(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 102460398 A
(43)申请公布日 2012.05.16

(21)申请号 201080024491.3

(22)申请日 2010.05.26

(30)优先权数据
12/480587 2009.06.08 US

(85)PCT申请进入国家阶段日
2011.12.02

(86)PCT申请的申请数据
PCT/US2010/036260 2010.05.26

(87)PCT申请的公布数据
WO2010/144260 EN 2010.12.16

(71)申请人 赛门铁克公司
地址 美国加利福尼亚州

(72)发明人 M·蔡司 W·吴

(74)专利代理机构 中原信达知识产权代理有限责任公司 11219
代理人 周亚荣 安翔

(51)Int.Cl.
G06F 11/14(2006.01)

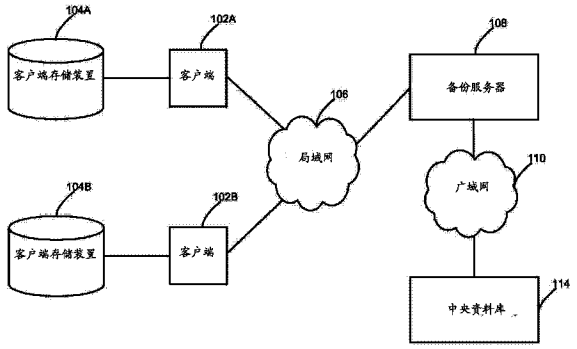
权利要求书 3 页 说明书 8 页 附图 6 页

(54)发明名称

用于在备份操作中执行去重复的源分类

(57)摘要

在此披露了一种用于使用去重复将数据从备份源备份到中央资料库的计算机实施的系统、方法以及计算机程序产品，其中该数据包括多个源数据段。接收一个指纹缓存，该指纹缓存包括存储在该中央资料库中的多个数据段的多个指纹，其中这些数据段是先前从该备份源备份的。生成多个源数据指纹，这些源数据指纹包括这些源数据段的多个指纹（例如多个散列值）。将这些源数据指纹与该指纹缓存中的这些指纹进行比较。与未在该指纹缓存中的多个指纹相对应的这些源数据段当前可以未存储在该中央资料库中。在进一步查询了该中央资料库后，响应于该比较而将这些源数据段中的一个或多个发送到该中央资料库用于存储。



CN 102460398 A

1. 一种用于使用去重复将数据从备份源备份到中央资料库的计算机实施的方法，该数据包括多个源数据段，该方法包括：

接收一个指纹缓存，该指纹缓存包括存储在该中央资料库中的多个数据段的多个指纹，这些数据段是先前从该备份源中备份的；

生成多个源数据指纹，这些源数据指纹包括这些源数据段的多个指纹；

将这些源数据指纹与该指纹缓存中的这些指纹进行比较；并且

响应于该比较，将这些源数据段中的一个或多个发送到该中央资料库用于存储。

2. 如权利要求 1 所述的计算机实施的方法，其中将这些源数据指纹与该指纹缓存中的这些指纹进行比较包括：

对未包含在该指纹缓存中的一个源数据指纹进行识别。

3. 如权利要求 2 所述的计算机实施的方法，其中响应于该比较而将这些源数据段中的一个或多个发送到该中央资料库用于存储进一步包括：

查询带有所识别的源数据指纹的中央资料库，以确定与所识别的源数据指纹相对应的一个数据段是否被存储在该中央资料库中；并且

响应于与所识别的源数据指纹相对应的一个数据段未存储在该中央资料库中，将与该源数据指纹相对应的一个数据段发送到该中央资料库用于存储。

4. 如权利要求 1 所述的计算机实施的方法，其中一个数据段的一个指纹包括该数据段的一个散列值。

5. 如权利要求 1 所述的计算机实施的方法，其中接收了一系列指纹缓存，该系列指纹缓存中的每一个均包括先前从该备份源备份的多个数据段的指纹的一部分，并且其中对这些源数据指纹的不同部分与该系列指纹缓存中的每一个进行比较。

6. 如权利要求 1 所述的计算机实施的方法，进一步包括：

将该指纹缓存中的、未在这些源数据指纹中找到的指纹发送到该中央资料库用于对该中央资料库的一个数据库进行更新。

7. 如权利要求 1 所述的计算机实施的方法，其中该指纹缓存包括先前使用一种指定的备份策略而备份的多个数据段的多个指纹。

8. 一种用于使用去重复将数据从备份源备份到中央资料库的计算机系统，该数据包括多个源数据段，该系统包括：

一种存储多个可执行的计算机程序模块的计算机可读存储媒质，用于：

接收一个指纹缓存，该指纹缓存包括存储在该中央资料库中的多个数据段的多个指纹，这些数据段是先前从该备份源备份的；

生成多个源数据指纹，这些源数据指纹包括这些源数据段的多个指纹；

将这些源数据指纹与该指纹缓存中的这些指纹进行比较；并且

响应于该比较而将这些源数据段中的一个或多个发送到该中央资料库用于存储。

9. 如权利要求 8 所述的计算机系统，其中将这些源数据指纹与该指纹缓存中的这些指纹进行比较包括：

对未包含在该指纹缓存中的一个源数据指纹进行识别。

10. 如权利要求 9 所述的计算机系统，其中响应于该比较而将这些源数据段中的一个或多个发送到该中央资料库用于存储进一步包括：

查询带有所识别的源数据指纹的中央资料库，以确定与所识别的源数据指纹相对应的一个数据段是否被存储在该中央资料库中；并且

响应于与所识别的源数据指纹相对应的一个数据段未存储在该中央资料库中，将与该源数据指纹相对应的一个数据段发送到该中央资料库用于存储。

11. 如权利要求 8 所述的计算机系统，其中一个数据段的一个指纹包括该数据段的一个散列值。

12. 如权利要求 8 所述的计算机系统，其中接收了一系列指纹缓存，该系列指纹缓存中的每一个均包括先前从该备份源备份的多个数据段的指纹的一部分，并且其中对这些源数据指纹的不同部分与该系列指纹缓存中的每一个进行比较。

13. 如权利要求 8 所述的计算机系统，其中这些计算机程序模块进一步被配置用于：

将该指纹缓存中的、未在这些源数据指纹中找到的指纹发送到该中央资料库用于对该中央资料库的一个数据库进行更新。

14. 如权利要求 8 所述的计算机系统，其中该指纹缓存包括先前使用一种指定的备份策略而备份的多个数据段的多个指纹。

15. 一种用于使用去重复将数据从备份源备份到中央资料库的具有一种计算机可读存储媒质的计算机程序产品，该计算机可读存储媒质具有多个可执行的计算机程序指令，该数据包括多个源数据段，这些计算机程序指令进一步被配置用于：

