(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 102436408 A
(43)申请公布日 2012.05.02

(21)申请号 201110305170.7

(22)申请日 2011.10.10

(71)申请人 上海交通大学
地址 200240 上海市闵行区东川路 800 号

(72)发明人 邹恒明 金娟 文珊珊

(74)专利代理机构 上海旭诚知识产权代理有限公司 31220
代理人 王萍萍

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

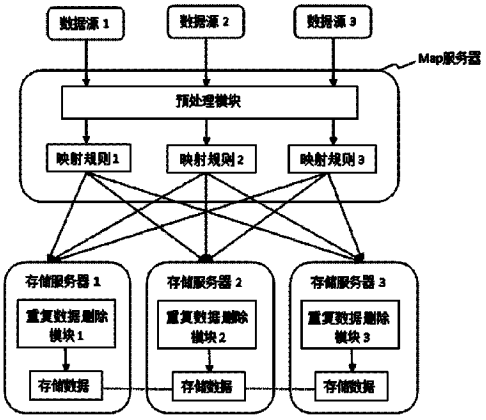
权利要求书 3 页 说明书 8 页 附图 3 页

(54)发明名称

基于 Map/Dedup 的数据存储云化和云备份方法

(57)摘要

本发明公开了一种基于 Map/Dedup 的数据存储云化和云备份方法，包括数据存储云化的步骤和服务器数据备份的步骤，其中数据存储云化的步骤包括：定制 Map 模块的映射规则以及重复数据删除模块的重复数据删除方式；Map 服务器的预处理模块对获取的数据进行预处理，得到结构化数据；Map 服务器将结构化数据按映射规则映射到存储服务器上；重复数据删除模块按重复数据删除方式对数据进行重复数据删除；存储数据。服务器数据备份步骤包括：扫描源文件的历史备份，打开源文件的增量备份的索引文件和内容文件；根据索引文件，建立索引网络；根据索引网络，读出增量备份的各数据块到内存；将数据块的数据与源文件中对应的数据作比较，对此数据块生成新的增量备份的索引文件。



CN 102436408 A

1. 一种基于 Map/Dedup 的数据存储云化和云备份方法，其特征在于，包括数据存储云化的步骤和服务器数据备份的步骤，其中：

所述数据存储云化的步骤进一步包括如下步骤：

步骤 (11) 定制映射模块的映射规则以及重复数据删除模块的重复数据删除方式；

步骤 (12) 映射服务器抓取或者接收数据源的数据；

步骤 (13) 所述映射服务器判断所述数据源的所述数据是否需要进行预处理，如果不需要进行预处理，直接执行步骤 (14)，如果需要进行预处理，则通过所述映射服务器的预处理模块对所述数据进行预处理，得到结构化数据；

