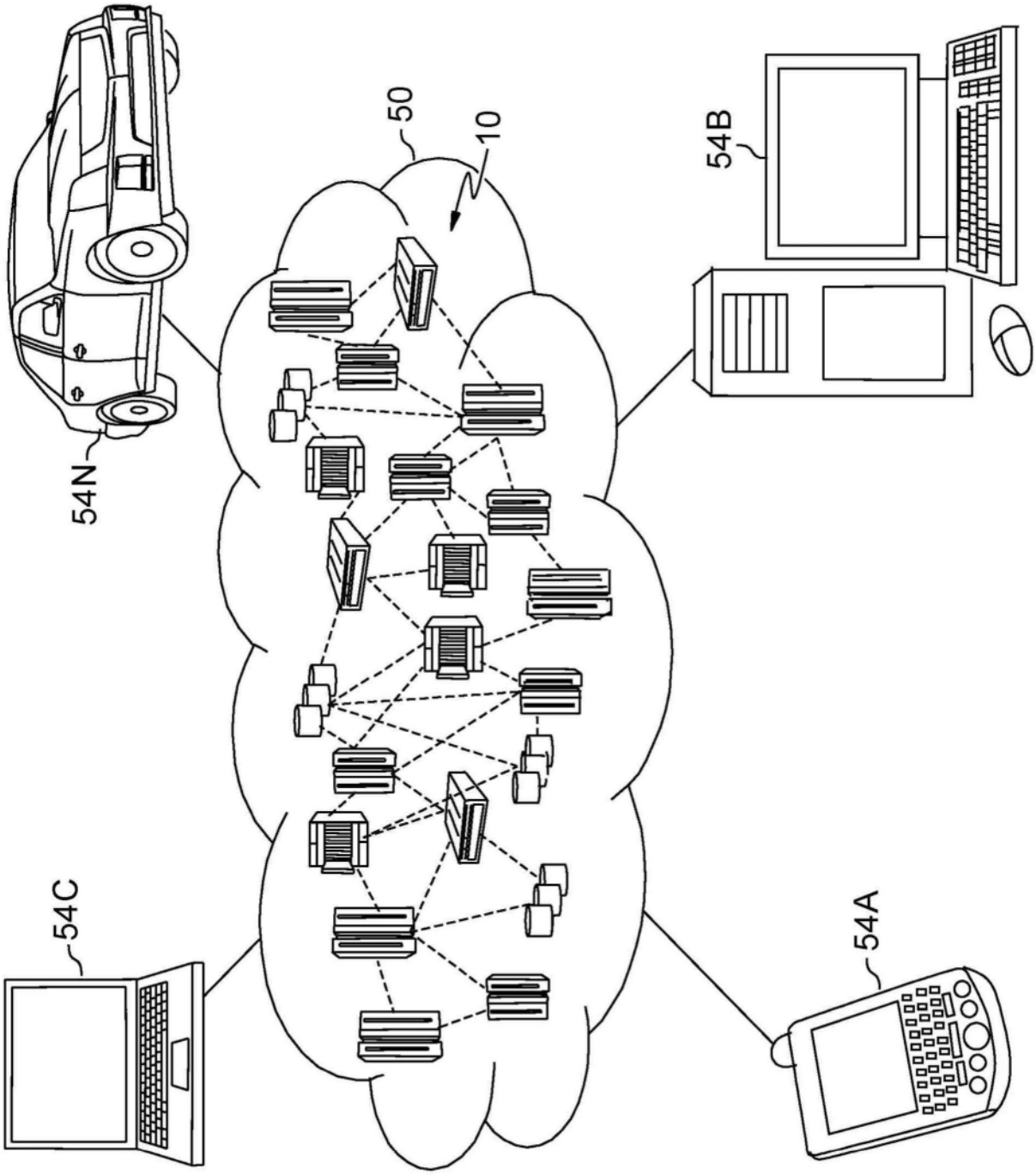


图3

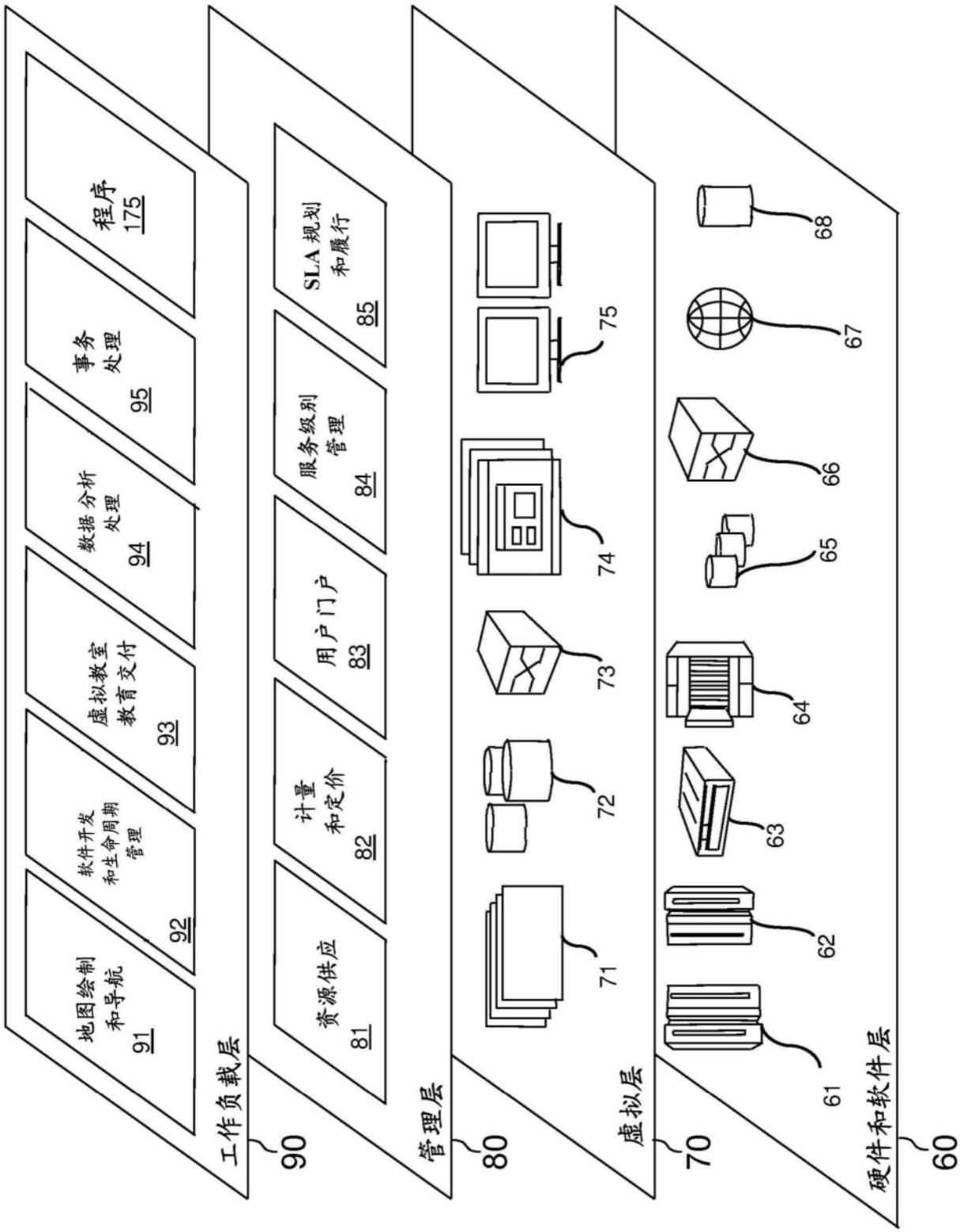

工作负载层
90
地图绘制和导航
91
软件开发和生命周期管理
92
虚拟教室教育交付
93
数据分析处理
94
事务处理
95
程序
175
管理层
80
资源供应
81
计量和定价
82
用户门户
83
服务级别管理
84
SLA规划和履行
85
虚拟层
70
硬件和软件层
60
61
62
63
64
65
66
67
68
71
72
73
74
75

图4