接收一个指纹缓存，该指纹缓存包括存储在该中央资料库中的多个数据段的多个指纹，这些数据段是先前从该备份源备份的；

生成多个源数据指纹，这些源数据指纹包括这些源数据段的多个指纹；

将这些源数据指纹与该指纹缓存中的这些指纹进行比较；并且

响应于该比较而将这些源数据段中的一个或多个发送到该中央资料库用于存储。

16. 如权利要求 15 所述的计算机程序产品，其中将这些源数据指纹与该指纹缓存中的这些指纹进行比较包括：

对未包含在该指纹缓存中的一个源数据指纹进行识别。

17. 如权利要求 16 所述的计算机程序产品，其中响应于该比较而将这些源数据段中的一个或多个发送到该中央资料库用于存储进一步包括：

查询带有所识别的源数据指纹的中央资料库，以确定与所识别的源数据指纹相对应的一个数据段是否被存储在该中央资料库中；并且

响应于与所识别的源数据指纹相对应的一个数据段未存储在该中央资料库中，将与该源数据指纹相对应的一个数据段发送到该中央资料库用于存储。

18. 如权利要求 15 所述的计算机程序产品，其中接收了一系列指纹缓存，该系列指纹缓存中的每一个均包括先前从该备份源备份的多个数据段的指纹的一部分，并且其中对这些源数据指纹的不同部分与该系列指纹缓存中的每一个进行比较。

19. 如权利要求 15 所述的计算机程序产品，其中这些计算机程序指令进一步被配置用于：

将该指纹缓存中的、未在这些源数据指纹中找到的指纹发送到该中央资料库用于对该中央资料库的一个数据库进行更新。

20. 如权利要求 15 所述的计算机程序产品，其中该指纹缓存包括先前使用一种指定的

备份策略而备份的多个数据段的多个指纹。

用于在备份操作中执行去重复的源分类

技术领域

[0001] 本发明总体上涉及对数字数据进行备份。

背景技术

[0002] 在一种包括若干客户端计算机的环境中，诸如企业局域网 (LAN)，常采用一种集中式备份系统。该集中式备份系统由系统管理员来配置以便自动地在这些客户端计算机的存储装置上对数据进行备份。该集中式备份系统可以包括一个备份服务器，该备份服务器通过将数据段（例如多个文件或文件的多个部分）从这些客户端计算机的存储装置拷贝到中央资料库来周期性地对每个客户端计算机进行备份。这个中央资料库具有足够大的存储容量并且通常被定位于远离该备份服务器（以及客户端计算机）。其结果是，常需要在一段显著的距离上通过一个可以受限的带宽链路来传输有待备份的数据段。

[0003] 一种减少传输到并存储在中央资料库中的数据量的技术被称为“去重复”。采用去重复的备份系统利用了以下事实：备份期间在一个客户端计算机中找到的一个数据段可以已经被存储在中央资料库中。因为在前一次备份期间它已从客户端计算机中备份并且从那时起在该客户端计算机上的数据段尚未被修改，所以它可以已经在中央资料库中。因为先前已从具有相同数据段的另一个客户端计算机中备份过该数据段，所以它也可以已经在中央资料库中。

[0004] 当使用去重复从一个客户端计算机中备份一个数据段时，备份服务器生成了对该数据段进行识别的信息。备份服务器将这个识别信息传输到中央资料库并且中央资料库发送一个响应来表明它是否已经包含所识别的数据段。只有该响应表明该数据段尚未包含在资料库中，备份服务器然后才将实际的数据段传输到中央资料库。其结果是，减少了传输到中央资料库的数据段的数目。

[0005] 然而，因为备份服务器要为每个数据段将识别信息传输到中央资料库，所以在备份中仍有涉及大量使用计算和网络资源。同样，对于每个数据段中央资料库都要使用计算资源来生成一个响应并且中央资料库将该响应传输到备份服务器。因此，在本领域中对于使用去重复来减少备份所需要的计算和网络资源的方法存在一种需要。

发明内容

[0006] 上述需要可以由一种用于使用去重复将数据从备份源备份到中央资料库的计算机实施的系统、方法以及计算机程序产品来满足，其中该数据包括多个源数据段。在一个实施方案中，接收一个指纹缓存，该指纹缓存包括存储在该中央资料库中的多个数据段的多个指纹，其中这些数据段是先前从该备份源备份的。生成多个源数据指纹，这些源数据指纹包括这些源数据段的多个指纹（例如多个散列值）。将这些源数据指纹与该指纹缓存中的这些指纹进行比较。与未在该指纹缓存中的多个指纹相对应的这些源数据段当前可以未存储在该中央资料库中。在进一步查询了该中央资料库后，响应于该比较而将这些源数据段中的一个或多个发送到该中央资料库用于存储。

附图说明

[0007] 图 1 是一个高级图，展示了在一个实施方案中的包括用于使用去重复来有效地对多个客户端进行备份的备份服务器和中央资料库的一种网络环境。

[0008] 图 2 是一个框图，展示了在一个实施方案中的能够作为客户端、备份服务器或中央资料库的实施方案的一种计算机。

[0009] 图 3 展示了在一个实施方案中的用于对多个客户端存储装置进行备份的的备份服务器的一个逻辑视图。

[0010] 图 4 展示了在一个实施方案中的用于对从多个客户端存储装置备份的数据段进行存储的中央资料库的一个逻辑视图。

[0011] 图 5 是一个流程图，展示了在一个实施方案中的用于使用去重复将数据从一个客户端存储装置备份到中央资料库的一种方法。

[0012] 图 6 是一个流程图，展示了在一个实施方案中的用于使用去重复在中央资料库对数据进行备份的一种方法。

[0013] 这些图示仅为展示的目的而描绘了一种实施方案。本领域的普通技术人员将容易地从以下说明中认识到，可以使用在此所展示的这些结构以及方法的替代实施方案而不背离在此说明的原理。

具体实施方式

[0014] 图 1 是一个高级图，展示了在一个实施方案中的包括用于使用去重复来有效地对多个客户端 102 进行备份的备份服务器 108 和中央资料库 114 的网络环境 100。备份服务器 108 与客户端 102 由局域网 106 来连接。虽然图中仅示出了两个客户端 102，但在网络 106 中可以有若干个。备份服务器 108 还可以通过的广域网 110（诸如互联网）而连接到中央资料库 114 上。相比广域网 110 而言，局域网 106 允许更高的带宽以及更便宜的通信。在一个实施方案中，备份服务器 108 与多个客户端 102 的定位在物理上非常接近，而中央资料库 114 定位在其他地方。在一个实施方案中，多个备份服务器 108 与中央资料库 114 进行通信。

[0015] 这些客户端 102 是不同的计算装置，诸如局域网 106 上的用户工作站、服务器（例如网络或应用服务器）或路由器。这些客户端 102 具有多个客户端存储装置 104，诸如硬盘驱动器。在一个实施方案中（未示出），一些客户端存储装置 104 被直接连接到网络 106 上。这些客户端存储装置 104 包含可以在一个文件系统中被存储为多个文件的数据。令人希望的是周期性地对来自客户端存储装置 104 的数据进行备份以防数据丢失。某些类型的文件（诸如操作系统文件）极少随着时间的推移而变化并且它们在环境 100 中的多个客户端存储装置 104 上作为完全相同的副本而存在。一个客户端 102 或客户端存储装置 104 也被称为一个备份源。

[0016] 备份服务器 108 周期性地（例如每天）在客户端存储装置 104 上进行不同类型的数据备份。有待从一个特定的客户端存储装置 104 备份的数据可以作为一个流（也被称为一个备份流）而由备份服务器 108 在局域网 106 上访问。这个备份流可以由备份服务器 108 划分为多个数据段，其中一个数据段（例如）对应于一个文件、一个文件的一部分或者一个