步骤 (14) 所述映射模块将所述结构化数据按所述映射规则映射到相应的存储服务器上；

步骤 (15) 所述存储服务器读取所述结构化数据，所述重复数据删除模块按所述重复数据删除方式对所述结构化数据进行重复数据删除；

步骤 (16) 所述存储服务器存储重复数据删除完成后的数据；

所述服务器数据备份的步骤进一步包括如下步骤：

步骤 (21) 所述存储服务器扫描源文件的历史备份，决定是否备份所述源文件及备份方式；

步骤 (22) 所述存储服务器打开所述源文件的增量备份的索引文件和内容文件，以及所述源文件的完全备份文件，以备读取；

步骤 (23) 所述存储服务器根据所述增量备份的所述索引文件，建立索引网络；

步骤 (24) 所述存储服务器根据所述索引网络，从始至末依次将所述增量备份中的各数据块中的数据读出到内存；

步骤 (25) 将所述存储服务器读出的所述数据块中的数据与所述源文件对应位置的数据作比较，为所述数据块生成新的增量备份的索引文件，释放所述数据块所占的内存空间；

步骤 (26) 所述存储服务器释放所述索引网络所占的内存空间。

2. 如权利要求 1 所述的数据存储云化和云备份方法，其中所述步骤 (11) 进一步包括如下步骤：

步骤 (111) 根据所述映射服务器上数据源的存储需求，定制映射规则；

步骤 (112) 根据所述映射规则，在所述映射服务器上生成对应的结构化数据模板；

步骤 (113) 根据所述结构化数据模板的特点，定制重复数据删除模块的重复数据删除方式。

3. 如权利要求 2 所述的数据存储云化和云备份方法，其中所述步骤 (13) 进一步包括如下步骤：

步骤 (131) 根据所述数据源类型，读取所述数据源对应的、由步骤 (112) 生成的结构化数据模板；

步骤 (132) 将所述映射服务器上获取的数据的数据结构与步骤 (131) 读取的所述结构化数据模板进行比对，如果两者相符，则直接执行步骤 (14)，如果两者的不相符，则执行步骤 (133)；

步骤 (133) 将所述获取的数据按照所述结构化数据模板进行预处理，生成结构化数

据。

4. 如权利要求 3 所述的数据存储云化和云备份方法，其中所述步骤 (14) 进

一步包括如下步骤：

步骤 (141) 所述映射服务器根据所述数据源，读取步骤 (111) 中定制的所述映射规则；

步骤 (142) 所述映射服务器根据所述映射规则将步骤 (13) 的所述结构化数据映射到所述存储服务器上。

5. 如权利要求 4 所述的数据存储云化和云备份方法，其中所述步骤 (15) 进一步包括如下步骤：

步骤 (151) 所述存储服务器的所述重复数据删除模块采用一个 Hash 函数为数据块计算校验和，所述 Hash 函数唯一地识别数据；

步骤 (152) 所述重复数据删除模块将所述数据块记录在一张数据表中，所述数据表建立了从所述数据块的校验和到所述数据块的存储位置的映射，以及从所述数据块的校验和到所述数据块的引用次数的映射；

步骤 (153) 当所述数据块在所述存储服务器中已有数据备份时，仅增加所述数据块的引用次数；

步骤 (154) 当数据高度冗余时，以备份服务器、备份虚拟机映像或备份源码库的方式成倍减少空间消耗。

6. 如权利要求 5 所述的数据存储云化和云备份方法，其中所述步骤 (21) 进一步包括如下步骤：

步骤 (211) 所述存储服务器扫描寻找所述源文件的所有历史备份，所述历史备份包括最初的完全备份和最初的完全备份之后的所有增量备份，并将所述所有增量备份存入一个链表；

步骤 (212) 所述存储服务器取出所述链表中的最新一次增量备份，比较所述最新一次增量备份和所述源文件的修改日期，如果所述最新一次增量备份的日期较新，则放弃备份，执行步骤 (22)，否则执行步骤 (213)；

步骤 (213) 备份所述源文件，如果所述源文件以前从未做过备份，则此次执行完全备份。

7. 如权利要求 6 所述的数据存储云化和云备份方法，其中所述步骤 (24) 进一步包括如下步骤：

步骤 (241) 通过所述索引网络找到所述数据块的依赖块，读取所述依赖块；

步骤 (242) 如果所述依赖块是不匹配块，则将所述依赖块对应的所述增量备份的数据读出到内存，执行步骤 (25)；

步骤 (243) 如果所述依赖块是匹配块，继续向所述数据块的依赖块追溯，直到遇到不匹配块，执行步骤 (242)，如果遇到索引文件链中最开始的匹配块，则执行步骤 (25)。

8. 如权利要求 1 或 2 所述的数据存储云化和云备份方法，其中所述映射模块的映射规则包括根据数据类型进行映射、根据数据来源地进行映射、根据数据的结构进行映射以及根据时间进行映射。

