

一种基于云管理平台的向量处理方法以及云管理平台

本申请要求于 2024 年 1 月 18 日提交国家知识产权局、申请号为 202410077190.0、发明名称为“一种任务调度的方法、装置及其他设备”的中国专利申请的优先权，以及要求于 2024 年 4 月 30 日提交国家知识产权局、申请号为 202410543817.7、发明名称为“一种基于云管理平台的向量处理方法以及云管理平台”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

本申请实施例涉及云技术领域，尤其涉及一种基于云管理平台的向量处理方法以及云管理平台。

背景技术

随着云技术的快速发展，越来越多的租户选择将其非结构化的数据以向量表示的形式存储在云厂商提供的云服务系统中，以供租户随时访问和查询某个向量，从而满足租户的向量处理需求。

在相关技术中，以存算分离架构搭建的云服务系统可包含多个计算节点以及多个存储节点，这多个计算节点基于功能的不同被划分为不同类型的计算节点。例如，具备向量生成功能的计算节点可将租户的非结构化数据转换为向量，并存储于存储节点中，又如，具备向量查询功能的计算节点可基于租户的请求，从存储节点中为租户查找相应的向量，并返回给租户等等。

上述系统中，由于多个计算节点被划分为不同类型的计算节点，导致特定类型的计算节点仅能完成特定类型的向量处理任务，导致多个计算节点的资源无法共享，造成了一定的资源浪费，导致系统的整体成本过高。

发明内容

本申请实施例提供了一种基于云管理平台的向量处理方法以及云管理平台，其可令多个计算节点的资源在全局上实现共享，更加有效地利用了计算节点的资源，从而在一定程度上降低成本。

本申请实施例的第一方面提供了一种基于云管理平台的向量处理方法，用于实现该方法的云管理平台可管理提供为租户提供云服务的基础设施，基础设施包含服务于租户的多个计算节点以及多个存储节点，多个存储节点用于存储租户的多个数据，该方法包括：

当云管理平台接收到租户所触发的向量处理请求后，云管理平台可基于向量处理请求获取服务于租户的多个计算节点的状态信息。

得到多个计算节点的状态信息后，云管理平台可利用负载均衡算法对多个计算节点的状态信息进行分析，以从多个计算节点中选择状态最优（即负载最低）的计算节点作为第一计算节点，值得注意的是，对于多个计算节点中的任意一个计算节点而言，该计算节点均设置有处理向量处理请求的能力，也就是说，即使向量处理请求为向量生成请求、向量查询请求、向量增加请求以及向量删除请求中的任意一种，该计算节点都可执行向量处理请求。

确定第一计算节点后，云管理平台可向第一计算节点发送向量处理请求，以使得第一计算节点可执行向量处理请求，从而完成向量处理。

例如，当向量处理请求为来自租户的针对多个数据的向量生成请求时，第一计算节点可从多个存储节点处获取多个数据，以生成与多个数据对应的多个向量，并将多个向量存储于多个存储节点中，从而满足租户的向量生成需求。

又如，当向量处理请求为来自租户的针对第一向量的向量查询请求时，第一计算节点可从自身存储的第二向量或多个存储节点存储的多个向量中，获取与第一向量匹配的第三向量。其中，多个存储节点所存储的多个向量包含第一计算节点所存储的第二向量，从而满足租户的向量查询需求。

又如，当向量处理请求为针对第一数据的向量增加请求时，第一计算节点可从多个存储节点处获取第一数据，以生成与第一数据对应的第四向量，并将第四向量存储于多个存储节点中。其中，第一数据为租户写入多个存储节点中的新数据。

又如，当向量处理请求为针对第二数据的向量删除请求时，第一计算节点可从多个存储节点存储的多个向量中，将与第二数据对应的第五向量删除。其中，第二数据为租户从多个存储节点存储的多个数据中删除的数据。

从上述方法可以看出：对于租户所触发的向量处理请求，云管理平台可获取服务于租户的各个计算节点的状态信息，并基于此从这多个计算节点中选出状态最优的计算节点来执行向量处理请求。无论向量处理请求为向量生成请求、向量查询请求、向量增加请求以及向量删除请求中的任意一种，从云管理平台的角度上看，多个计算节点的功能并无差别，每个计算节点均可执行向量处理请求，故云管理平台可从多个计算节点中选择状态最优的某个计算节点，以调用该计算节点执行该向量处理请求。由此可见，由于云管理平台具备针对多个计算节点的任务调度功能，可令每个计算节点可执行各种类型的向量处理请求，从而使得多个计算节点的资源在全局上实现共享，更加有效地利用了计算节点的资源，从而在一定程度上降低云服务系统的成本。

在一种可能实现的方式中，向量增加请求或向量删除请求来自多个存储节点。前述实现方式中，当租户向多个存储节点写入第一数据后，多个存储节点可自行生成针对第一数据的向量增加请求，并发送至云管理平台，以使得云管理平台调用第一计算节点来执行向量增加请求，从而自行完成向量增加需求。同理，当租户从多个存储节点删除第一数据后，多个存储节点可自行生成针对第二数据的向量删除请求，并发送至云管理平台，以使得云管理平台调用第一计算节点来执行向量删除请求，从而自行完成向量删除需求。由此可见，向量增加请求和向量删除请求可由云管理平台、计算节点以及存储节点自行生成和处理，不需要租户来向云管理平台发送请求，可降低租户的操作量，提高租户体验。

在一种可能实现的方式中，向量生成请求，用于指示：第一计算节点将多个向量划分为多个向量组，每个向量组所包含向量的数量等于预置的数值；第一计算节点以向量组为单位，将多个向量组分别写入多个存储节点中。前述实现方式中，由于租户的数据有多个，第一计算节点可生成多个数据中第一个数据对应的向量，并存储该向量，接着再生成第二个数据对应的向量，并存储该向量，以此类推，当第一计算节点所存储的向量的数量等于预置的数值时，第一计算节点可将存储的这些向量划分为一个向量组，并将该向量组存储于对象存储池中的某个存储节点。不断重复前述操作，第一计算节点可将多个数据均转换为相应的向量，并将多个向量划分为多个向量组存储于对象存储池中，从而满足租户的向量生成需求。

在一种可能实现的方式中，第二向量以第一向量组的形式存储在第一计算节点中，向量生成请求，还用于指示：在得到多个向量组后，第一计算节点可从多个向量组中选择第一向量组，并存储第一向量组；第一计算节点向云管理平台通知第一计算节点存储有第一向量组。前述实现方式中，在得到多个向量组的过程中，第一计算节点可自行选择所需存储的向量组，即第一向量组，并在存储第一向量组，第一计算节点可向云管理平台汇报其存储有第一向量组，以使得云管理平台记录第一计算节点的向量存储情况，以在后续获取第一计算节点的状态信息，进而对第一计算节点进行准确的状态评估。

在一种可能实现的方式中，第二向量以第一向量组的形式存储在第一计算节点中，向量查询请求，用于指示：第一计算节点检测第一向量组中，是否存在与第一向量匹配的第三向量；若第一向量组中存在第三向量，第一计算节点从第一向量组中获取第三向量；或者，若第一向量组中不存在第三向量，第一计算节点从多个存储节点处获取多个向量组中除第一向量组之外的第二向量组，并从第二向量组中获取第三向量，其中，多个向量组基于多个向量划分得到。前述实现方式中，当租户需要查询第一向量时，第一计算节点可从先检测自身存储的第一向量组是否存在与第一向量匹配的第三向量，若存在，第一计算节点可从自身存储的第一向量组获取第三向量，并返回给租户，若不存在，第一计算节点可从多个存储节点处获取除第一向量组之外的第二向量组，并从第二向量组中获取第三向量，并返回给租户，从而满足租户的向量查询需求。

在一种可能实现的方式中，第二向量以第一向量组的形式存储在第一计算节点中，向量查询请求，还用于指示：第一计算节点从第二向量组中，确定第三向量所在的第三向量组，并基于第三向量组对第一向量组进行更新，得到更新后的第一向量组，更新后的第一向量组包含第三向量组；第一计算节点向云管理平台通知第一计算节点存储有更新后的第一向量组。前述实现方式中，由于第一计算节点自身并未存储第三向量所在的第三向量组，当第一计算节点确定需要存储第三向量组时，第一计算节点可令第

三向量组代替第一向量组中的某个向量组，从而得到更新后的第一向量组，并向云管理平台通知其存储有更新后的第一向量组，以使得云管理平台更新第一计算节点的最新向量存储情况，以在后续获取第一计算节点的状态信息，进而对第一计算节点进行更加准确的状态评估。

5 在一种可能实现的方式中，向量增加请求，用于指示：第一计算节点生成与第一数据对应的第四向量，并存储第四向量；若第一计算节点存储的未划分成向量组的向量的数量等于预置的数值，第一计算节点将未划分成向量组的向量划分为第四向量组，并将第四向量组写入多个存储节点，未划分成向量组的向量包含第四向量。前述实现方式中，当需要增加与第一数据对应的第四向量时，第一计算节点可从先从多个存储节点处获取第一数据，并生成与第一数据对应的第四向量，再存储第四向量。接着，第一计算节点可判断自身存储的未划分成向量组的向量（包含第四向量）的数量是否等于预置的数值，若是，
10 第一计算节点则将这些向量划分为第四向量组，并将第四向量组写入多个存储节点，从而自行完成向量增加。

在一种可能实现的方式中，第二向量以第一向量组的形式存储在第一计算节点中，向量增加请求，还用于指示：第一计算节点基于第四向量组对第一向量组进行更新，得到更新后的第一向量组，更新后的第一向量组包含第四向量组；第一计算节点向云管理平台通知第一计算节点存储有更新后的第一向量组。前述实现方式中，由于第一计算节点自身并未存储新生成的第四向量组，当第一计算节点确定需要存储第四向量组时，第一计算节点可令第四向量组代替第一向量组中的某个向量组，从而得到更新后的第一向量组，并向云管理平台通知其存储有更新后的第一向量组，以使得云管理平台更新第一计算节点的最新向量存储情况，以在后续获取第一计算节点的状态信息，进而对第一计算节点进行更加准确的状态评估。
15

20 在一种可能实现的方式中，多个向量以多个向量组的形式存储在多个存储节点中，向量删除请求，用于指示：第一计算节点从多个存储节点存储的多个向量组中，获取与第二数据对应的第五向量所在的第五向量组，并从多个存储节点存储的多个向量组中，删除第五向量组；第一计算节点基于第五向量组中除第五向量之外的第六向量生成第六向量组，并将第六向量组存储于多个存储节点中。前述实现方式中，当需要删除与第二数据对应的第五向量时，第一计算节点可从先从多个存储节点存储的多个向量组中，确定与第五向量所在的第五向量组，并删除第五向量组。此外，第一计算节点还可基于第五向量组中除第五向量之外的第六向量生成第六向量组，并将第六向量组写入多个存储节点，从而自行完成向量删除和回收。
25

在一种可能实现的方式中，向量删除请求，还用于指示：第一计算节点基于第六向量组对第一向量组进行更新，得到更新后的第一向量组，更新后的第一向量组包含第六向量组；第一计算节点向云管理平台通知第一计算节点存储有更新后的第一向量组。前述实现方式中，由于第一计算节点自身并未存储新生成的第六向量组，当第一计算节点确定需要存储第六向量组时，第一计算节点可令第六向量组代替第一向量组中的某个向量组，从而得到更新后的第一向量组，并向云管理平台通知其存储有更新后的第一向量组，以使得云管理平台更新第一计算节点的最新向量存储情况，以在后续获取第一计算节点的状态信息，进而对第一计算节点进行更加准确的状态评估。
30