磁盘扇区。如果数据段尚未存储在中央资料库中，那么备份服务器 108 采用多种去重复技术将这些数据段发送到该中央资料库 114 中用于存储。备份服务器 108 通过对当前存储在该中央资料库中的可以在该当前备份流中遇到的多个数据段进行识别而从中央资料库 114 中检索一个指纹缓存以完成此动作。备份服务器 108 使用这个指纹缓存来确定在该当前备份流中的哪些数据段应被发送到中央资料库 114 中，而无须针对每个数据段来查询该中央资料库。

[0017] 中央资料库 114 在备份操作期间对从备份服务器 108 中接收的多个数据段的副本进行存储。中央资料库 114 可以对最初来自数千个客户端存储装置 104 的数百万个数据段进行存储。在必须将所备份的数据恢复到客户端存储装置 104 中的情况下可以访问存储在中央资料库 114 中的这些数据段。因为该备份操作采用了去重复，所以在中央资料库 114 中的一个单一的数据段当前可以存在于多个客户端存储装置 104 上。为了支持使用去重复的备份操作，中央资料库 114 将关于哪些数据段当前被存储在该资料库中的信息提供给备份服务器 108。

[0018] 图 2 是根据一个实施方案的用作客户端 102、备份服务器 108、和 / 或中央资料库 114 的计算机 200 的高级框图。所展示的是至少一个处理器 202 连接到芯片组 204 上。同样连接到芯片组 204 上的是存储器 206、存储装置 208、键盘 210、图形适配器 212、定点装置 214 以及网络适配器 216。显示器 218 连接到图形适配器 212 上。在一个实施方案中，由存储器控制器集线器 220 与 I/O 控制器集线器 222 来提供芯片组 204 的功能。在另一个实施方案中，存储器 206 直接连接到处理器 202 上而不是芯片组 204 上。

[0019] 存储装置 208 是任何计算机可读存储媒质，诸如硬盘驱动器、光盘只读存储器 (CD-ROM)、DVD 或者固态存储器装置。在一个实施方案中，客户端 102 的存储装置 208 是客户端存储装置 104。存储器 206 保存由处理器 202 使用的指令与数据。定点装置 214 可以是鼠标、轨迹球、或其他类型的定点装置，并且与键盘 210 一起使用以便将数据输入计算机系统 200 中。图形适配器 212 在显示器 218 上显示图像和其他信息。网络适配器 216 将计算机系统 200 连接到一个局域或广域网上。

[0020] 如在本领域中所知的，计算机 200 可具有不同于图 2 所示的部件和 / 或其他部件。另外，计算机 200 可以缺少所展示的某些部件。在一个实施方案中，用作备份服务器 108 的计算机 200 缺少键盘 210、定点装置 214、图形适配器 212、和 / 或显示器 218。此外，存储装置 208 可以在定位于计算机 200 中和 / 或使其与之远离（诸如实施在一个存储区域网络 (SAN) 内）。

[0021] 如本领域中所知的，计算机 200 被适配为执行多个计算机程序模块用于提供在此所述的功能。如在此所使用的，术语“模块”是指用来提供特定功能的计算机程序逻辑。这样，模块可以在硬件、固件和 / 或软件中实施。在一个实施方案中，多个程序模块存储在存储装置 208 上，加载到存储器 206 中，并且由处理器 202 来执行。

[0022] 在此所述的这些实体的实施方案可以包括其他的和 / 或与在此所述不同的模块。另外，归因于这些模块的功能可以在其他的实施方案中由其他的或不同的模块来执行。而且，为了明晰与方便的目的，这种描述偶尔省略了术语“模块”。

[0023] 图 3 展示了在一个实施方案中的用于对多个客户端存储装置 104 进行备份的备份服务器 108 的一个逻辑视图。备份服务器 108 基于多种备份规范 310 来进行备份。备份规

范 310 包括该备份源的标识，诸如网络协议 (IP) 地址或客户端 102 的主机名称。备份规范 310 还可以包括表明该备份是否是全备份或增量备份的备份类型。备份规范 310 还可以包括进一步限定关于该备份的细节的备份策略。该备份策略可以表明有待备份的文件类型（例如，文本文件或不可执行文件）。它可以表明应该对该文件系统中某些位置处（例如，在某些目录中）的文件进行备份。该备份策略可以指定该备份是原始格式备份（例如，其中数据段对应于磁盘扇区而不是文件）。

[0024] 备份规范 310 可以由系统管理员输入到备份服务器 108 中并且可以被存储在该备份服务器上。备份服务器 108 可以被配置为根据多种备份规范 310 来进行多个备份，以便对多个客户端 102 进行备份和 / 或在一个特定的客户端上进行不同类型的备份。备份服务器 108 还可以被配置为以一个指定的频率（例如，每天）根据每种备份规范 310 来进行备份。备份规范 310 也被称为源分类，因为它对所备份的数据的源进行分类。

[0025] 备份模块 304 在一个实施方案中被配置为基于备份规范 310 使用去重复来进行一次备份。备份模块 304 从客户端存储装置 104 创建一个备份流，其中该客户端存储装置（或相关联的客户端 102）是在备份规范 310 中识别的备份源。备份模块 304 将该备份流分割为多个数据段。该备份流的内容以及该备份流的分割是基于在备份规范 310 中的备份类型和备份策略。该备份流可以是不受限制的并且包括任何数据量并且具有任何数目的段。

[0026] 指纹模块 302 创建一个当前指纹集 312，该指纹集包含该备份流的数据段的多个指纹。一个数据段的一个指纹是一片数据，该片数据唯一地标识该数据段但是它显著地小于该数据段本身。在一个实施方案中，一个数据段的一个指纹是基于该数据段的内容的一个散列的值。可以通过将一个散列函数（诸如消息摘要算法 5(MD5)）应用到该数据段上而从一个数据段中生成一个指纹。

[0027] 资料库通信模块 308 从中央资料库 114 ＝请求一个指纹缓存。该请求包括当前备份规范 310，该规范可以包括备份源、类型和策略。中央资料库 114 可以通过提供由资料库通信模块 308 接收的一个指纹缓存 306 来响应。在指纹缓存 306 中的全部指纹对应于当前存储在中央资料库 114 中的多个数据段。

[0028] 指纹缓存 306 是存储在中央资料库 114 上的指纹（可能是数百万计的）的一个子集。如以下在中央资料库的描述中所进一步讨论的，指纹缓存 306 是由中央资料库 114 基于当前的备份规范 310 选择的。简言之，使用与在当前备份中相同或相似的备份规范 310 来选择指纹缓存 306 以包含与在先前的备份中遇到的多个数据段相对应的多个指纹。其结果是，所接收的指纹缓存 306 可以对应于在该当前备份中的这些数据段。如果备份规范 310 对于中央资料库 114 而言是未知的，那么它可以用一个空指纹缓存 306 来响应。

[0029] 段处理模块 314 对在当前指纹集 312 中的和在指纹缓存 306 中的指纹进行比较。如果来自当前指纹集 312 的一个指纹也出现在指纹缓存 306 中，那么对应于该指纹的数据段已经存储在中央资料库 114 中并且不需要被发送到那里。该数据段从使用相同备份规范的前一个备份起可以一直是未改变的（例如，在客户端存储装置 104 上的从该客户端存储装置的上次备份起就未改变的一个文件）。