9. 如权利要求 8 所述的数据存储云化和云备份方法，其中所述重复数据删除模块的重

复数据删除方式是对大小固定的数据块的重复数据删除。

10. 如权利要求 8 所述的数据存储云化和云备份方法，其中所述重复数据删除模块的重复数据删除方式是对大小可变的数据块的重复数据删除。

基于 Map/Dedup 的数据存储云化和云备份方法

技术领域

[0001] 本发明涉及一种计算机存储领域的海量数据存储和备份方法，尤其涉及一种基于 Map/Dedup（映射 / 重复数据删除）的数据存储云化和云备份方法。

背景技术

[0002] 在当今的信息化、网络化社会里，计算机在工作和生活中扮演着极其重要的角色。越来越多的企业、商家、政府机关和个人通过计算机来获取信息、处理信息，同时将大量的信息以数据文件的形式保存在计算机中。随着信息社会的发展，越来越多的信息被数据化，尤其是伴随着 Internet、云计算、物联网等技术的发展，数据呈爆炸式增长。由此产生的海量数据给以数据为中心的各大中型企业的高效存储带来了新的挑战包括：(1) 面对计算机产生的各行各业的数据，服务器需要进行高效的管理；(2) 面对部分重要的数据丢失或者发生其他不可预见的事故，个人、商家、企业乃至政府机关需要最小化此类事故造成的损失。

[0003] 由此，人们开始关注如何高效存储数据以及如何确保数据完好的问题。面对信息量越来越丰富、数据量越来越大，很多海量数据平台正逐步出现数据存储瓶颈、数据备份恢复庞大而缓慢等问题，而数据存储云化和云备份技术因此应运而生。

[0004] 申请号为 7650331 的美国专利“高效大规模数据处理系统和方法”，针对大规模数据的计算提出了一种高效数据处理系统和方法，即 MapReduce 技术：Map 模块读取输入数据，并对数据进行特定于应用的 Map 操作，生成中间数据值，再由 Reduce 模块整合中间值得到最后计算结果进行输出。MapReduce 是一种用于分布系统的编程模型，支持在计算机集群中对超大数据集进行分布式处理。该系统与方法具有极大的扩展性和极强的容错性，同时为高效地处理海量信息提供了支持，特别适合需要高吞吐量访问的具有超大数据集的应用程序，但是，它仅仅只是应用于大规模数据的查询和计算，并没有为存储领域的海量数据的存储与备份提供很好的解决方案。

[0005] 申请号为 200610001299.8 的中国专利“数据恢复方法”提出了一种数据备份方法，将主计算机的数据备份在至少一台备份机算机上。该方法要求本地与远程之间进行完整的文件交换，从而产生了大量的网络传输，在延长了数据恢复时间的同时还带来了传输安全隐患，更重要的是，反复存储同一文件的不同版本（但这些不同版本的大部分内容是相同的），会导致大量的存储空间浪费。

[0006] 因此，本领域的技术人员致力于开发一种数据存储云化和云备份方法及其系统，使得在保证海量数据高效存储的同时，尽可能高效地利用存储空间，同时保证数据的安全。

发明内容

[0007] 有鉴于现有技术的上述缺陷，本发明所要解决的技术问题是提供一种数据存储云化和云备份方法及其系统，通过将数据存储云化，即通过分布式架构，并且通过并行计算提高采集数据的存储效率；通过备份手段，更好地对海量数据进行备份恢复以保障数据的安

全。

[0008] 为实现上述目的，本发明提供了一种数据存储云化和云备份方法，其特征在于，包括数据存储云化的步骤和服务器数据备份的步骤，其中：

[0009] 所述数据存储云化的步骤包括如下步骤：

[0010] 步骤 11，定制映射 (Map) 模块的映射规则以及重复数据删除 (Dedup) 模块的重复数据删除方式；

[0011] 步骤 12，Map 服务器抓取或者接收数据源的数据；

[0012] 步骤 13，所述 Map 服务器判断所述数据源的所述数据是否需要进行预处理，如果不需要进行预处理，直接执行步骤 (14)，如果需要进行预处理，则通过所述 Map 服务器的预处理模块对所述数据进行预处理，得到结构化数据；