35 在一种可能实现的方式中，云管理平台基于向量处理请求以及多个计算节点的状态信息，根据负载均衡算法从多个计算节点中确定负载最低的第一计算节点包括：云管理平台基于向量处理请求，根据负载均衡算法对多个计算节点的状态信息进行计算，得到多个计算节点的评估值，多个计算节点的评估值用于指示多个计算节点的负载；云管理平台从多个计算节点中，将评估值最小的计算节点确定为第一计算节点；其中，对于多个计算节点中的任意一个计算节点，计算节点的状态信息包含以下至少一项：计算节点加载自身存储的向量所需的时间，计算节点加载自身存储的向量的索引所需的时间，计算节点加载多个存储节点存储的向量所需的时间，计算节点加载多个存储节点存储的向量的索引所需的时间，以及计算节点执行向量处理请求所需的时间。前述实现方式中，对于多个计算节点中的任意一个计算节点，云管理平台可通过负载均衡算法对该计算节点加载自身存储的向量所需的时间，该计算节点加载自身存储的向量的索引所需的时间，该计算节点加载多个存储节点存储的向量所需的时间，该计算节点加载多个存储节点存储的向量的索引所需的时间，以及该计算节点执行向量处理请求所需的时间等该计算节点的状态信息进行计算，从而得到该计算节点的评估值，对于其余计算节点，也是如此，故云管理平台可
40
45

得到多个计算节点的评估值。由于评估值越小，说明计算节点的状态越好（即计算节点的负载越低），云管理平台可选择评估值最小的计算节点作为第一计算节点，以高效地执行向量处理请求。

在一种可能实现的方式中，该方法还包括以下至少一项：云管理平台对计算节点存储的向量以及多个存储节点存储的向量进行分析，得到计算节点加载自身存储的向量所需的时间以及计算节点加载多个存储节点存储的向量所需的时间；云管理平台对计算节点存储的向量的索引以及多个存储节点存储的向量的索引进行分析，得到计算节点加载自身存储的索引所需的时间以及计算节点加载多个存储节点存储的索引所需的时间；云管理平台对计算节点的资源使用情况进行分析，得到计算节点执行向量处理请求所需的时间。前述实现方式中，由于云管理平台具备向量状态表、向量索引状态表以及计算节点状态表，向量状态表记录着各个计算节点所存储的向量，向量索引状态表记录着各个计算节点所存储的向量的索引，且计算节点状态表记录着各个计算节点的资源使用情况，故云管理平台可利用这三个表进行分析，从而准确得到各个计算节点的状态信息，进而准确评估各个计算节点。

本申请实施例的第二方面提供了一种云管理平台，云管理平台用于管理提供云服务的基础设施，基础设施包含多个计算节点以及多个存储节点，多个存储节点用于存储租户的多个数据，云管理平台包括：接收模块，用于接收来自租户的针对多个数据的向量生成请求；处理模块，用于基于向量生成请求以及多个计算节点的状态信息，根据负载均衡算法从多个计算节点中确定负载最低的第一计算节点，多个计算节点均设置有处理向量生成请求的能力；处理模块，还用于向第一计算节点发送向量生成请求，向量生成请求用于指示第一计算节点生成与多个数据对应的多个向量，并将多个向量存储于多个存储节点中。

在一种可能实现的方式中，接收模块，还用于接收来自租户的针对第一向量的向量查询请求；处理模块，还用于基于向量查询请求以及多个计算节点的状态信息，根据负载均衡算法从多个计算节点中确定负载最低的第一计算节点，多个计算节点均设置有处理向量查询请求的能力；处理模块，还用于向第一计算节点发送向量查询请求，向量查询请求用于指示第一计算节点从自身存储的第二向量或多个存储节点存储的多个向量中，获取与第一向量匹配的第三向量，多个向量包含第二向量。

在一种可能实现的方式中，接收模块，还用于接收针对第一数据的向量增加请求，第一数据为租户写入多个存储节点中的新数据；处理模块，还用于基于向量增加请求以及多个计算节点的状态信息，根据负载均衡算法从多个计算节点中确定负载最低的第一计算节点，多个计算节点均设置有处理向量增加请求的能力；处理模块，还用于向第一计算节点发送向量增加请求，向量增加请求用于指示第一计算节点生成与第一数据对应的第四向量，并将第四向量存储于多个存储节点中。

在一种可能实现的方式中，接收模块，还用于接收针对第二数据的向量删除请求，第二数据为租户从多个存储节点存储的多个数据中删除的数据；处理模块，还用于基于向量删除请求以及多个计算节点的状态信息，根据负载均衡算法从多个计算节点中确定负载最低的第一计算节点，多个计算节点均设置有处理向量删除请求的能力；处理模块，还用于向第一计算节点发送向量删除请求，向量删除请求用于指示第一计算节点从多个存储节点存储的多个向量中，将与第二数据对应的第五向量删除。

在一种可能实现的方式中，向量增加请求或向量删除请求来自多个存储节点。

在一种可能实现的方式中，向量生成请求，用于指示：第一计算节点将多个向量划分为多个向量组，每个向量组所包含向量的数量等于预置的数值；第一计算节点以向量组为单位，将多个向量组分别写入多个存储节点中。

在一种可能实现的方式中，第二向量以第一向量组的形式存储在所述第一计算节点中，向量生成请求，还用于指示：在得到多个向量组后，第一计算节点可从多个向量组中选择第一向量组，并存储第一向量组；第一计算节点向云管理平台通知第一计算节点存储有第一向量组。

在一种可能实现的方式中，第二向量以第一向量组的形式存储在所述第一计算节点中，向量查询请求，用于指示：第一计算节点检测第一向量组中，是否存在与第一向量匹配的第三向量；若第一向量组中存在第三向量，第一计算节点从第一向量组中获取第三向量；或者，若第一向量组中不存在第三向量，第一计算节点从多个存储节点处获取多个向量组中除第一向量组之外的第二向量组，并从第二向量组中获取第三向量，其中，多个向量组基于多个向量划分得到。

在一种可能实现的方式中，第二向量以第一向量组的形式存储在所述第一计算节点中，向量查询请求，还用于指示：第一计算节点从第二向量组中，确定第三向量所在的第三向量组，并基于第三向量组对第

一向量组进行更新，得到更新后的第一向量组，更新后的第一向量组包含第三向量组；第一计算节点向云管理平台通知第一计算节点存储有更新后的第一向量组。

在一种可能实现的方式中，向量增加请求，用于指示：第一计算节点生成与第一数据对应的第四向量，并存储第四向量；若第一计算节点存储的未划分成向量组的向量的数量等于预置的数值，第一计算节点将未划分成向量组的向量划分为第四向量组，并将第四向量组写入多个存储节点，未划分成向量组的向量包含第四向量。

在一种可能实现的方式中，第二向量以第一向量组的形式存储在多个存储节点中，向量增加请求，还用于指示：第一计算节点基于第四向量组对第一向量组进行更新，得到更新后的第一向量组，更新后的第一向量组包含第四向量组；第一计算节点向云管理平台通知第一计算节点存储有更新后的第一向量组。

在一种可能实现的方式中，多个向量以多个向量组的形式存储在多个存储节点中，向量删除请求，用于指示：第一计算节点从多个存储节点存储的多个向量组中，获取与第二数据对应的第五向量所在的第五向量组，并从多个存储节点存储的多个向量组中，删除第五向量组；第一计算节点基于第五向量组中除第五向量之外的第六向量生成第六向量组，并将第六向量组存储于多个存储节点中。

在一种可能实现的方式中，向量删除请求，还用于指示：第一计算节点基于第六向量组对第一向量组进行更新，得到更新后的第一向量组，更新后的第一向量组包含第六向量组；第一计算节点向云管理平台通知第一计算节点存储有更新后的第一向量组。

在一种可能实现的方式中，处理模块，用于：基于向量生成请求，根据负载均衡算法对多个计算节点的状态信息进行计算，得到多个计算节点的评估值，多个计算节点的评估值用于指示多个计算节点的负载；从多个计算节点中，将评估值最小的计算节点确定为第一计算节点；其中，对于多个计算节点中的任意一个计算节点，计算节点的状态信息包含以下至少一项：计算节点加载自身存储的向量所需的时间，计算节点加载自身存储的向量的索引所需的时间，计算节点加载多个存储节点存储的向量所需的时间，计算节点加载多个存储节点存储的向量的索引所需的时间，以及计算节点执行向量处理请求所需的时间。

在一种可能实现的方式中，云管理平台还包括：分析模块，分析模块用于执行以下至少一项：对计算节点存储的向量以及多个存储节点存储的向量进行分析，得到计算节点加载自身存储的向量所需的时间以及计算节点加载多个存储节点存储的向量所需的时间；对计算节点存储的向量的索引以及多个存储节点存储的向量的索引进行分析，得到计算节点加载自身存储的索引所需的时间以及计算节点加载多个存储节点存储的索引所需的时间；对计算节点的资源使用情况进行分析，得到计算节点执行向量处理请求所需的时间。

本申请实施例的第三方面提供了一种计算设备集群，计算设备集群包括至少一个计算设备，每个计算设备包括处理器以及存储器：存储器用于存储指令；处理器用于根据指令，以使得计算设备集群执行第一方面或第一方面中任意一种可能实现的方式所述的方法。

本申请实施例的第四方面提供了一种计算机存储介质，计算机存储介质存储有一个或多个指令，指令在由一个或多个计算机执行时使得一个或多个计算机实施第一方面或第一方面中任意一种可能实现的方式所述的方法。

本申请实施例的第五方面提供了一种计算机程序产品，计算机程序产品存储有指令，指令在由计算机执行时，使得计算机实施第一方面或第一方面中任意一种可能实现的方式所述的方法。

本申请实施例中，对于租户所触发的向量处理请求，云管理平台可获取服务于租户的各个计算节点的状态信息，并基于此从这多个计算节点中选出状态最优的计算节点来执行向量处理请求。无论向量处理请求为向量生成请求、向量查询请求、向量增加请求以及向量删除请求中的任意一种，从云管理平台的角度上看，多个计算节点的功能并无差别，每个计算节点均可执行向量处理请求，故云管理平台可从多个计算节点中选择状态最优的某个计算节点，以调用该计算节点执行该向量处理请求。由此可见，由于云管理平台具备针对多个计算节点的任务调度功能，可令每个计算节点可执行各种类型的向量处理请求，从而使得多个计算节点的资源在全局上实现共享，更加有效地利用了计算节点的资源，从而在一定程度上降低云服务系统的成本。