[0030] 如果来自当前指纹集 312 的一个指纹未出现在指纹缓存 306 中，那么仍有坑的是对应于该指纹的数据段已存储在中央资料库 114 中。虽然先前未使用当前备份规范 310 对该数据段进行备份，但是可以先前已经使用不同的备份规范对其进行了备份（例如，从一

个不同的客户端存储装置中被备份）。

[0031] 段处理模块314致使资料库通信模块308将多个请求发送到中央资料库114，这些请求包含来自当前指纹集312的、未出现在指纹缓存306中的多个指纹。在一个实施方案中，包含全部这些指纹的一个单一请求被发送到中央资料库114以降低网络带宽。不管相关的备份规范，中央资料库114进行一次搜索来确定它是否包含与这些指纹相对应的数据段。

[0032] 中央资料库114发送一个响应（或多个响应）来表明哪些指纹对应于当前存储在该中央资料库中的数据段。如果该响应表明一个指纹已存储在该中央资料库中，那么无需由备份服务器108对该指纹做进一步处理。然而，如果该响应表明一个指纹未存储在该中央资料库114中，那么资料库通信模块308将与该指纹相对应的数据段发送到该中央资料库用于存储。

[0033] 由于使用了指纹缓存306用于去重复，通常关于多个单独指纹的仅有较小数目的请求从备份服务器108中被发送到中央资料库114。指纹缓存306常常包含当前指纹集312中的大部分指纹。这是因为对于一个给定的备份规范310（例如，一个给定的客户端与备份类型），包括一个备份的数据段从一个备份到下一个典型地不会变化太大，尤其是如果是时常进行备份的话。如上所述的备份服务器108利用了这一备份历史来减少指纹请求的数目。跨广域网110的网络带宽也可以被减少。使用指纹缓存306可以减少到中央资料库114的关于多个单独指纹的请求的发送以及来自该中央资料库的多个响应的接收。另外地，在备份服务器108和中央资料库114上用于生成请求、响应请求以及处理响应的处理资源可以被减少。

[0034] 如果在从中央资料库114接收的指纹缓存306中的一个指纹不在当前指纹集312中，那么该指纹可以对应于不再与当前备份规范310相关的一个数据段。例如，从客户端存储装置104的上次备份起一个文件可以已经从该客户端存储装置中被删除。在一个实施方案中，段处理模块314使资料库通信模块308发送一个通知报文到中央资料库114，该通知报文识别在指纹缓存306中的但不在当前指纹集312的多个指纹。该报文可以由中央资料库114使用来更新其多个数据库并且删除不再需要的已存储的多个数据段。

[0035] 在这些指纹与指纹缓存306中的指纹进行比较之前，指纹模块302并不一定要创建一个包含在该备份流中的全部数据段的指纹的当前指纹集312。如上面所提及的，该备份流是不受限制的并且可以具有很多数据段。一个基于流的处理途径可以被使用，其中当这些数据段从客户端102被接收时由指纹模块302创建这些数据段的指纹。这些指纹可被传输到段处理模块314中是因为它们被创建用于与指纹缓存306的比较。当有需要时关于多个特定指纹的请求可被发送到中央资料库114中。在一个实施方案中，在多个数据段的基于流的处理开始之前从中央资料库114中接收指纹缓存306。

[0036] 在一个实施方案中，对应于一种备份规范310的一系列指纹缓存306响应于对于一个指纹缓存的请求由资料库通信模块308从中央资料库114中来发送。这可以被完成来限制被发送到备份服务器108的每个指纹缓存306的大小，该备份服务器可以具有有限的存储器和处理资源。在此实施方案中，一个第一指纹缓存306可以覆盖该备份流的初始部分（例如，最初的10,000个数据段）。一旦段处理模块314已处理了该备份流的初始部分，则一个请求可被发送到中央资料库114中用于在该系列中的下一个指纹缓存，并且该备份

流的下一部分可被处理，以此类推。

[0037] 在一个实施方案中，如图 3 中所展示的某些模块可以被定位于一个客户端 102 而不是备份服务器 108 上。例如，指纹模块 302 可被定位在每个客户端 102 上。其中，指纹集 312 被创建在客户端 102 上并且被发送到备份服务器 108 中用于进一步地处理。在此实施方案中，由于所传输的是指纹而不是实际的数据段，因此较少数据可以在局域网 106 上被传输。然而，局域网 106 典型地是一个高带宽网络，在其中进行通信是不昂贵的，因此在该网络上传输数据段是可以接受的。在一个实施方案中，备份服务器 108 的全部部件都在一个客户端 102 上。在此实施方案中，客户端 102 直接与中央资料库 114 进行通信，并且一个单独的备份服务器 108 并不用于该备份操作。

[0038] 图 4 展示了在一个实施方案中的用于存储来自客户端存储装置 104 的被备份的数据段的中央资料库 114 的一个逻辑视图。段存储器 402 包含被备份的数据段并且可以包括数百万个数据段。指纹存储器 410 包含与在段存储器 402 中每个段相对应的指纹。

[0039] 在一个实施方案中，数据库 404 包括对在段存储器 402 中的数据段以及对在指纹存储器 410 中对应指纹的多个索引，这些索引基于多种备份规范参数使能多个查询。例如，一个查询可以包括带有一个特定备份类型（例如，全备份）和策略（例如，非可执行文件）的一个特定备份源（例如，客户端主机名称）。数据库 404 返回与备份规范 310 相关的指纹和 / 或数据段。数据库 404 还能够有效地确定一个被查询的特定指纹是否被包含在指纹存储器 410 中。数据库 404 还可以保有对每个数据段以及相应指纹的参引计数来表明当前具有每个数据段的客户端存储装置 104 的数目。如果一个参引计数达到零，那么该数据段以及指纹可分别从段存储器 402 与指纹存储器 410 中被去除。

[0040] 备份服务器通信模块 408 处理与备份服务器 108 的通信。一种类型的通信是基于一种备份规范 310 的对于一个指纹从备份服务器 108 中接收的请求。备份服务器通信模块 408 将此请求传输到指纹缓存发生器 406 中。

[0041] 指纹缓存发生器 406 用备份规范 310 的这些参数来查询数据库 404 以生成一组指纹。指纹缓存发生器 406 将该组指纹作为指纹缓存 306 发送到备份服务器通信模块 408 中，该备份服务器通信模块将其发送给备份服务器 108。在一个实施方案中，指纹缓存发生器 406 创建一系列如上述被发送给该备份服务器的大小有限的指纹高速度缓存 306。

[0042] 如果用一个新的备份规范的一次备份已被执行，那么用备份规范 310 的参数的数据库 404 的查询可以导致无匹配。其中，指纹缓存发生器 406 可生成一个空的指纹缓存 306。在一个实施方案中，取代返回一个空的指纹缓存 306，指纹缓存发生器 406 初始化一个包含多个指纹的新的指纹缓存，这些指纹可以会包含在该备份中。数据库 404 的一个查询可使用少于备份规范 310 的全部参数来执行或可使用不同但相似的参数来执行。