[0013] 步骤 14，所述 Map 模块将所述结构化数据按所述映射规则映射到相应的存储服务器上；

[0014] 步骤 15，所述存储服务器读取所述结构化数据，所述重复数据删除模块按所述重复数据删除方式对所述结构化数据进行重复数据删除；

[0015] 步骤 16，所述存储服务器存储从复数据删除完成后的数据；

[0016] 所述服务器数据备份的步骤包括如下步骤：

[0017] 步骤 21，所述存储服务器扫描源文件的历史备份，决定是否备份所述源文件及备份方式；

[0018] 步骤 22，所述存储服务器打开所述源文件的增量备份的索引文件和内容文件，以及所述源文件的完全备份文件，以备读取；

[0019] 步骤 23，所述存储服务器根据所述增量备份的所述索引文件，建立索引网络；

[0020] 步骤 24，所述存储服务器根据所述索引网络，从始至末依次将所述增量备份中的各数据块中的数据读出到内存；

[0021] 步骤 25，将所述存储服务器将读出的所述数据块中的数据与所述源文件中对应位置的数据作比较，为所述数据块生成新的增量备份的索引文件，释放所述数据块所占的内存；

[0022] 步骤 26，所述存储服务器释放所述索引网络所占的内存空间。

[0023] 进一步地，其中所述步骤 11 进一步包括如下步骤：

[0024] 步骤 111，根据所述 Map 服务器上数据源的存储需求，定制映射规则；

[0025] 步骤 112，根据所述映射规则，在所述 Map 服务器上生成对应的结构化数据模板；

[0026] 步骤 113，根据所述结构化数据模板的特点，定制重复数据删除模块的重复数据删除方式。

[0027] 进一步地，其中所述步骤 13 进一步包括如下步骤：

[0028] 步骤 131，根据所述数据源类型，读取所述数据源对应的、由步骤 112 生成的结构化数据模板；

[0029] 步骤 132，将所述 Map 服务器上获取的数据的数据结构与步骤 131 读取的所述结构化数据模板进行比对，如果两者相符，则直接执行步骤 14，如果两者的不相符，则执行步骤 133；

[0030] 步骤 133，将所述获取的数据按照所述结构化数据模板进行预处理，生成结构化数

据。

[0031] 进一步地，其中所述步骤 14 进一步包括如下步骤：

[0032] 步骤 141，所述 Map 服务器根据所述数据源，读取步骤 111 中定制的所述映射规则；

[0033] 步骤 142，所述 Map 服务器根据所述映射规则将步骤 13 的所述结构化数据映射到所述存储服务器上。

[0034] 进一步地，其中所述步骤 15 进一步包括如下步骤：

[0035] 步骤 151，所述存储服务器的所述重复数据删除模块采用一个 Hash 函数为数据块计算校验和，所述 Hash 函数唯一地识别数据；

[0036] 步骤 152，所述重复数据删除模块将所述数据块记录在一张数据表中，所述数据表建立了从所述数据块的校验和到所述数据块的存储位置的映射，以及从所述数据块的校验和到所述数据块的引用次数的映射；

[0037] 步骤 153，当所述数据块在所述存储服务器中已有数据备份时，仅增加所述数据块的引用次数；

[0038] 步骤 154，当数据高度冗余时，以备份服务器、备份虚拟机映像或备份源码库的方式成倍减少空间消耗。

[0039] 进一步地，其中所述步骤 21 进一步包括如下步骤：

[0040] 步骤 211，所述存储服务器扫描寻找所述源文件的所有历史备份，所述历史备份包括最初的完全备份和最初的完全备份之后的所有增量备份，并将所述所有增量备份存入一个链表；

[0041] 步骤 212，所述存储服务器取出所述链表中的最新一次增量备份，比较所述最新一次增量备份和所述源文件的修改日期，如果所述最新一次增量备份文件的日期较新，则放弃备份，执行步骤 22，否则执行步骤 213；