附图说明

- 图 1 为本申请实施例提供的云服务系统的一个结构示意图；
图 2 为本申请实施例提供的基于云管理平台的向量处理方法的一个流程示意图；
5 图 3 为本申请实施例提供的云服务系统的另一结构示意图；
图 4 为本申请实施例提供的基于云管理平台的向量处理方法的另一流程示意图；
图 5 为本申请实施例提供的云服务系统的另一结构示意图；
图 6 为本申请实施例提供的基于云管理平台的向量处理方法的另一流程示意图；
图 7 为本申请实施例提供的云服务系统的另一结构示意图；
10 图 8 为本申请实施例提供的基于云管理平台的向量处理方法的另一流程示意图；
图 9 为本申请实施例提供的云服务系统的另一结构示意图；
图 10 为本申请实施例提供的云管理平台的一个结构示意图；
图 11 为本申请实施例提供的计算设备的一个结构示意图；
图 12 为本申请实施例提供的计算设备集群的一个结构示意图；
15 图 13 为本申请实施例提供的计算机集群中的计算机设备通过网络连接的一个示意图。

具体实施方式

本申请实施例提供了一种基于云管理平台的向量处理方法以及云管理平台，其可令多个计算节点的资源在全局上实现共享，更加有效地利用了计算节点的资源，从而在一定程度上降低成本。

- 20 本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的术语在适当情况下可以互换，这仅仅是描述本申请的实施例中对于相同属性的对象在描述时所采用的区分方式。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，以便包含一系列单元的过程、方法、系统、产品或设备不必限于那些单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它单元。

- 25 随着云技术的快速发展，越来越多的租户选择将其非结构化的数据（例如，图像、文本以及语音等等）以向量表示的形式存储在云厂商提供的云服务系统中，以供租户随时访问和查询某个向量，从而满足租户的向量处理需求。

- 在相关技术中，以存算分离架构搭建的云服务系统可包含多个计算节点以及多个存储节点，这多个计算节点基于功能的不同被划分为不同类型的计算节点。例如，具备向量生成功能的计算节点可将租户的非结构化数据转换为向量，并存储于存储节点中，又如，具备索引生成功能的计算节点可生成向量的索引，并将索引存储于存储节点中。又如，具备向量查询功能的计算节点可基于租户的请求，从存储节点中为租户查找相应的向量，并返回给租户等等。

- 上述系统中，由于多个计算节点被划分为不同类型的计算节点，导致特定类型的计算节点仅能完成特定类型的向量处理任务，无法完成其它类型的向量处理任务，例如，具备向量生成功能的计算节点无法完成向量查询任务。这样会导致多个计算节点的资源无法全局共用，当系统中某种向量处理任务大量突发时，会有一部分计算节点闲置，造成了一定的资源浪费，导致系统的整体成本过高。

- 为了解决上述问题，本申请实施例提供了一种基于云管理平台的向量处理方法，该方法可通过云服务系统实现，图 1 为本申请实施例提供的云服务系统的一个结构示意图，如图 1 所示，包括可提供云服务的基础设施以及管理这些基础设施的云管理平台。下文分别对云管理平台以及基础设施分别进行介绍：

- 40 云管理平台，可对整个云服务系统中的基础设施进行统筹管理（例如，在基础设施中，按照租户的指示，为租户创建服务于租户的多个计算节点以及多个存储节点，这些计算节点以及存储节点可为租户提供向量生成服务、向量查询服务、向量增加服务以及向量删除服务等等），也可对云服务系统外的租户开放，并响应他们的请求。例如，云管理平台可提供登录接口以及处理接口等各种接口，以供租户的客户端（例如，租户所使用的终端设备或终端设备上的浏览器等等）接入。其中，云管理平台可通过登
45 录接口，对租户的客户端进行身份验证，身份验证成功后可允许租户的客户端登录云管理平台。又如，云管理平台还可以通过处理接口，允许租户的客户端向云管理平台发送租户针对租户的多个（非结构化

的)数据(这多个数据已提前存储在服务于租户的多个存储节点中)的向量生成请求,云管理平台可从服务于租户的多个计算节点中选择某个计算节点,以使得该计算节点处理基于该向量生成请求,生成与租户的多个数据对应的多个向量,并将这多个向量存储在多个存储节点中。又如,云管理平台还可以通过处理接口,允许租户的客户端向云管理平台发送租户针对目标向量的向量查询请求,云管理平台可从多个计算节点中选择某个计算节点,以使得该计算节点处理基于该向量查询请求,从自身存储的若干个向量或从多个存储节点存储的多个向量中,找到并向租户返回与目标向量匹配的另一个输出向量。又如,云管理平台还可以通过处理接口,允许租户通过其客户端向多个存储节点写入新数据,以使得多个存储节点向云管理平台发送针对新数据的向量添加请求,云管理平台可从多个计算节点中选择某个计算节点,以使得该计算节点处理基于该向量添加请求,生成与新数据对应的新向量,并将新向量存储于多个存储节点中。又如,云管理平台还可以通过处理接口,允许租户通过其客户端从多个存储节点所存储的多个数据中删除某个数据,以使得多个存储节点向云管理平台发送针对已删除数据的向量删除请求,云管理平台可从多个计算节点中选择某个计算节点,以使得该计算节点处理基于该向量删除请求,从多个存储节点所存储的多个向量中,对与已删除数据对应的待删除向量进行删除等等。

基础设施包含服务于租户的多个计算节点以及多个存储节点。在多个计算节点中,对于任意一个计算节点而言,该计算节点可包含向量缓存(cache)以及向量索引缓存,向量缓存用于存储与租户的多个数据对应的多个向量中的若干个向量,这若干个向量通常以若干个向量组(也可以称为向量片段(vectorsegment))的形式进行存储,每个向量组包含至少一个向量,向量索引缓存用于存储这若干个向量组的索引,这若干个向量组的索引可供该计算节点获取这若干个向量组。对于多个存储节点而言,与租户的多个数据所对应的多个向量可以多个向量组的形式存储在这多个存储节点中,每个向量组包含多个向量中的至少一个向量。而且,这多个向量组的索引也存储在这多个存储节点中,需要说明的是,每个存储节点可存储这多个向量组中的至少一个向量组以及相应的索引(当然,每个存储节点也可存储有这多个向量组以及这多个向量组的索引)。

值得注意的是,服务于租户的多个存储节点,可以划分为两部分存储节点,其中,一部分存储节点构成内存池,另一部分存储节点构成对象存储池。对于内存池中的存储节点,这些存储节点可存储有租户的多个向量组中的一部分向量组以及相应的索引,需要说明的是,内存池所存储的向量组的数量通常大于任意一个计算节点所存储的向量组的数量。对于对象存储池中的存储节点,这些存储节点不仅可存储有租户的多个数据,还可存储有这多个数据对应的多个向量组以及相应的索引。也就是说,对象存储池存储有租户的全部向量组,计算节点和内存池存储有租户的一部分向量组。

还值得注意的是,云管理平台具备任务调度功能,在该功能的基础上,云管理平台可记录以下数据:

(1) 向量状态表:向量状态表可记录每个计算节点所存储的若干个向量组,以及多个存储节点所存储的多个向量组。可以理解的是,对于任意一个计算节点而言,向量状态表记录有该计算节点所存储的若干个向量组,由于该计算节点所存储的这若干个向量组为多个存储节点存储的多个向量组中的一部分向量组,故云管理平台可对该计算节点所存储的这若干个向量组,以及多个存储节点所存储的除这若干个向量组之外的其余向量组进行综合分析(例如,该计算节点存储有多个向量组中的某部分向量组,但未存储有多个向量组中另一部分向量组),从而计算出该计算节点加载(获取)其自身所存储的这若干个向量组所需的时间(即该计算节点加载自身所存储的向量所需的时间),以及该计算节点从多个存储节点处加载除这若干个向量组之外的其余向量组所需的时间(即该计算节点加载多个存储节点所存储的向量所需的时间),这些信息可作为该计算节点的状态信息。

例如,向量状态表如表1所示,设服务于租户的计算节点有计算节点1、计算节点2等计算节点,服务于租户的有多个存储节点。设按照租户的要求,为租户生成了向量组1,向量组2,向量组3,向量组4,向量组5以及向量组6等向量组,并存储在远端的多个存储节点中。且在之后的过程中,计算节点1缓存有向量组1,向量组2以及向量组5等向量组,计算节点2缓存有向量组2,向量组3,向量组5等向量组,故云管理平台可生成如表1所示的向量组状态表,以记录当前各个计算节点以及存储节点的向量存储情况(向量存储情况通常由各个计算节点以及存储节点来汇报给云管理平台,需要说明的是,该例子多个存储节点视为一个整体,向量组1至向量组6等向量组可分别部署在相同或不同的存储节点中)。

表 1

计算节点	本地向量组	远端向量组
1	向量组 1, 向量组 2, 向量组 5, ...	向量组 1, 向量组 2, 向量组 3, 向量组 4, 向量组 5, 向量组 6...
2	向量组 2, 向量组 3, 向量组 5, ...	
...	...	

那么，基于向量状态表，云管理平台可确定计算节点 1 加载自身存储的向量组 1，向量组 2 以及向量组 5 等向量组所需要的时间，以及计算节点 1 加载多个存储节点存储的向量组 3、向量组 4 以及向量组 6 等向量组所需要的时间。同理，云管理平台还确定计算节点 2 等计算节点的这些时间，此处不再赘述。

(2) 向量索引状态表：向量索引状态表可记录每个计算节点所存储的若干个向量组的索引，以及多个存储节点所存储的多个向量组的索引。可以理解的是，对于任意一个计算节点而言，向量索引状态表记录有该计算节点所存储的若干个向量组的索引，由于该计算节点所存储的这若干个向量组的索引为多个存储节点存储的多个向量组的索引中的一部分索引，故云管理平台可基于该计算节点所存储的若干个向量组的索引，以及多个存储节点所存储的除这若干个向量组的索引之外的其余向量组的索引进行综合分析（例如，该计算节点存储有多个向量组中的某部分向量组的索引，但未存储有多个向量组中另一部分向量组的索引），计算出该计算节点加载其自身所存储的这若干个向量组的索引所需的时间（即该计算节点加载自身所存储的索引所需的时间），以及该计算节点从多个存储节点处加载除这若干个向量组的索引之外的其余向量组的索引所需的时间（即该计算节点加载多个存储节点所存储的索引所需的时间），这些信息也可作为该计算节点的状态信息。

依旧如上述例子，向量索引状态表如表 2 所示，云管理平台还可生成如表 2 所示的向量索引状态表，以记录当前各个计算节点以及存储节点的向量索引存储情况（向量索引存储情况通常由各个计算节点以及存储节点来汇报给云管理平台）。

表 2

计算节点	本地向量组的索引	远端向量组的索引
1	向量组 1 的索引, 向量组 2 的索引, 向量组 5 的索引, ...	向量组 1 的索引, 向量组 2 的索引, 向量组 3 的索引, 向量组 4 的索引, 向量组 5 的索引, 向量组 6 的索引...
2	向量组 2 的索引, 向量组 3 的索引, 向量组 5 的索引, ...	
...	...	