[0043] 备份服务器通信模块 408 还处理来自备份服务器 108 的多个请求，用于确定特定的指纹是否出现在中央资料库 114 中。备份服务器通信模块 408 可查询带有多个特定指纹的数据库 404 并且将这些结果发送给备份服务器 108。并且，备份服务器通信模块 408 可将这些指纹传输到更新模块 412 中用于对数据库 404 进行更新。如果备份服务器 108 确定数据段需要被加入到中央资料库 114 中，那么备份服务器通信模块 408 还可从备份服务器 108 中接收实际的数据段。备份服务器通信模块 408 将这些数据段传输到更新模块 412 中。并且，备份服务器通信模块 408 可从备份服务器 108 中接收出现在指纹缓存 306 中但不在

当前指纹集 312 中的一个指纹列表。备份服务器通信模块 408 将这些指纹传输到更新模块 412 中。

[0044] 更新模块 412 基于从备份服务器 108 接收的通信对数据库 404、段存储器 402 和指纹存储器 410 进行更新。如上述所提及的，这些通信从备份服务器通信模块 408 中被传输到更新模块 412 中。来自备份服务器 108 的关于特定指纹的存在的一个请求表明由于这些特定的指纹不在所提供的指纹缓存 306 中所以它们当前与当前备份规范 310 无关。这些特定的指纹中的一些可以已经在指纹存储器 410 中（相对应的数据段在段存储器 402 中）。对于这些指纹，更新模块 412 对数据库 404 进行更新，这样使得这些指纹与当前备份规范 310 的参数相关，同时保留与已经相关的备份规范的参数的关联。

[0045] 当备份服务器 108 将一个数据段发送到中央资料库 114 用于存储时，更新模块 412 将该段加入到段存储器 402 中并且将一个相对应的指纹加入到指纹存储器 410 中。这个指纹可以已经随同该数据段被发送或者可以由中央资料库 114 中的一个指纹模块 302（未示出）来生成。更新模块 412 还对数据库 404 进行更新，这样使得这些当前备份规范参数参引了新加入的数据段和指纹。

[0046] 来自备份服务器 108 的包含一个指纹列表（该指纹列表包含在指纹缓存 306 中但在该备份流的数据段中找不到）的一个报文表明这些指纹与当前备份规范 310 不再有关联。更新模块 412 可以去除这些当前备份规范参数与在数据库 404 中所识别的指纹之间的关联。由该更新模块执行的这一维护操作可以有益地防止了多个旧指纹被包括在指纹缓存中。

[0047] 备份服务器 108 和中央资料库 114 还可以包含用于在客户端 102 上执行备份存储操作的多个模块。这些模块在此就不做描述了但是它们可使用本领域中皆知的技术来实现。

[0048] 图 5 是展示了在一个实施方案中的用于使用去重复将数据从客户端存储装置 104 备份到中央资料库 114 的方法的一个流程图。在 502 中，一种备份规范 310 由备份服务器 108 来接收。这个备份规范 310 可以已由系统管理员创建并且可以识别备份源、备份类型以及备份策略。资料库通信模块 308 基于备份规范 310 请求来自中央资料库 114 的一个指纹缓存。中央资料库 114 响应，并且在 504 中一个指纹缓存 306 被接收。备份模块 304 基于备份规范 310 启动来自一个客户端存储装置 104 的一个备份流，并且在 506 中，指纹模块 302 生成在该备份流中的数据段的多个指纹。

[0049] 这些数据段的指纹与指纹缓存 306 中的指纹进行比较，并且在 508 中确定在该缓存中未找到的这些数据段指纹。在 510 中，资料库通信模块 308 通过将一个请求发送给中央资料库 114 来关于这些数据段指纹对该中央资料库进行查询。中央资料库 114 响应该请求来表明哪些数据段指纹对应于已存储在该中央资料库中的数据段。在 512 中，该备份服务器将其他的数据段（即，尚未存储在该中央资料库中的这些数据段）发送到中央资料库 114 中用于存储。

[0050] 图 6 是展示了在一个实施方案中的用于使用去重复在一个中央资料库 114 备份数据的方法的一个流程图。在 602 中，中央资料库 114 接收来自备份服务器 108 的一种备份规范。在 604 中，指纹缓存发生器 406 通过用来自备份规范 310 的参数查询数据库 404 来生成一个指纹缓存 306。在 606 中，备份服务器通信模块 408 将该指纹缓存发送给备份服

务器 108。备份服务器 108 处理指纹缓存 306，并且在 608 中，发送关于多个单独指纹的由中央资料库 114 接收的多个查询。在 610 中，备份服务器通信模块 408 在这些单独指纹的数据库 404 中进行一次查找并且将一个响应发送给备份服务器 108。备份服务器 108 处理该响应，并且在 612 中将被接收并存储的多个数据段发送给中央资料库 114。更新模块 412 基于关于单独指纹的这些查询以及所接收的数据段对数据库 404 进行更新。

[0051] 以上说明被包括用以展示优选实施方案的操作并且它无意限制本发明的范围。本发明的范围仅由以下权利要求限定。从以上讨论中，很多变体对于相关领域的技术人员将是清楚的，这些变体仍应被本发明的精神和范围所涵盖。