[0042] 步骤 213，备份所述源文件，如果所述源文件以前从未做过备份，则此次执行完全备份。

[0043] 进一步地，其中所述步骤 24 进一步包括如下步骤：

[0044] 步骤 241，通过所述索引网络找到所述数据块的依赖块，读取所述依赖块；

[0045] 步骤 242，如果所述依赖块是不匹配块，则将所述依赖块对应的所述增量备份的数据读出到内存，执行步骤 25；

[0046] 步骤 243，如果所述依赖块是匹配块，继续向所述块的依赖块追溯，直到遇到不匹配块，执行步骤 242，如果遇到索引文件链中最开始的匹配块，则执行步骤 25。

[0047] 进一步地，其中所述 Map 模块的映射规则包括根据数据类型进行映射、根据数据来源地进行映射、根据数据的结构进行映射以及根据时间进行映射。

[0048] 进一步地，其中所述重复数据删除模块的重复数据删除方式是固定大小块的重复数据删除。

[0049] 进一步地，其中所述重复数据删除模块的重复数据删除方式是变大小块的重复数据删除。

[0050] 在本发明的较佳实施方式中，本发明的数据存储云化和云备份方法，包括数据存储云化的步骤和服务器数据备份的步骤。其中，数据存储云化的步骤涉及两个模块：Map 模

块和重复数据删除模块。Map 服务器根据数据源的存储需求，定制 Map 模块的映射规则以及重复数据删除模块的重复数据删除方式；Map 服务器分别抓取或者接收数据源后，Map 服务器的预处理模块对数据源的数据按存储需求进行预处理，使其成为符合映射规则的结构化数据；Map 服务器根据预先定制的映射规则，将结构化数据映射到对应的各存储服务器上。各个存储服务器依据映射规则获取数据后，使用定制的重复数据删除模块的重复数据删除方式对获取的数据进行重复数据删除，确保每台服务器在存储时删除了冗余数据。其中 Map 模块的映射规则可以定制和变更，重复数据删除模块的重复数据删除方式可以进行置换。服务器数据备份的步骤为：存储服务器扫描源文件的历史备份，决定是否备份源文件及备份方式；存储服务器打开源文件的所有增量备份的索引文件和内容文件，以及完全备份文件，以备读取；存储服务器根据增量备份的索引文件，建立索引网络；存储服务器根据索引网络，从始至末依次读出增量备份中的各数据块中的数据到内存；每读出一个数据块，都和源文件中对应位置的数据作比较，逐块生成新的增量备份的索引文件，并对其中不匹配的部分生成新的增量备份的内容文件；在读取下一个数据块内容之前，先释放这个数据块所占的内存。

[0051] 与现有技术相比，本发明的特点在于：

[0052] (1) 使用 Map 技术提高了海量数据存储效率，满足了海量数据入库的实时性要求。同时，采用重复数据删除技术将结构化数据进行冗余删除，有效地提高了空间利用率；

[0053] (2) 对各个类型的数据源可以制定不同映射规则和重复数据删除方式，具有较高的灵活性，能根据数据的特性以及需求更高效地进行海量数据的存储；

[0054] (3) 数据管理简单方便。本发明的重复数据删除模块在各存储服务器上进行重复数据删除操作，节省了重复数据删除模块设置在 Map 模块之前时的分布式复杂管理的开销；

[0055] (4) 本发明中的存储服务器仅与 Map 模块有联系，而 Map 模块不会限制存储服务器容量的增长。倘若有需求加入新的存储服务器，只需要修改映射规则的配置及网络配置；

[0056] (5) 由于本发明中存储服务器的可扩展性，可带来的经济效应不容忽视。对未来的投资保护上，将解决价格及更新问题。由于智能存储单元数量不受限，将极大降低成本，并随着 PC 技术的发展而发展。

[0057] (6) 本发明非常适合电信、数据中心、游戏运营商的业务需求。本发明使用 Map 技术将海量的源数据分配到各个存储服务器进行并发存储，可以大大提高存储效率，满足数据入库的实时性要求。而且，电信、数据中心、游戏运营商的业务数据大多具有结构化或者半结构化的特性，还能在一定程度上加快本发明的 Map 模块的数据映射速度。

[0058] (7) 在分布式系统中，尤其是在服务器速度成为瓶颈的时候，本发明可以把备份数据发送到多台服务器上，采用分布式并行恢复，无须额外代价，并且可以带来性能的提升。