那么，基于向量索引状态表，云管理平台可确定计算节点 1 加载自身存储的向量组 1 的索引，向量组 2 的索引以及向量组 5 的索引等索引所需要的时间，以及计算节点 1 加载多个存储节点存储的向量组 3 的索引、向量组 4 的索引以及向量组 6 的索引等索引所需要的时间。同理，云管理平台还确定计算节点 2 等计算节点的这些时间，此处不再赘述。

(3) 计算节点状态表：计算节点状态表可记录每个计算节点的待处理任务（待处理的向量处理请求）的数量，中央处理器（central processing unit, CPU）的已用量以及内存的已用量等等。可以理解的是，对于任意一个计算节点而言，计算节点状态表记录有该计算节点的待处理任务的数量，该计算节点的 CPU 的已用量以及内存的已用量等数据（用于指示该计算节点的资源使用情况），故云管理平台可基于这些数据进行分析，计算出该计算节点执行（处理）某个任务（某个向量处理请求）所需的时间，这些信息也可作为该计算节点的状态信息。

依旧如上述例子，计算节点状态表如表 3 所示，云管理平台还可生成如表 3 所示的计算节点状态表，以记录当前各个计算节点的资源使用情况（资源使用情况通常由各个计算节点来汇报给云管理平台）。

表 3

计算节点	待处理任务的数量	CPU 的已用量	内存的已用量
1	5	80%	70%
2	10	30%	30%
.....			

那么，对于某个待处理任务，云管理平台可基于向量状态表中计算节点 1 的待处理任务的数量、CPU 的已用量以及内存的已用量等数据来确定计算节点 1 执行该待处理任务所需的时间。同理，云管理平台还确定计算节点 2 等计算节点的该时间，此处不再赘述。

基于此，对于租户所触发的某个向量处理请求，云管理平台可基于向量状态表、向量索引状态表以及计算节点状态表，来确定服务于租户的各个计算节点的状态信息，以基于此从服务于租户的多个计算节点中选出状态最优（即负载最低）的计算节点来执行该向量处理请求。也就是说，无论该向量处理请求为向量生成请求、向量查询请求、向量增加请求以及向量删除请求中的任意一种，从云管理平台的角度上看，多个计算节点的功能并无差别，每个计算节点均具备执行该向量处理请求的能力，故云管理平台可从多个计算节点中选择状态最优的某个计算节点，以调用该计算节点执行该向量处理请求。由此可见，由于云管理平台具备针对多个计算节点的任务调度功能，可令每个计算节点可执行各种类型的向量处理请求，从而使得多个计算节点的资源在全局上实现共享，更加有效地利用了计算节点的资源，从而在一定程度上降低云服务系统的成本。为了进一步了解云服务系统的工作流程，下文结合多个实施例对该流程做进一步的介绍，首先对第一个实施例进行介绍，第一个实施例如图 2 所示（图 2 为本申请实施例提供的基于云管理平台的向量处理方法的一个流程示意图），第一个实施例可通过如图 1 所示的云服务系统实现，云服务系统包含为租户提供云服务的基础设施以及管理这些基础设施的云管理平台，这些基础设施包含服务于租户的多个计算节点以及多个存储节点，多个存储节点用于存储租户的多个数据。第一个实施例包括：

201、云管理平台接收来自租户的针对多个数据的向量生成请求。

本实施例中，当租户需要将其存储于多个存储节点中的多个数据转换为多个向量时，租户可通过其客户端向云管理平台发送针对多个数据的向量生成请求。接收到针对多个数据的向量生成请求后，云管理平台可确定需要将租户的多个数据转换为多个向量。

例如，如图 3 所示（图 3 为本申请实施例提供的云服务系统的另一结构示意图），设租户已提前通过云管理平台，在对象存储池中存储了多个（非结构化）数据。当租户需要将这多个数据以向量的形式进行存储时，租户可向云管理平台发送针对这多个数据的向量生成请求。

202、云管理平台基于向量生成请求以及多个计算节点的状态信息，根据负载均衡算法从多个计算节点中确定负载最低的第一计算节点，并向第一计算节点发送向量生成请求，多个计算节点均设置有处理向量生成请求的能力。

得到针对多个数据的向量生成请求后，云管理平台可基于向量生成请求，通过向量状态表、向量索引状态表以及计算节点状态表来获取服务于租户的多个计算节点的状态信息，并基于这多个计算节点的状态信息，根据负载均衡算法从这多个计算节点中选择状态最优（即负载最低的）的计算节点作为第一计算节点，并将向量生成请求发送至第一计算节点。需要说明的是，对于多个计算节点中的任意一个计算节点而言，该计算节点在被云管理平台创建时，已被云管理平台设置有处理向量生成请求的能力。

具体地，云管理平台可通过以下方式从多个计算节点中确定第一计算节点：

在接收到向量生成请求后，对于服务于租户的多个计算节点中的任意一个计算节点而言，云管理平台可基于向量状态表、向量索引状态表以及计算节点状态表，获取该计算节点加载自身所存储的向量所需的时间（由于此时该计算节点并未存储有任何向量组，故该时间为零），该计算节点加载多个存储节点所存储的向量所需的时间（由于此时多个存储节点并未存储有任何向量组，故该时间为零），该计算节点加载自身所存储的索引所需的时间（由于此时该计算节点并未存储有任何向量组的索引，故该时间为零），该计算节点加载多个存储节点所存储的索引所需的时间（由于此时多个存储节点并未存储有任何向量组的索引，故该时间为零），以及该计算节点执行向量生成请求所需的时间等该计算节点的状态

信息，那么，云管理平台可通过负载均衡算法对该计算节点的状态信息进行计算（该负载均衡算法可以有多种，此处不做限制），从而得到该计算节点的评估值，该计算节点的评估值用于指示该计算节点的状态的优劣程度，也就是该计算节点的负载。该评估值越大，说明该计算节点的状态越恶劣，即该计算节点的负载越高，该评估值越小，说明该计算节点的状态越良好，即该计算节点的负载越低。

5 同理，对于多个计算节点除该计算节点之外的其余计算节点，云管理平台也可对其余计算节点执行如同对该计算节点所执行的操作，故最终云管理平台可得到多个计算节点的评估值。那么，在多个计算节点中，云管理平台可将评估值最小的计算节点确定为用于执行向量生成请求的第一计算节点。

依旧如上述例子，云管理平台接收向量生成请求后，可将向量生成请求视为一个向量生成任务，并从向量状态表、向量索引状态表以及计算节点状态表所记载的数据，来确定计算节点 1 加载自身存储的向量组所需要的时间（为零），计算节点 1 加载多个存储节点存储的向量组所需的时间（为零），计算节点 1 加载自身存储的向量组的索引所需要的时间（为零），计算节点 1 加载多个存储节点存储的向量组的索引所需的时间（为零），以及计算节点 1 执行向量生成任务所需要的时间。那么，云管理平台可将这五个时间进行相加（即前述的负载均衡算法），从而得到计算节点 1 的评估值。同样地，云管理平台也可对计算节点 2 等其余计算节点进行类似的操作，故最终云管理平台可得到计算节点 1 以及计算节点 15 2 等所有计算节点的评估值。

在这些计算节点中，由于计算节点 1 的评估值最小（说明计算节点 1 的负载最低，其状态最优），故云管理平台可将计算节点 1 确定为执行向量生成任务的计算节点，并将向量生成任务调度给计算节点 1 执行。

203、第一计算节点基于向量生成请求生成与多个数据对应的多个向量，并将多个向量存储于多个 20 存储节点中。

得到针对多个数据的向量生成请求后，第一计算节点可基于向量生成请求，从多个存储节点处获取多个数据，并对多个数据进行映射（embedding），从而得到与多个数据对应的多个向量，其中，一个数据可对应至少一个向量。然后，第一计算节点可将多个向量写入多个存储节点中（例如，通过内存池写入对象存储池，又如，直接写入对象存储池等等），以使得多个存储节点存储有这多个向量，从而满足 25 租户的向量生成需求。

具体地，第一计算节点可通过以下方式将多个向量存储于多个存储节点中：

得到针对多个数据的向量生成请求后，第一计算节点可生成第一个数据对应的向量，并存储该向量，接着再生成第二个数据对应的向量，并存储该向量，以此类推，当第一计算节点所存储的向量的数量等于预置的数值（该数值的大小可根据实际需求进行设置，此处不做限制）时，第一计算节点可将存储的 30 这些向量划分为一个向量组，并将该向量组存储于对象存储池中的某个存储节点。不断重复前述操作，第一计算节点可将多个数据均转换为相应的向量，并将多个向量划分为多个向量组存储于对象存储池中。

进一步地，第一计算节点每将一个向量组写入对象存储池的某个存储节点后，还可生成该向量组的索引，并将该向量组的索引也写入对象存储池的该存储节点中，如此一来，第一计算节点可将多个向量组的索引均写入对象存储池中。

35 值得注意的是，第一计算节点每得到一个向量组，可按照预置的策略来选择是否将该向量组存储在自身中，无论第一计算节点选择存储该向量组还是选择不存储该向量组，在第一计算节点将该向量组写入对象存储池后，均可将自身已存储的若干个向量组汇报给云管理平台，以使得云管理平台更新向量状态表。需要说明的是，第一计算节点并不会存储其生成的多个向量组，仅会选择性地将多个向量组中的部分向量组（即后续实施例中的第一向量组，而这部分向量组所包含的向量也就是后续实施例中的第二向 40 量）。此外，由于向量组写入对象存储池的过程会经过内存池，故内存池中的存储节点也可存储有多个向量组中的部分向量组。

还值得注意的是，第一计算节点每得到一个向量组的索引，可按照是否已存储该向量组来选择是否将该向量组的索引存储在自身中，无论第一计算节点选择存储该向量组的索引还是选择不存储该向量组的索引，在第一计算节点将该向量组的索引写入对象存储池后，均可将自身已存储的若干个向量组的索引汇报给云管理平台，以使得云管理平台更新向量索引状态表。需要说明的是，第一计算节点仅会选择性 45 地将多个向量组中的部分向量组的索引。此外，由于向量组的索引写入对象存储池的过程会经过内存池，

故内存池中的存储节点也可存储有多个向量组中部分向量组的索引。

依旧如上述例子，计算节点1接收到向量生成任务后，可基于向量生成任务，从对象存储池中读取租户的多个数据，包含数据1至数据36。接着，计算节点1可利用映射模型对数据1进行处理，从而得到数据1对应的向量1，再对数据2进行处理，得到数据2对应的向量2，以此类推，由于6个向量可组成一个向量组，故在得到向量6后，计算节点1可将向量1至向量6作为向量组1，并将向量组1通过内存池写入对象存储池中。接着，计算节点1还可选择将向量组1缓存在自身中，并向云管理平台汇报计算节点1当前已存储的向量组，即向量组1。然后，计算节点1还可生成向量组1的索引，并将向量组1的索引通过内存池写入对象存储池。随后，计算节点1还可选择将向量组1的索引缓存在自身中，并向云管理平台汇报计算节点1当前已存储的向量组的索引，即向量组1的索引。

重复以上操作，计算节点1可继续生成向量7至向量36，即向量组2至向量组5，以及向量组2的索引至向量组5的索引，并先后将向量组2至向量组5，向量组2的索引至向量组5的索引，写入到对象存储池中，此处不再赘述。值得注意的是，设计计算节点1仅能缓存三个向量组，基于预置的策略，计算节点1选择缓存向量组1、向量组2、向量组5以及这三个向量组的索引，并实时汇报给云管理平台，故云管理平台可生成前述表1中关于计算节点1的向量存储情况以及表2中关于计算节点2的向量索引存储情况。

以上是对第一个实施例所进行的介绍，以下将对第二个实施例进行介绍，图4为本申请实施例提供的基于云管理平台的向量处理方法的另一流程示意图，如同4所示，第二个实施例可通过如图1所示的云服务系统实现，云服务系统包含为租户提供云服务的基础设施以及管理这些基础设施的云管理平台，这些基础设施包含服务于租户的多个计算节点以及多个存储节点，多个存储节点用于存储租户的多个数据。第二个实施例包括：