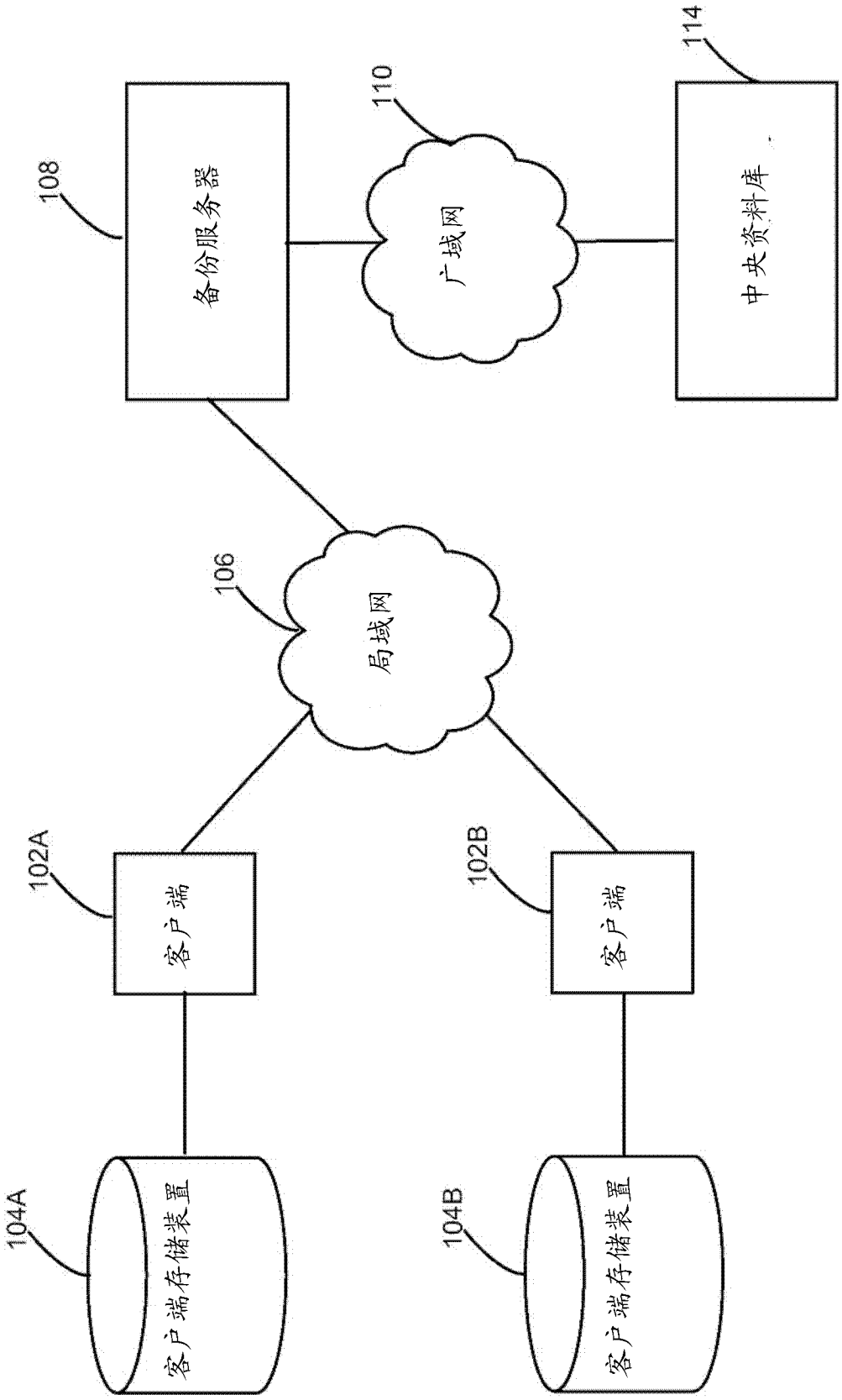

108
备份服务器
110
广域网
114
中央资料库
106
局域网
102A
客户端
102B
客户端
104A
客户端存储装置
104B
客户端存储装置

图 1

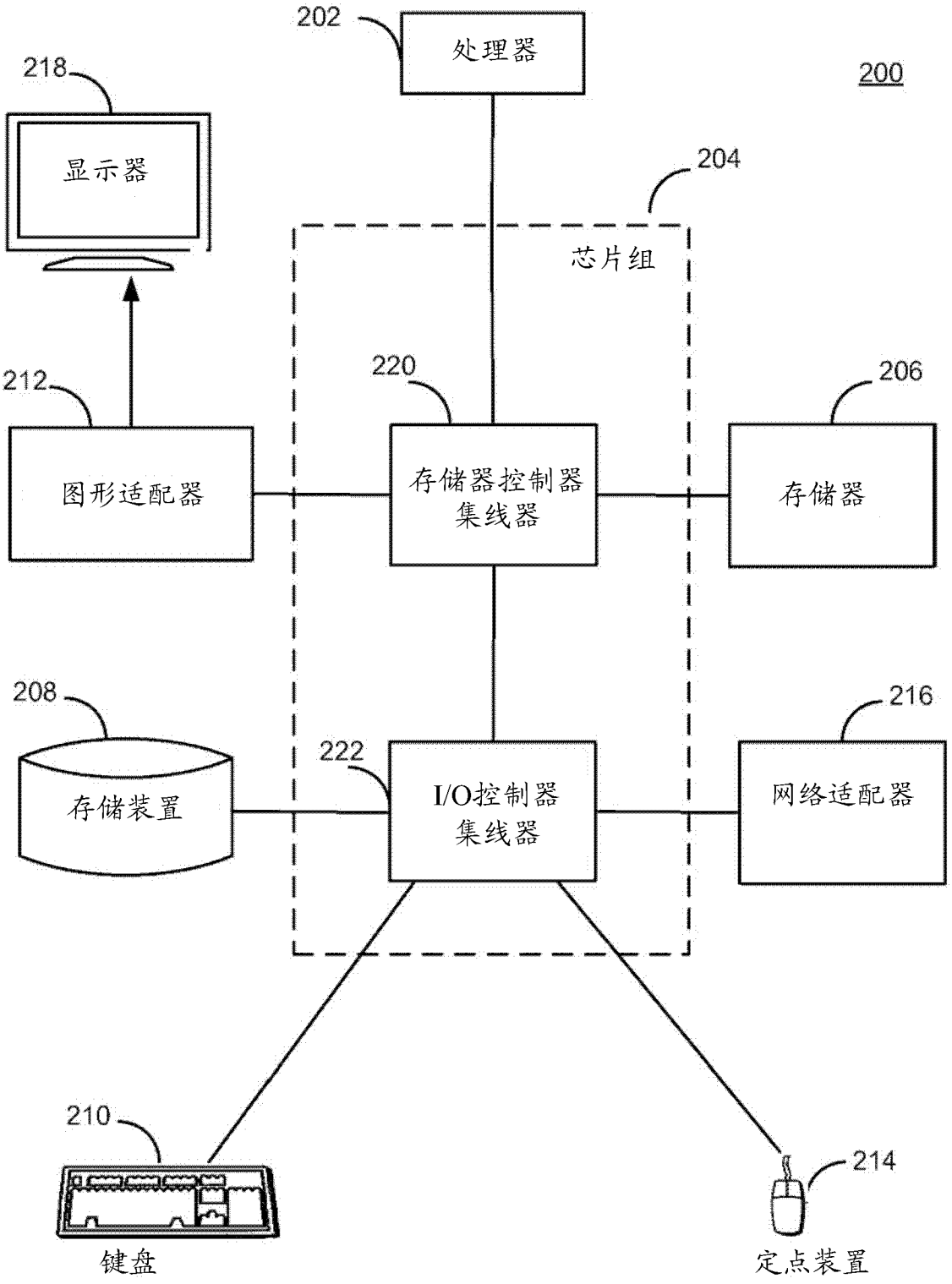

202
处理器
200
218
显示器
204
芯片组
212
220
206
图形适配器
存储器控制器集线器
存储器
208
216
222
存储装置
I/O控制器集线器
网络适配器
210
214
键盘
定点装置

图 2

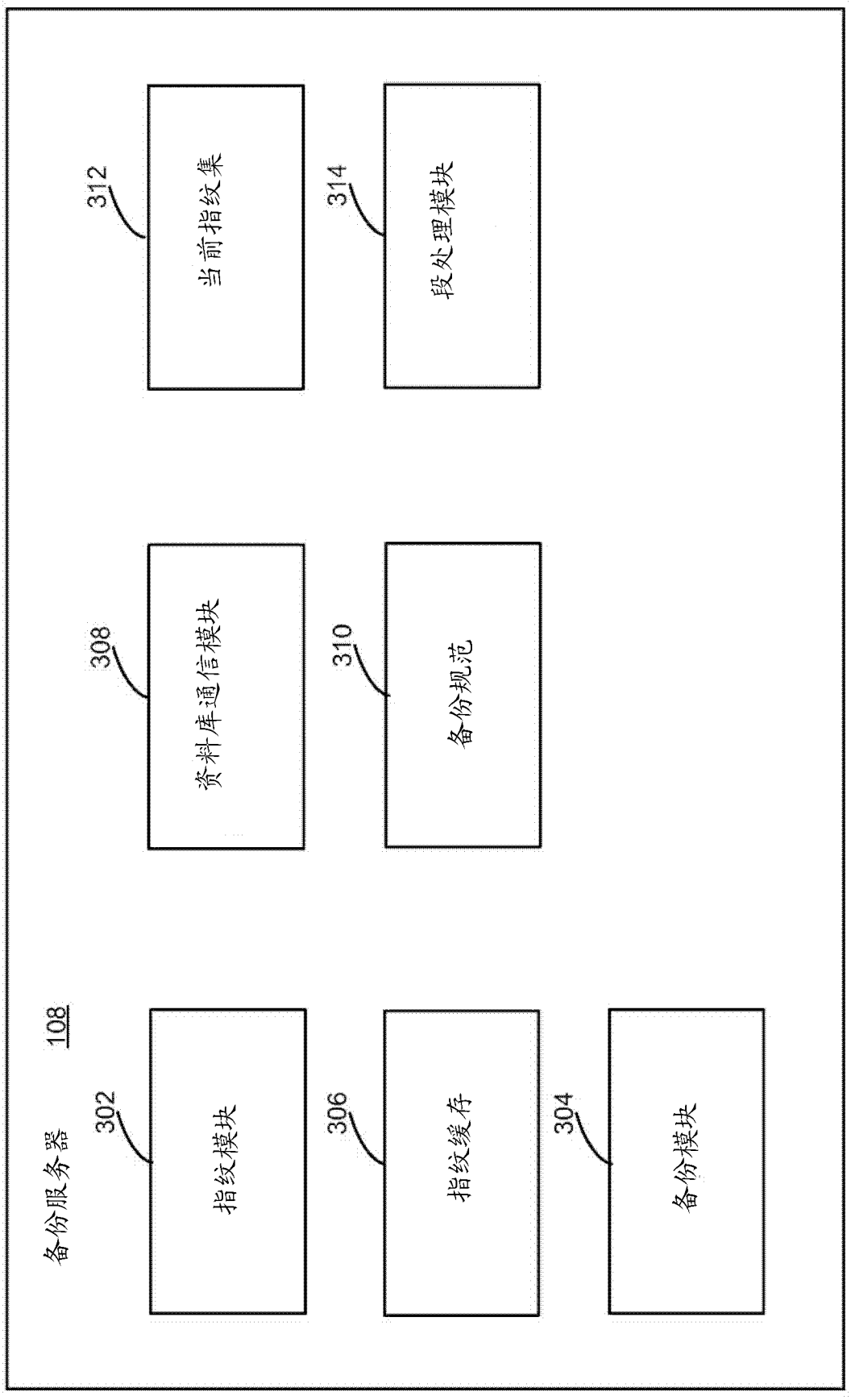

备份服务器 108
302
指纹模块
306
指纹缓存
304
备份模块
308
资料库通信模块
310
备份规范
312
当前指纹集
314
段处理模块

图 3

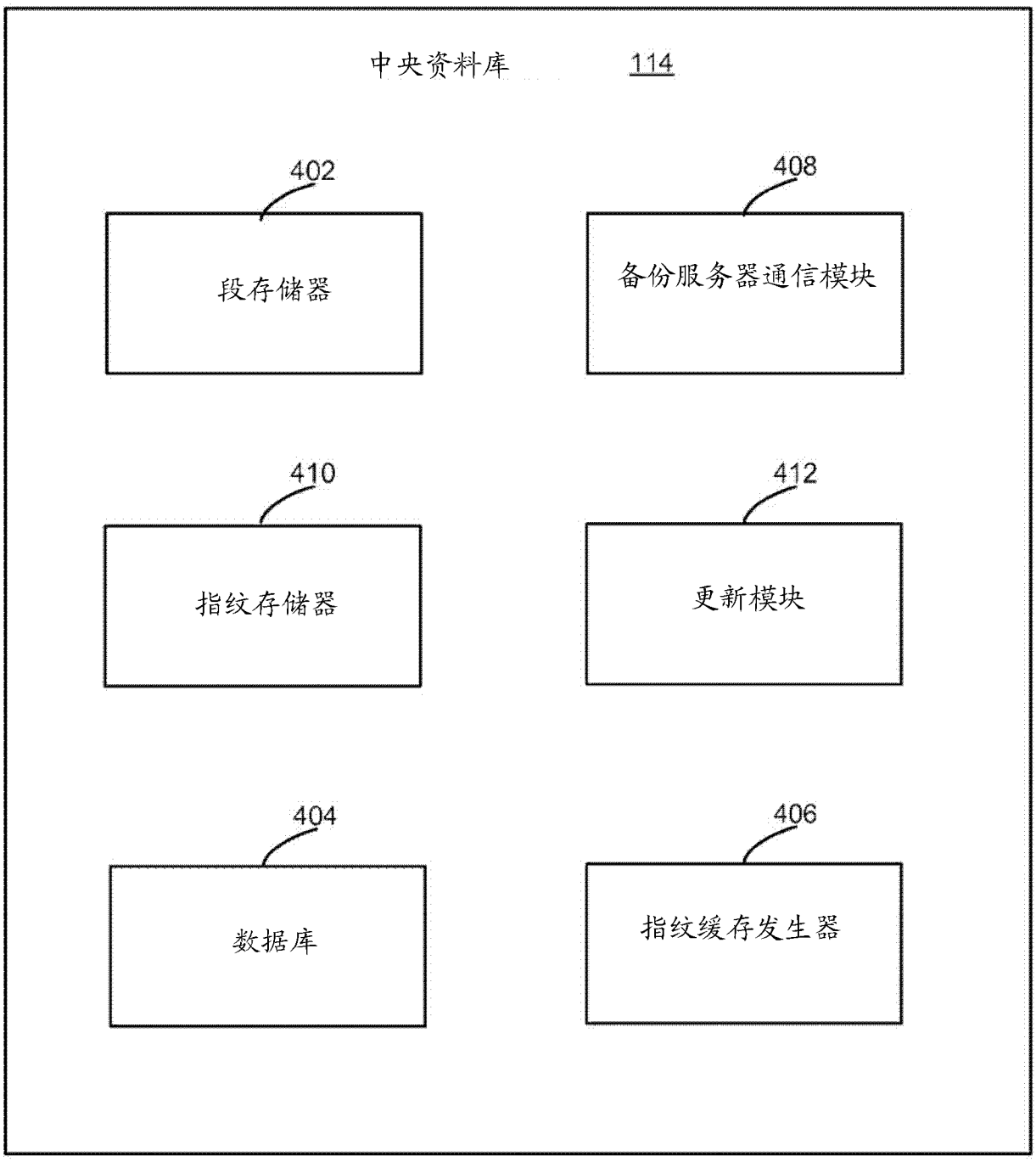

中央资料库 114
402
段存储器
408
备份服务器通信模块
410
指纹存储器
412
更新模块
404
数据库
406
指纹缓存发生器

图 4