401、云管理平台接收来自租户的针对第一向量的向量查询请求。

本实施例中，当租户需要查询第一向量（即目标向量）时，租户可通过其客户端向云管理平台发送针对第一向量的向量查询请求。接收到针对第一向量的向量生成请求后，云管理平台可确定需要将租户查询并返回第一向量。

例如，如图5所示（图5为本申请实施例提供的云服务系统的另一结构示意图），设租户需要查询目标向量，租户可向云管理平台发送针对这目标向量的向量查询请求。

402、云管理平台基于向量查询请求以及多个计算节点的状态信息，根据负载均衡算法从多个计算节点中确定负载最低的第一计算节点，并向第一计算节点发送向量查询请求，多个计算节点均设置有处理向量查询请求的能力。

得到针对第一向量的向量查询请求后，云管理平台可基于向量查询请求，通过向量状态表、向量索引状态表以及计算节点状态表来获取服务于租户的多个计算节点的状态信息，并基于这多个计算节点的状态信息，根据负载均衡算法从这多个计算节点中选择状态最优（即负载最低的）的计算节点，比如，该计算节点依旧为第一计算节点，故云管理平台可将向量查询请求发送至第一计算节点。需要说明的是，对于多个计算节点中的任意一个计算节点而言，该计算节点在被云管理平台创建时，已被云管理平台设置有处理向量查询请求的能力。

具体地，云管理平台可通过以下方式从多个计算节点中确定第一计算节点：

在接收到向量生成请求后，对于服务于租户的多个计算节点中的任意一个计算节点而言，云管理平台可基于向量状态表、向量索引状态表以及计算节点状态表，获取该计算节点加载自身所存储的向量所需的时间（由于此时该计算节点通常已存储有多个向量组中的部分向量组，故该时间不为零），该计算节点加载多个存储节点所存储的向量所需的时间（由于此时多个存储节点已存储有多个向量组，故该时间不为零），该计算节点加载自身所存储的索引所需的时间（由于此时该计算节点已存储有多个向量组中的部分向量组的索引，故该时间不为零），该计算节点加载多个存储节点所存储的索引所需的时间（由于此时多个存储节点已存储有多个向量组的索引，故该时间不为零），以及该计算节点执行向量查询请求所需的时间等该计算节点的状态信息，那么，云管理平台可通过负载均衡算法对该计算节点的状态信息进行计算，从而得到该计算节点的评估值。

同理，对于多个计算节点除该计算节点之外的其余计算节点，云管理平台也可对其余计算节点执行

如同对该计算节点所执行的操作，故最终云管理平台可得到多个计算节点的评估值。那么，在多个计算节点中，云管理平台可选出评估值最小的计算节点，比如，该计算节点依旧为第一计算节点，其可执行向量查询请求。

依旧如上述例子，云管理平台接收向量查询请求后，可将向量查询请求视为一个向量查询任务，并
5 从向量状态表、向量索引状态表以及计算节点状态表所记载的数据，来确定计算节点 1 加载自身存储的
向量组所需要的时间（不为零），计算节点 1 加载多个存储节点存储的向量组所需的时间（不为零），计
算节点 1 加载自身存储的向量组的索引所需要的时间（不为零），计算节点 1 加载多个存储节点存储的
向量组的索引所需的时间（不为零），以及计算节点 1 执行向量查询任务所需要的时间。那么，云管理
平台可将这五个时间进行相加，从而得到计算节点 1 的评估值。同样地，云管理平台也可对计算节点 2
10 等其余计算节点进行类似的操作，故最终云管理平台可得到计算节点 1 以及计算节点 2 等所有计算节点
的评估值。

在这些计算节点中，由于计算节点 1 的评估值最小，故云管理平台可将计算节点 1 确定为执行向量
查询任务的计算节点，并将向量查询任务调度给计算节点 1 执行。

403、第一计算节点基于向量查询请求，从自身存储的第二向量或多个存储节点存储的多个向量中，
15 获取与第一向量匹配的第三向量，多个向量包含第二向量。

得到针对第一向量的向量查询请求后，第一计算节点可基于向量查询请求，从自身存储的第二向量
（即第一计算节点自身已存储的若干个向量，这若干个向量为多个存储节点所存储的多个向量中的部分
向量）或多个存储节点存储的多个向量中，获取与第一向量相匹配的第三向量（即输出向量，输出向量
与目标向量可相同或相似），并返回给租户的客户端，以供租户使用，从而满足租户的向量查询需求。

20 具体地，第一计算节点可通过以下方式获取第三向量：

基于前述实施例可知，第一计算节点自身存储的若干个向量可以若干个向量组的形式进行存储的，
即第一计算节点自身存储的第二向量可以第一向量组的形式进行存储，且第一向量组也就是对象存储池
存储的多个向量组中的部分向量组。那么，在接收到向量查询请求后，第一计算节点可基于自身已存储
的第一向量组的索引，获取自身已存储的第一向量组，并从第一向量组中检测是否存在与第一向量匹配
25 的第三向量。若存在，说明第一计算节点自身已存储有第三向量，故第一计算节点可从第一向量组中获
取第三向量，并返回给租户。若不存在，说明第一计算节点自身未存储有第三向量，故第一计算节点可
从内存池或对象存储池处获取多个向量组中除第一向量组之外的第二向量组（即多个向量组中，除第一
计算节点存储的若干个向量组之外的其余向量组）以及第二向量组的索引，并基于第二向量组的索引获
取第二向量组，再从第二向量组中获取第三向量，然后将第三向量返回给租户。

30 值得注意的是，当第一计算节点自身未存储有第三向量时，第一计算节点可根据预置的策略来从第
二向量组中确定第三向量所在的第三向量组，并进一步确定是否存储第三向量组，若第一计算节点选择
存储第三向量组，第一计算节点可按照预置的策略，令第三向量组替换自身已存储的若干个向量组（即
第一向量组）中的某个向量组（当然，也可以不替换，而是令第三向量组直接加入第一向量组），并实
时将自身最新存储的若干个向量组（即更新后的第一向量组）汇报给云管理平台，以使得云管理平台更
35 新向量状态表。

相应地，若第一计算节点选择存储第三向量所在的向量组，第一计算节点还可令第三向量所在的第
三向量组的索引替换自身已存储的若干个向量组中的某个向量组的索引（当然，也可以不替换），并实
时将自身最新存储的若干个向量组的索引（即更新后的第一向量组的索引）汇报给云管理平台，以使得
云管理平台更新向量索引状态表。

40 基于此，对于多个计算节点中除第一计算节点之外的其余计算节点，一旦其余计算节点也执行了其
它的向量查询请求，也会使得其余计算节点存储有若干个向量组以及相应的索引，此处不再赘述。

依旧如上述例子，得到向量查询任务后，由于计算节点 1 自身缓存有向量组 1、向量组 2 和向量组
5，当与租户所需的目标向量匹配的是向量 8 时，其位于向量组 2 中，故计算节点 1 可从自身缓存的向
量组 2 中读取向量 8，并返回给租户，当与租户所需的目标向量匹配的是向量 12 时，其位于向量组 3
45 中，故计算节点 1 可从内存池或对象存储池处读取向量组 3、向量组 4、向量组 6 以及相应的索引，以
从向量组 3 中获取向量 12，再返回给租户。

此外，当计算节点 1 未缓存有向量 12 所在的向量组 3 时，计算节点 1 可按照预置的策略选择是否缓存向量组 3，若选择缓存，计算节点 1 可按照预置的策略，选择向量组 3 替代自身已存储的向量组 5，并令向量组 3 的索引替代自身已存储的向量组 5 的索引。此时，计算节点 1 则缓存有向量组 1、向量组 2、向量组 3 以及相应的索引，故计算节点 1 可实时汇报最新的向量存储情况以及索引存储情况给云管理平台，以使得云管理平台更新向量状态表以及向量索引状态表。

以上是对第二个实施例所进行的介绍，以下将对第三个实施例进行介绍，图 6 为本申请实施例提供的基于云管理平台的向量处理方法的另一流程示意图，如同 6 所示，第三个实施例可通过如图 1 所示的云服务系统实现，云服务系统包含为租户提供云服务的基础设施以及管理这些基础设施的云管理平台，这些基础设施包含服务于租户的多个计算节点以及多个存储节点，多个存储节点用于存储租户的多个数据。第三个实施例包括：

601、云管理平台接收针对第一数据的向量增加请求，第一数据为租户写入多个存储节点中的新数据。

本实施例中，当租户向多个存储节点写入某个新数据时，即第一数据，可触发多个存储节点向云管理平台发送针对第一数据的向量增加请求。接收到针对第一数据的向量增加请求后，云管理平台可确定需要将第一数据转换为第四向量（即新向量）。

例如，如图 7 所示（图 7 为本申请实施例提供的云服务系统的另一结构示意图），设租户向对象存储池写入了新数据，触发对象存储池向云管理平台发送针对新数据的向量增加请求。

602、云管理平台基于向量增加请求以及多个计算节点的状态信息，根据负载均衡算法从多个计算节点中确定负载最低的第一计算节点，并向第一计算节点发送向量增加请求，多个计算节点均设置有处理向量增加请求的能力。

得到针对第一数据的向量增加请求后，云管理平台可基于向量增加请求，通过向量状态表、向量索引状态表以及计算节点状态表来获取服务于租户的多个计算节点的状态信息，并基于这多个计算节点的状态信息，根据负载均衡算法从这多个计算节点中选择状态最优（即负载最低的）的计算节点，比如，该计算节点依旧为第一计算节点，故云管理平台可将向量增加请求发送至第一计算节点。需要说明的是，对于多个计算节点中的任意一个计算节点而言，该计算节点在被云管理平台创建时，已被云管理平台设置有处理向量增加请求的能力。

关于步骤 602 的介绍，可参考图 4 所示实施例中步骤 402 的相关说明部分，此处不再赘述。

603、第一计算节点基于向量增加请求，生成与第一数据对应的第四向量，并将第四向量存储于多个存储节点中。

得到针对第一数据的向量增加请求后，第一计算节点可基于向量增加请求，从多个存储节点处获取第一数据，并对第一数据进行映射，从而得到与第一数据对应的第四向量。然后，第一计算节点可将第四向量存储于多个存储节点中，从而自动满足租户的向量增加需求。

具体地，第一计算节点可通过以下方式将第四向量写入多个存储节点中：

在接收到向量增加请求后，第一计算节点可从对象存储池中读取第一数据，并对第一数据进行映射从而得到与第一数据对应的第四向量，并存储第四向量。此时，第一计算节点会判定自身所存储的未划分成向量组的向量（包含第四向量）的数量是否等于预置的数据，若等于，第一计算节点可将这些向量划分为第四向量组（新向量组），并将第四向量组写入对象存储池中。

进一步地，第一计算节点还可生成第四向量组的索引，并将第四向量组的索引写入对象存储池中。

值得注意的是，第一计算节点可根据预置的策略来确定是否存储第四向量组，若第一计算节点选择存储第四向量组，第一计算节点可按照预置的策略，令第四向量组替换自身已存储的若干个向量组（即第一向量组）中的某个向量组（当然，也可以不替换），并实时将自身最新存储的若干个向量组（即更新后的第一向量组）汇报给云管理平台，以使得云管理平台更新向量状态表。

相应地，第一计算节点还可令第四向量组的索引替换自身已存储的若干个向量组中的某个向量组的索引（当然，也可以不替换），并实时将自身最新存储的若干个向量组的索引（即更新后的第一向量组的索引）汇报给云管理平台，以使得云管理平台更新向量索引状态表。

依旧如上述例子，得到向量增加任务后，计算节点 1 可从对象存储池中读取新数据，并对新数据进

行映射,从而得到新向量,并缓存新向量。由于计算节点1自身已缓存有一些还未划分成向量组的向量(包含前面的新向量),当这些向量的数量等于预置的数值时,计算节点1可将这些向量划分为新向量组,即向量组7,并将向量组7写入对象存储池。然后,计算节点1还可生成向量组7的索引,并将向量组7的索引写入对象存储池。

5 此外,计算节点1可按照预置的策略选择是否缓存向量组7,若选择缓存,计算节点1可按照预置的策略,选择向量组7替代自身已存储的向量组5,并令向量组7的索引替代自身已存储的向量组5的索引。此时,计算节点1则缓存有向量组1、向量组2、向量组7以及相应的索引,故计算节点1可实时汇报给云管理平台,以使得云管理平台更新向量状态表以及向量索引状态表。

10 以上是对第三个实施例所进行的介绍,以下将对第四个实施例进行介绍,图8为本申请实施例提供的基于云管理平台的向量处理方法的另一流程示意图,如同8所示,第四个实施例可通过如图1所示的云服务系统实现,云服务系统包含为租户提供云服务的基础设施以及管理这些基础设施的云管理平台,这些基础设施包含服务于租户的多个计算节点以及多个存储节点,多个存储节点用于存储租户的多个数据。第四个实施例包括:

15 801、云管理平台接收针对第二向量的向量删除请求,第二数据为租户从多个存储节点存储的多个数据中删除的数据。

本实施例中,当租户从多个存储节点存储的多个数据中删除某个数据时,即第二数据,可触发多个存储节点向云管理平台发送针对第二数据的向量删除请求。接收到针对第二数据的向量删除请求后,云管理平台可确定需要将第二数据对应的第五向量删除(即待删除向量)。

20 例如,如图9所示(图9为本申请实施例提供的云服务系统的另一结构示意图),设租户从对象存储池中删除了某个数据,触发对象存储池向云管理平台发送针对已删除数据的向量删除请求。

802、云管理平台基于向量删除请求以及多个计算节点的状态信息,根据负载均衡算法从多个计算节点中确定负载最低的第一计算节点,并向第一计算节点发送向量删除请求,多个计算节点均设置有处理向量删除请求的能力。

25 得到针对第二数据的向量删除请求后,云管理平台可基于向量删除请求,通过向量状态表、向量索引状态表以及计算节点状态表来获取服务于租户的多个计算节点的状态信息,并基于这多个计算节点的状态信息,根据负载均衡算法从这多个计算节点中选择状态最优(即负载最低的)的计算节点,比如,该计算节点依旧为第一计算节点,故云管理平台可将向量删除请求发送至第一计算节点。需要说明的是,对于多个计算节点中的任意一个计算节点而言,该计算节点在被云管理平台创建时,已被云管理平台设置有处理向量删除请求的能力。

30 关于步骤802的介绍,可参考图4所示实施例中步骤402的相关说明部分,此处不再赘述。

803、第一计算节点基于向量删除请求,从多个存储节点存储的多个向量中,将与第二数据对应的第五向量删除。

得到针对第二数据的向量删除请求后,第一计算节点可基于向量删除请求,从多个存储节点存储的多个向量中确定与第二数据对应的第五向量,并删除第五向量。

35 具体地,第一计算节点可通过以下方式删除第五向量:

在接收到向量增加请求后,第一计算节点可从对象存储池中所存储的多个向量组中,确定与第二数据对应的第五向量所在的第五向量组(即待删除的向量组),并删除第五向量组,在删除之前,第一计算节点可读取第五向量组中除第五向量之外的第六向量(即待删除的向量组中,除待删除向量之外的其余向量)。此后,第一计算节点可基于第六向量生成第六向量组(新向量组),例如,第一计算节点可第六向量以及其余向量(这些向量可以是第一计算节点所存储未划分为向量组的向量)进行合并,从而得到第六向量组。然后,第一计算节点可将第六向量组写入对象存储池中。

进一步地,第一计算节点还可生成第六量组的索引,并将第六向量组的索引写入对象存储池中。

45 值得注意的是,第一计算节点可根据预置的策略来确定是否存储第六向量组,若第一计算节点选择存储第六向量组,第一计算节点可按照预置的策略,令第六向量组替换自身已存储的若干个向量组(即第一向量组)中的某个向量组(当然,也可以不替换),并实时将自身最新存储的若干个向量组(即更新后的第一向量组)汇报给云管理平台,以使得云管理平台更新向量状态表。

相应地，第一计算节点还可令第六向量组的索引替换自身已存储的若干个向量组中的某个向量组的索引（当然，也可以不替换），并实时将自身最新存储的若干个向量组的索引（即更新后的第一向量组的索引）汇报给云管理平台，以使得云管理平台更新向量索引状态表。

依旧如上述例子，得到向量删除任务后，计算节点 1 可从对象存储池中的向量组 1 至向量组 6 中，
5 确定已删除数据对应的向量 31 位于向量组 6（待删除向量组）中，计算节点 1 可在对象存储池中，删除向量组 6，且自身缓存向量 32 至向量 36。由于计算节点 1 还缓存有其它未划分成向量组的向量，计算节点 1 可从中选取一个向量与向量 32 至向量 36 合并，得到新向量组，即向量组 8，接着计算节点 1 还可生成向量组 8 的索引。然后，计算节点 1 可将向量组 8 以及向量组 8 的索引存储于对象存储池中。

此外，计算节点 1 可按照预置的策略选择是否缓存向量组 8，若选择缓存，计算节点 1 可按照预置
10 的策略，选择向量组 8 替代自身已存储的向量组 5，并令向量组 8 的索引替代自身已存储的向量组 5 的索引。此时，计算节点 1 则缓存有向量组 1、向量组 2、向量组 8 以及相应的索引，故计算节点 1 可实时汇报给云管理平台，以使得云管理平台更新向量状态表以及向量索引状态表。

本申请实施例中，对于租户所触发的向量处理请求，云管理平台可获取服务于租户的各个计算节点的状态信息，并基于此从这多个计算节点中选出状态最优的计算节点来执行向量处理请求。无论向量处理请求为向量生成请求、向量查询请求、向量增加请求以及向量删除请求中的任意一种，从云管理平台的角度上看，多个计算节点的功能并无差别，每个计算节点均可执行向量处理请求，故云管理平台可从多个计算节点中选择状态最优的某个计算节点，以调用该计算节点执行该向量处理请求。由此可见，由于云管理平台具备针对多个计算节点的任务调度功能，可令每个计算节点可执行各种类型的向量处理请求，从而使得多个计算节点的资源在全局上实现共享，更加有效地利用了计算节点的资源，从而在一定程度上降低云服务系统的成本。

以上是对本申请实施例提供的基于云管理平台的向量处理方法所进行的详细说明，以下将对本申请实施例提供的云管理平台进行介绍。图 10 为本申请实施例提供的云管理平台的一个结构示意图，如图 10 所示，云管理平台用于管理提供云服务的基础设施，基础设施包含多个计算节点以及多个存储节点，多个存储节点用于存储租户的多个数据，云管理平台包括：

25 接收模块 1001，用于接收来自租户的针对多个数据的向量生成请求；

处理模块 1002，用于基于向量生成请求以及多个计算节点的状态信息，根据负载均衡算法从多个计算节点中确定负载最低的第一计算节点，多个计算节点均设置有处理向量生成请求的能力。

处理模块 1002，还用于向第一计算节点发送向量生成请求，向量生成请求用于指示第一计算节点生成与多个数据对应的多个向量，并将多个向量存储于多个存储节点中。

30 在一种可能实现的方式中，接收模块 1001，还用于接收来自租户的针对第一向量的向量查询请求；处理模块 1002，还用于基于向量查询请求以及多个计算节点的状态信息，根据负载均衡算法从多个计算节点中确定负载最低的第一计算节点，多个计算节点均设置有处理向量查询请求的能力；处理模块 1002，还用于向第一计算节点发送向量查询请求，向量查询请求用于指示第一计算节点从自身存储的第二向量或多个存储节点存储的多个向量中，获取与第一向量匹配的第三向量，多个向量包含第二向量。

35 在一种可能实现的方式中，接收模块 1001，还用于接收针对第一数据的向量增加请求，第一数据为租户写入多个存储节点中的新数据；处理模块 1002，还用于基于向量增加请求以及多个计算节点的状态信息，根据负载均衡算法从多个计算节点中确定负载最低的第一计算节点，多个计算节点均设置有处理向量增加请求的能力；处理模块 1002，还用于向第一计算节点发送向量增加请求，向量增加请求用于指示第一计算节点生成与第一数据对应的第四向量，并将第四向量存储于多个存储节点中。

40 在一种可能实现的方式中，接收模块 1001，还用于接收针对第二数据的向量删除请求，第二数据为租户从多个存储节点存储的多个数据中删除的数据；处理模块 1002，还用于基于向量删除请求以及多个计算节点的状态信息，根据负载均衡算法从多个计算节点中确定负载最低的第一计算节点，多个计算节点均设置有处理向量删除请求的能力；处理模块 1002，还用于向第一计算节点发送向量删除请求，向量删除请求用于指示第一计算节点从多个存储节点存储的多个向量中，将与第二数据对应的第五向量删除。

45 在一种可能实现的方式中，向量增加请求或向量删除请求来自多个存储节点。

在一种可能实现的方式中，向量生成请求，用于指示：第一计算节点将多个向量划分为多个向量组，

每个向量组所包含向量的数量等于预置的数值；第一计算节点以向量组为单位，将多个向量组分别写入多个存储节点中。

在一种可能实现的方式中，第二向量以第一向量组的形式存储在所述第一计算节点中，向量生成请求，还用于指示：在得到多个向量组后，第一计算节点可从多个向量组中选择第一向量组，并存储第一向量组；第一计算节点向云管理平台通知第一计算节点存储有第一向量组。

在一种可能实现的方式中，第二向量以第一向量组的形式存储在所述第一计算节点中，向量查询请求，用于指示：第一计算节点检测第一向量组中，是否存在与第一向量匹配的第三向量；若第一向量组中存在第三向量，第一计算节点从第一向量组中获取第三向量；或者，若第一向量组中不存在第三向量，第一计算节点从多个存储节点处获取多个向量组中除第一向量组之外的第二向量组，并从第二向量组中获取第三向量，其中，多个向量组基于多个向量划分得到。

在一种可能实现的方式中，第二向量以第一向量组的形式存储在所述第一计算节点中，向量查询请求，还用于指示：第一计算节点从第二向量组中，确定第三向量所在的第三向量组，并基于第三向量组对第一向量组进行更新，得到更新后的第一向量组，更新后的第一向量组包含第三向量组；第一计算节点向云管理平台通知第一计算节点存储有更新后的第一向量组。