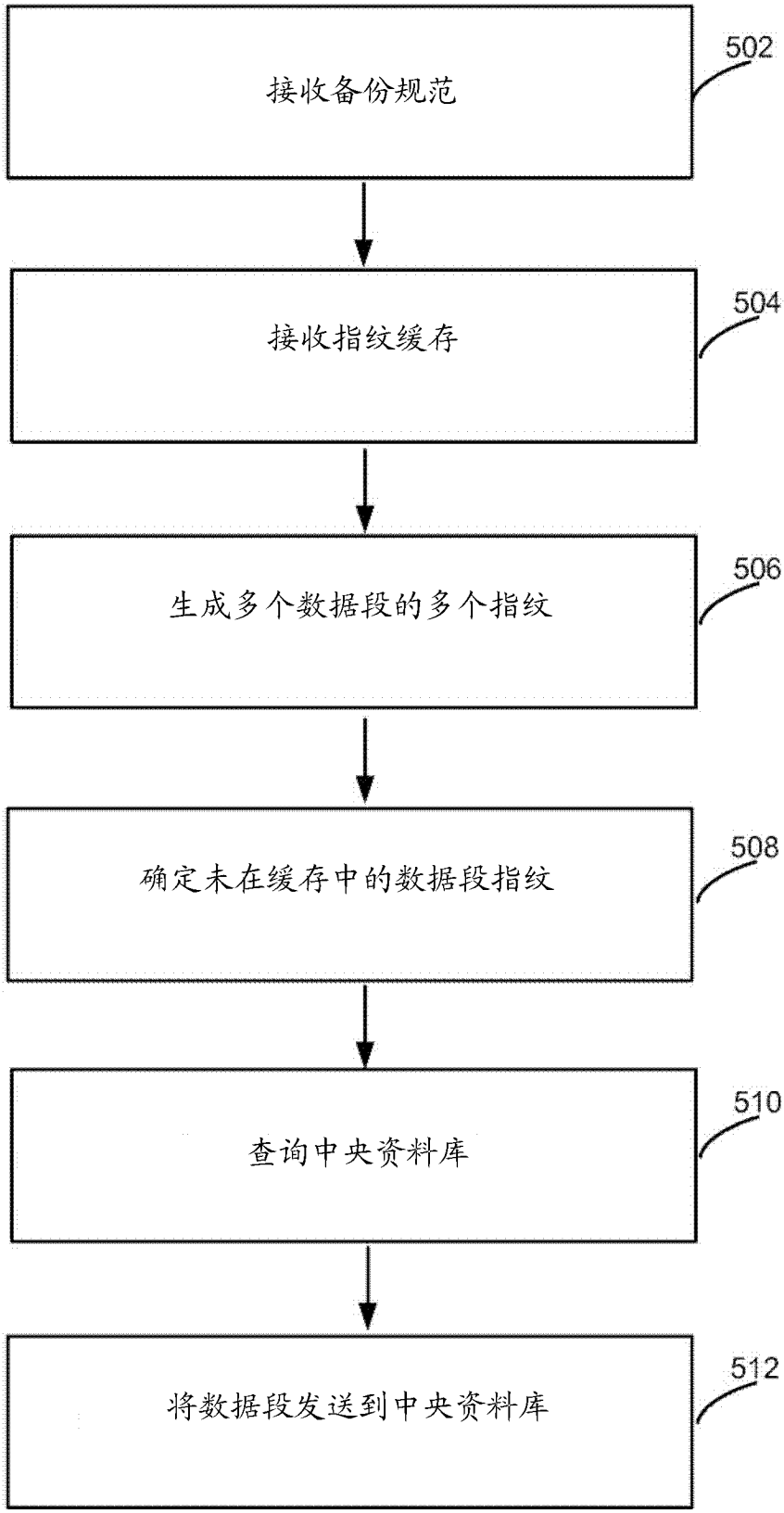

接收备份规范
502
接收指纹缓存
504
生成多个数据段的多个指纹
506
确定未在缓存中的数据段指纹
508
查询中央资料库
510
将数据段发送到中央资料库
512

图 5

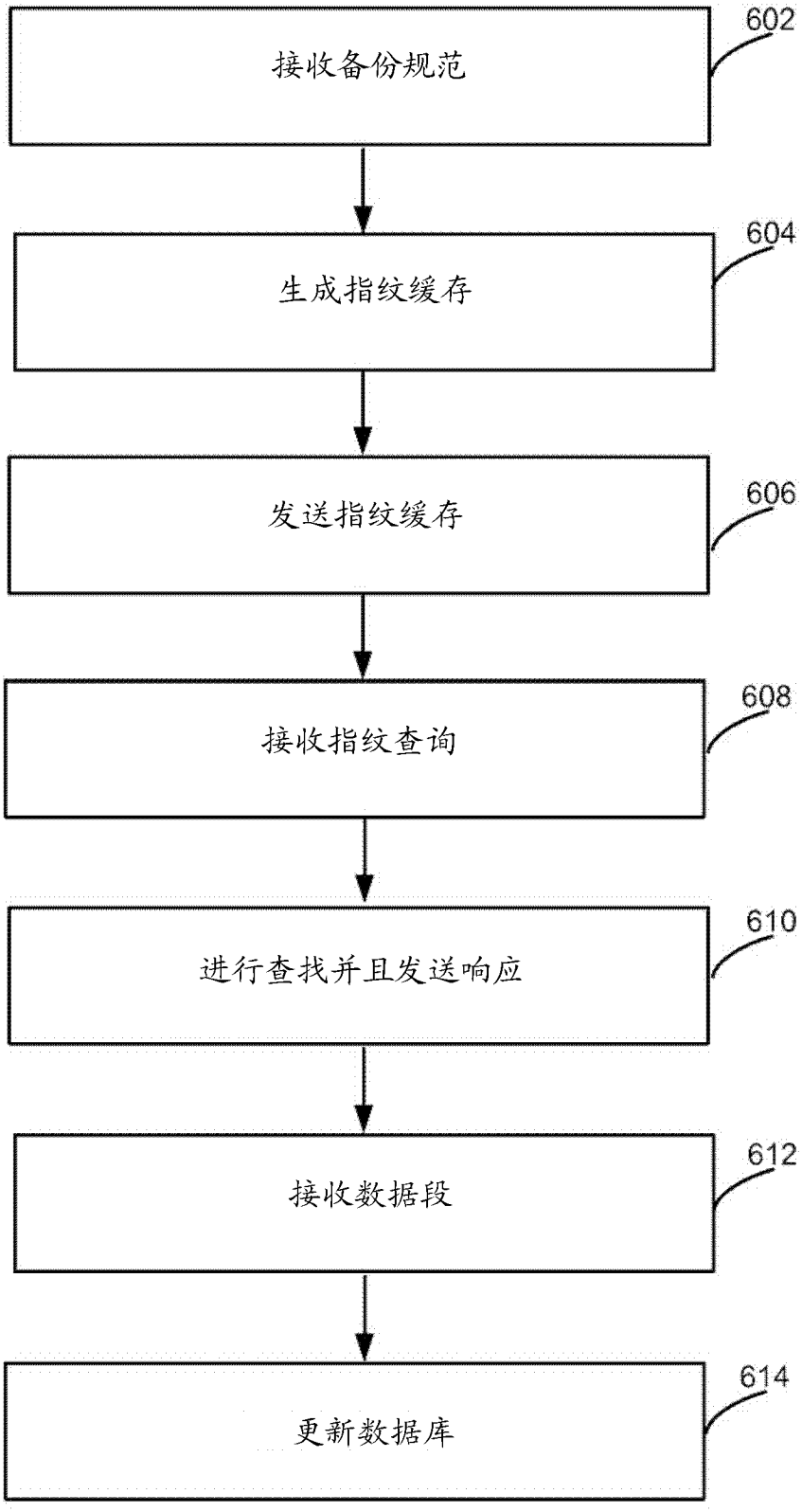

接收备份规范
602
生成指纹缓存
604
发送指纹缓存
606
接收指纹查询
608
进行查找并且发送响应
610
接收数据段
612
更新数据库
614

图 6