在一种可能实现的方式中，向量增加请求，用于指示：第一计算节点生成与第一数据对应的第四向量，并存储第四向量；若第一计算节点存储的未划分成向量组的向量的数量等于预置的数值，第一计算节点将未划分成向量组的向量划分为第四向量组，并将第四向量组写入多个存储节点，未划分成向量组的向量包含第四向量。

在一种可能实现的方式中，第二向量以第一向量组的形式存储在所述第一计算节点中，向量增加请求，还用于指示：第一计算节点基于第四向量组对第一向量组进行更新，得到更新后的第一向量组，更新后的第一向量组包含第四向量组；第一计算节点向云管理平台通知第一计算节点存储有更新后的第一向量组。

在一种可能实现的方式中，多个向量以多个向量组的形式存储在多个存储节点中，向量删除请求，用于指示：第一计算节点从多个存储节点存储的多个向量组中，获取与第二数据对应的第五向量所在的第五向量组，并从多个存储节点存储的多个向量组中，删除第五向量组；第一计算节点基于第五向量组中除第五向量之外的第六向量生成第六向量组，并将第六向量组存储于多个存储节点中。

在一种可能实现的方式中，向量删除请求，还用于指示：第一计算节点基于第六向量组对第一向量组进行更新，得到更新后的第一向量组，更新后的第一向量组包含第六向量组；第一计算节点向云管理平台通知第一计算节点存储有更新后的第一向量组。

在一种可能实现的方式中，处理模块，用于：基于向量生成请求，根据负载均衡算法对多个计算节点的状态信息进行计算，得到多个计算节点的评估值，多个计算节点的评估值用于指示多个计算节点的负载；从多个计算节点中，将评估值最小的计算节点确定为第一计算节点；其中，对于多个计算节点中的任意一个计算节点，计算节点的状态信息包含以下至少一项：计算节点加载自身存储的向量所需的时间，计算节点加载自身存储的向量的索引所需的时间，计算节点加载多个存储节点存储的向量所需的时间，计算节点加载多个存储节点存储的向量的索引所需的时间，以及计算节点执行向量处理请求所需的时间。

在一种可能实现的方式中，云管理平台还包括：分析模块，分析模块用于执行以下至少一项：对计算节点存储的向量以及多个存储节点存储的向量进行分析，得到计算节点加载自身存储的向量所需的时间以及计算节点加载多个存储节点存储的向量所需的时间；对计算节点存储的向量的索引以及多个存储节点存储的向量的索引进行分析，得到计算节点加载自身存储的索引所需的时间以及计算节点加载多个存储节点存储的索引所需的时间；对计算节点的资源使用情况进行分析，得到计算节点执行向量处理请求所需的时间。

需要说明的是，上述装置各模块/单元之间的信息交互、实现过程等内容，由于与本申请方法实施例基于同一构思，其带来的技术效果与本申请方法实施例相同，具体内容可参考本申请实施例前述所示的方法实施例中的叙述，此处不再赘述。

请参阅图 11，图 11 为本申请实施例提供的计算设备的一个结构示意图。如图 11 所示，该计算设备

1100（可用于呈现前述的云管理平台）包括：处理器 1101、存储器 1102、通信接口 1103 和总线 1104，处理器 1101、存储器 1102 与通信接口 1103 通过总线（图中未标注）耦合。存储器 1102 存储有指令，当存储器 1102 中的执行指令被执行时，计算设备 1100 执行上述方法实施例中云管理平台所执行的方法。

计算设备 1100 可以是配置成实施以上方法的一个或多个集成电路，例如：一个或多个特定集成电路（application specific integrated circuit, ASIC），或，一个或多个微处理器（digital signal processor, DSP），或，一个或者多个现场可编程门阵列（field programmable gate array, FPGA），或这些集成电路形式中至少两种的组合。再如，当装置中的单元可以通过处理元件调度程序的形式实现时，该处理元件可以是通用处理器，例如中央处理器（central processing unit, CPU）或其它可以调用程序的处理器。再如，这些单元可以集成在一起，以片上系统（system-on-a-chip, SOC）的形式实现。

处理器 1101 可以是中央处理单元（central processing unit, CPU），还可以是其它通用处理器、数字信号处理器（digital signal processor, DSP）、专用集成电路（application specific integrated circuit, ASIC）、现场可编程门阵列（field programmable gate array, FPGA）或者其它可编程逻辑器件、晶体管逻辑器件，硬件部件或者其任意组合。通用处理器可以是微处理器，也可以是任何常规的处理器。

存储器 1102 可以是易失性存储器或非易失性存储器，或可包括易失性和非易失性存储器两者。其中，非易失性存储器可以是只读存储器（read-only memory, ROM）、可编程只读存储器（programmable ROM, PROM）、可擦除可编程只读存储器（erasable PROM, EPROM）、电可擦除可编程只读存储器（electrically EPROM, EEPROM）或闪存。易失性存储器可以是随机存取存储器（random access memory, RAM），其用作外部高速缓存。通过示例性但不是限制性说明，许多形式的 RAM 可用，例如静态随机存取存储器（static RAM, SRAM）、动态随机存取存储器（DRAM）、同步动态随机存取存储器（synchronous DRAM, SDRAM）、双倍数据速率同步动态随机存取存储器（double data rate SDRAM, DDR SDRAM）、增强型同步动态随机存取存储器（enhanced SDRAM, ESDRAM）、同步连接动态随机存取存储器（synchlink DRAM, SLDRAM）和直接内存总线随机存取存储器（direct rambus RAM, DRAM）。

存储器 1102 中存储有可执行的程序代码，处理器 1101 执行该可执行的程序代码以分别实现前述接收模块和处理模块等模块的功能，从而实现上述基于云管理平台的向量处理方法。也即，存储器 1102 上存有用于执行上述基于云管理平台的向量处理方法的指令。

通信接口 1103 使用例如但不限于网络接口卡、收发器一类的收发模块，来实现计算设备 1100 与其他设备或通信网络之间的通信。

总线 1104 除包括数据总线之外，还可以包括电源总线、控制总线和状态信号总线等。总线可以是快捷外围部件互连标准（peripheral component interconnect express, PCIe）总线，或扩展工业标准结构（extended industry standard architecture, EISA）总线、统一总线（unified bus, Ubus 或 UB）、计算机快速链接（compute express link, CXL）、缓存一致互联协议（cache coherent interconnect for accelerators, CCIX）等。总线可以分为地址总线、数据总线、控制总线等。

请参阅图 12，图 12 为本申请实施例提供的计算设备集群的一个结构示意图。如图 12 所示，该计算设备集群 1200 包括至少一台计算设备 1100。

如图 12 所示，所述计算设备集群 1200 包括至少一个计算设备 1100。计算设备集群 1200 中的一个或多个计算设备 1100 中的存储器 1102 中可以存有相同的用于执行上述基于云管理平台的向量处理方法的指令。

在一些可能的实现方式中，该计算设备集群 1200 中的一个或多个计算设备 1100 的存储器 1102 中也可以分别存有用于执行上述基于云管理平台的向量处理方法的部分指令。换言之，一个或多个计算设备 1100 的组合可以共同执行用于执行上述基于云管理平台的向量处理方法。

需要说明的是，计算设备集群 1200 中的不同的计算设备 1100 中的存储器 1102 可以存储不同的指令，分别用于执行上述云管理平台的部分功能。也即，不同的计算设备 1100 中的存储器 1102 存储的指令可以实现接收模块和处理模块等等模块中的一个或多个模块的功能。

在一些可能的实现方式中，计算设备集群 1200 中的一个或多个计算设备 1100 可以通过网络连接。

其中，所述网络可以是广域网或局域网等等。

请参阅图 13，图 13 为本申请实施例提供的计算机集群中的计算机设备通过网络连接的一个示意图。如图 13 所示，两个计算设备 1100A 和 1100B 之间通过网络进行连接。具体地，通过各个计算设备中的通信接口与所述网络进行连接。

在一种可能的实现方式中，计算设备 1100A 中的存储器中存有执行接收模块等等模块的功能的指令。同时，计算设备 1100B 中的存储器中存有执行处理模块等等模块的功能的指令。

应理解，图 13 中示出的计算设备 1100A 的功能也可以由多个计算设备完成。同样，计算设备 1100B 的功能也可以由多个计算设备完成。

本申请实施例还涉及一种计算机存储介质，该计算机可读存储介质中存储有用于进行信号处理的程序，当其在计算机上运行时，使得计算机执行如图 2、图 4、图 6 或图 8 所示实施例中云管理平台所执行的步骤。

本申请实施例还涉及一种计算机程序产品，该计算机程序产品存储有指令，该指令在由计算机执行时使得计算机执行如图 2、图 4、图 6 或图 8 所示实施例中云管理平台所执行的步骤。

所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统，装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

在本申请所提供的几个实施例中，应该理解到，所揭露的系统，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U 盘、移动硬盘、只读存储器（ROM，read-only memory）、随机存取存储器（RAM，random access memory）、磁碟或者光盘等各种可以存储程序代码的介质。

权 利 要 求 书

1. 一种基于云管理平台的向量处理方法，其特征在于，所述云管理平台用于管理提供云服务的基础设施，所述基础设施包含多个计算节点以及多个存储节点，所述多个存储节点用于存储租户的多个数据，所述方法包括：

5 所述云管理平台接收来自所述租户的针对所述多个数据的向量生成请求；

 所述云管理平台基于所述向量生成请求以及所述多个计算节点的状态信息，根据负载均衡算法从所述多个计算节点中确定负载最低的第一计算节点，所述多个计算节点均设置有处理所述向量生成请求的能力；

10 所述云管理平台向所述第一计算节点发送所述向量生成请求，所述向量生成请求用于指示所述第一计算节点生成与所述多个数据对应的多个向量，并将所述多个向量存储于所述多个存储节点中。

2. 根据权利要求1所述的方法，其特征在于，所述方法还包括：

 所述云管理平台接收来自所述租户的针对第一向量的向量查询请求；

15 所述云管理平台基于所述向量查询请求以及所述多个计算节点的状态信息，根据所述负载均衡算法从所述多个计算节点中确定负载最低的所述第一计算节点，所述多个计算节点均设置有处理所述向量查询请求的能力；

 所述云管理平台向所述第一计算节点发送所述向量查询请求，所述向量查询请求用于指示所述第一计算节点从自身存储的第二向量或所述多个存储节点存储的所述多个向量中，获取与所述第一向量匹配的第三向量，所述多个向量包含所述第二向量。

3. 根据权利要求1或2所述的方法，其特征在于，所述方法还包括：

20 所述云管理平台接收针对第一数据的向量增加请求，所述第一数据为所述租户写入所述多个存储节点中的新数据；

 所述云管理平台基于所述向量增加请求以及所述多个计算节点的状态信息，根据所述负载均衡算法从所述多个计算节点中确定负载最低的所述第一计算节点，所述多个计算节点均设置有处理所述向量增加请求的能力；

25 所述云管理平台向所述第一计算节点发送所述向量增加请求，所述向量增加请求用于指示所述第一计算节点生成与所述第一数据对应的第四向量，并将所述第四向量存储于所述多个存储节点中。

4. 根据权利要求1至3任意一项所述的方法，其特征在于，所述方法还包括：

 所述云管理平台接收针对第二数据的向量删除请求，所述第二数据为所述租户从所述多个存储节点存储的所述多个数据中删除的数据；

30 所述云管理平台基于所述向量删除请求以及所述多个计算节点的状态信息，根据所述负载均衡算法从所述多个计算节点中确定负载最低的所述第一计算节点，所述多个计算节点均设置有处理所述向量删除请求的能力；

 所述云管理平台向所述第一计算节点发送所述向量删除请求，所述向量删除请求用于指示所述第一计算节点从所述多个存储节点存储的所述多个向量中，将与所述第二数据对应的第五向量删除。

35 5. 根据权利要求3或4所述的方法，其特征在于，所述向量增加请求或所述向量删除请求来自所述多个存储节点。

6. 根据权利要求1至5任意一项所述的方法，其特征在于，所述向量生成请求，用于指示：

 所述第一计算节点将所述多个向量划分为多个向量组，每个向量组所包含向量的数量等于预置的数值；

40 所述第一计算节点以向量组为单位，将所述多个向量组分别写入所述多个存储节点中。

7. 根据权利要求6所述的方法，其特征在于，所述第二向量以第一向量组的形式存储在所述第一计算节点中，所述向量生成请求，还用于指示：

 在得到所述多个向量组后，所述第一计算节点可从所述多个向量组中选择所述第一向量组，并存储第一向量组；

45 所述第一计算节点向所述云管理平台通知所述第一计算节点存储有所述第一向量组。

8. 根据权利要求 2 所述的方法, 其特征在于, 所述第二向量以第一向量组的形式存储在所述第一计算节点中, 所述向量查询请求, 用于指示:

所述第一计算节点检测所述第一向量组中, 是否存在与所述第一向量匹配的第三向量;

若所述第一向量组中存在所述第三向量, 所述第一计算节点从所述第一向量组中获取所述第三向量;
5 或者,

若所述第一向量组中不存在所述第三向量, 所述第一计算节点从所述多个存储节点处获取多个向量组中除所述第一向量组之外的第二向量组, 并从所述第二向量组中获取所述第三向量, 其中, 所述多个向量组基于所述多个向量划分得到。

9. 根据权利要求 8 所述的方法, 其特征在于, 所述第二向量以第一向量组的形式存储在所述第一计算节点中, 所述向量查询请求, 还用于指示:

所述第一计算节点从所述第二向量组中, 确定所述第三向量所在的第三向量组, 并基于所述第三向量组对所述第一向量组进行更新, 得到更新后的第一向量组, 所述更新后的第一向量组包含所述第三向量组;

所述第一计算节点向所述云管理平台通知所述第一计算节点存储有所述更新后的第一向量组。

10. 根据权利要求 3 所述的方法, 其特征在于, 所述向量增加请求, 用于指示:

所述第一计算节点生成与所述第一数据对应的第四向量, 并存储所述第四向量;

若所述第一计算节点存储的未划分成向量组的向量的数量等于预置的数值, 所述第一计算节点将所述未划分成向量组的向量划分为第四向量组, 并将所述第四向量组写入所述多个存储节点, 所述未划分成向量组的向量包含所述第四向量。

11. 根据权利要求 10 所述的方法, 其特征在于, 所述第二向量以第一向量组的形式存储在所述第一计算节点中, 所述向量增加请求, 还用于指示:

所述第一计算节点基于所述第四向量组对所述第一向量组进行更新, 得到更新后的第一向量组, 所述更新后的第一向量组包含所述第四向量组;

所述第一计算节点向所述云管理平台通知所述第一计算节点存储有所述更新后的第一向量组。

12. 根据权利要求 4 所述的方法, 其特征在于, 所述多个向量以多个向量组的形式存储在所述多个存储节点中, 所述向量删除请求, 用于指示:

所述第一计算节点从所述多个存储节点存储的所述多个向量组中, 获取与所述第二数据对应的第五向量所在的第五向量组, 并从所述多个存储节点存储的所述多个向量组中, 删除所述第五向量组;

所述第一计算节点基于所述第五向量组中除所述第五向量之外的第六向量生成第六向量组, 并将所述第六向量组存储于所述多个存储节点中。

13. 根据权利要求 12 所述的方法, 其特征在于, 所述向量删除请求, 还用于指示:

所述第一计算节点基于所述第六向量组对所述第一向量组进行更新, 得到更新后的第一向量组, 所述更新后的第一向量组包含所述第六向量组;

所述第一计算节点向所述云管理平台通知所述第一计算节点存储有所述更新后的第一向量组。

14. 根据权利要求 1 至 13 任意一项所述的方法, 其特征在于, 所述云管理平台基于所述向量生成请求以及所述多个计算节点的状态信息, 根据负载均衡算法从所述多个计算节点中确定负载最低的第一计算节点包括:

所述云管理平台基于所述向量生成请求, 根据负载均衡算法对所述多个计算节点的状态信息进行计算, 得到所述多个计算节点的评估值, 所述多个计算节点的评估值用于指示所述多个计算节点的负载;

所述云管理平台从所述多个计算节点中, 将评估值最小的计算节点确定为第一计算节点;

其中, 对于所述多个计算节点中的任意一个计算节点, 所述计算节点的状态信息包含以下至少一项: 所述计算节点加载自身存储的向量所需的时间, 所述计算节点加载自身存储的向量的索引所需的时间, 所述计算节点加载所述多个存储节点存储的向量所需的时间, 所述计算节点加载所述多个存储节点存储的向量的索引所需的时间, 以及所述计算节点执行所述向量处理请求所需的时间。

15. 根据权利要求 14 所述的方法, 其特征在于, 所述方法还包括以下至少一项:

所述云管理平台对所述计算节点存储的向量以及所述多个存储节点存储的向量进行分析, 得到所述

计算节点加载自身存储的向量所需的时间以及所述计算节点加载所述多个存储节点存储的向量所需的时间；

所述云管理平台对所述计算节点存储的向量的索引以及所述多个存储节点存储的向量的索引进行分析，得到所述计算节点加载自身存储的索引所需的时间以及所述计算节点加载所述多个存储节点存储的索引所需的时间；

所述云管理平台对所述计算节点的资源使用情况进行分析，得到所述计算节点执行所述向量处理请求所需的时间。

16. 一种云管理平台，其特征在于，所述云管理平台用于管理提供云服务的基础设施，所述基础设施包含多个计算节点以及多个存储节点，所述多个存储节点用于存储租户的多个数据，所述云管理平台包括：

接收模块，用于接收来自所述租户的针对所述多个数据的向量生成请求；

处理模块，用于基于所述向量生成请求以及所述多个计算节点的状态信息，根据负载均衡算法从所述多个计算节点中确定负载最低的第一计算节点，所述多个计算节点均设置有处理所述向量生成请求的能力；

所述处理模块，还用于向所述第一计算节点发送所述向量生成请求，所述向量生成请求用于指示所述第一计算节点生成与所述多个数据对应的多个向量，并将所述多个向量存储于所述多个存储节点中。

17. 根据权利要求 16 所述的云管理平台，其特征在于，所述接收模块，还用于接收来自所述租户的针对第一向量的向量查询请求；

所述处理模块，还用于基于所述向量查询请求以及所述多个计算节点的状态信息，根据所述负载均衡算法从所述多个计算节点中确定负载最低的所述第一计算节点，所述多个计算节点均设置有处理所述向量查询请求的能力；

所述处理模块，还用于向所述第一计算节点发送所述向量查询请求，所述向量查询请求用于指示所述第一计算节点从自身存储的第二向量或所述多个存储节点存储的所述多个向量中，获取与所述第一向量匹配的第三向量，所述多个向量包含所述第二向量。

18. 根据权利要求 16 或 17 所述的云管理平台，其特征在于，所述接收模块，还用于接收针对第一数据的向量增加请求，所述第一数据为所述租户写入所述多个存储节点中的新数据；

所述处理模块，还用于基于所述向量增加请求以及所述多个计算节点的状态信息，根据所述负载均衡算法从所述多个计算节点中确定负载最低的所述第一计算节点，所述多个计算节点均设置有处理所述向量增加请求的能力；

所述处理模块，还用于向所述第一计算节点发送所述向量增加请求，所述向量增加请求用于指示所述第一计算节点生成与所述第一数据对应的第四向量，并将所述第四向量存储于所述多个存储节点中。

19. 根据权利要求 16 至 18 任意一项所述的云管理平台，其特征在于，所述接收模块，还用于接收针对第二数据的向量删除请求，所述第二数据为所述租户从所述多个存储节点存储的所述多个数据中删除的数据；

所述处理模块，还用于基于所述向量删除请求以及所述多个计算节点的状态信息，根据所述负载均衡算法从所述多个计算节点中确定负载最低的所述第一计算节点，所述多个计算节点均设置有处理所述向量删除请求的能力；

所述处理模块，还用于向所述第一计算节点发送所述向量删除请求，所述向量删除请求用于指示所述第一计算节点从所述多个存储节点存储的所述多个向量中，将与所述第二数据对应的第五向量删除。

20. 根据权利要求 18 或 19 所述的云管理平台，其特征在于，所述向量增加请求或所述向量删除请求来自所述多个存储节点。

21. 根据权利要求 16 至 20 任意一项所述的云管理平台，其特征在于，所述处理模块，用于：

基于所述向量生成请求，根据负载均衡算法对所述多个计算节点的状态信息进行计算，得到所述多个计算节点的评估值，所述多个计算节点的评估值用于指示所述多个计算节点的负载；

从所述多个计算节点中，将评估值最小的计算节点确定为第一计算节点；

其中，对于所述多个计算节点中的任意一个计算节点，所述计算节点的状态信息包含以下至少一项：

所述计算节点加载自身存储的向量所需的时间，所述计算节点加载自身存储的向量的索引所需的时间，所述计算节点加载所述多个存储节点存储的向量所需的时间，所述计算节点加载所述多个存储节点存储的向量的索引所需的时间，以及所述计算节点执行所述向量处理请求所需的时间。

22. 根据权利要求 21 所述的云管理平台，其特征在于，所述云管理平台还包括：分析模块，所述分析模块用于执行以下至少一项：

对所述计算节点存储的向量以及所述多个存储节点存储的向量进行分析，得到所述计算节点加载自身存储的向量所需的时间以及所述计算节点加载所述多个存储节点存储的向量所需的时间；

对所述计算节点存储的向量的索引以及所述多个存储节点存储的向量的索引进行分析，得到所述计算节点加载自身存储的索引所需的时间以及所述计算节点加载所述多个存储节点存储的索引所需的时间；

对所述计算节点的资源使用情况进行分析，得到所述计算节点执行所述向量处理请求所需的时间。

23. 一种计算设备集群，其特征在于，所述计算设备集群包括至少一个计算设备，每个计算设备包括处理器以及存储器：

所述存储器用于存储指令；

所述处理器用于根据所述指令，以使得所述计算设备集群执行权利要求 1 至 15 中任一项所述的方法。

24. 一种计算机存储介质，其特征在于，所述计算机存储介质存储有一个或多个指令，所述指令在由一个或多个计算机执行时使得所述一个或多个计算机实施权利要求 1 至 15 任一所述的方法。

25. 一种计算机程序产品，其特征在于，所述计算机程序产品存储有指令，所述指令在由计算机执行时，使得所述计算机实施权利要求 1 至 15 任意一项所述的方法。