[0059] 以下将结合附图对本发明的构思、具体结构及产生的技术效果作进一步说明，以充分地了解本发明的目的、特征和效果。

附图说明

[0060] 图 1 是本发明的数据存储云化和云备份方法的原理图。

[0061] 图 2 是本发明的数据存储云化和云备份方法的 Map 模块和重复数据删除模块的工

作流程图。

[0062] 图 3 是本发明的数据存储云化和云备份方法的服务器数据备份的步骤的流程图。

具体实施方式

[0063] 中国上海电信网络优化公司的网优平台在经历了 3 期的建设后，所采集的数据越来越丰富，数据量越来越大，这样一个海量数据平台正逐步出现数据存储瓶颈、数据备份恢复庞大而缓慢等等问题。通过实现数据存储云化和服务器数据备份技术，将数据存储云化，即通过分布式架构，并行计算提高采集数据的存储效率，通过云备份手段，更好的对海量数据进行备份恢复以保障数据的安全。

[0064] 如图 1 所示，本实施例涉及分布式环境下数据存储云化和云备份，包括数据存储云化的步骤和服务器数据备份的步骤，其中数据存储的步骤如图 2 所示，步骤如下：

[0065] 步骤 11：根据各个数据源的不同存储需求，定制 Map 模块的映射规则，定制重复数据删除模块的重复数据删除方式，包括步骤 111-113。

[0066] 步骤 111：根据所述 Map 服务器上数据源的存储需求，定制映射规则。

[0067] 所述 Map 模块的映射规则包括根据数据类型进行映射、根据数据来源地进行映射、根据数据的结构进行映射以及根据时间进行映射。针对客户的数据（按照名字、流量、省、地点、工作负载划分）定制映射规则，如图 1 所示，数据源 1 的映射方式为映射规则 1，数据源 2 的映射方式为映射规则 2，数据源 3 的映射方式为映射规则 3。

[0068] 步骤 112：根据步骤 111 中定制的映射规则，在 Map 服务器上生成对应的结构化数据模板。

[0069] 步骤 113：根据所述结构化数据模板的特点，定制重复数据删除模块的重复数据删除方式。

[0070] 如图 1 所示，指定并配置存储服务器 1、存储服务器 2、存储服务器 3 上的重复数据删除模块 1、重复数据删除模块 2、重复数据删除模块 3 的重复数据删除方式。重复数据删除模块的重复数据删除方式可以是固定大小块的重复数据删除，重复数据删除模块的重复数据删除方式也可以是变大小块的重复数据删除。

[0071] 步骤 12：Map 服务器抓取或者接收数据源的数据。

[0072] 如图 1 所示，Map 服务器从数据源 1、数据源 2 和数据源 3 抓取或者接收数据。

[0073] 步骤 13：Map 服务器判断所述数据源的数据是否需要进行预处理，如果不需要进行预处理，直接执行步骤 14，如果需要进行预处理，则通过 Map 服务器的预处理模块对所述数据进行预处理，得到结构化数据。包括步骤 131-133。

[0074] 步骤 131：Map 服务器的预处理模块根据数据源类型，读取数据源对应的、由步骤 112 生成的结构化数据模板。

[0075] 步骤 132：将 Map 服务器获取的数据的数据结构与步骤 131 读取的结构化数据模板进行比对，如果两者相符，则直接执行步骤 14，如果两者的不相符，则执行步骤 133。

[0076] 步骤 133：将获取的数据按照所述结构化数据模板进行预处理，生成结构化数据。

[0077] 步骤 14：Map 模块将所述结构化数据按所述映射规则映射到相应的存储服务器上，包括步骤 141-142。。

[0078] 步骤 141：Map 服务器根据所述数据源，读取步骤 111 中定制的映射规则。

[0079] 如图 1 所示，Map 服务器读取映射规则 1、映射规则 2、映射规则 3。

[0080] 步骤 142 :Map 服务器根据映射规则将步骤 13 的数据源的数据映射到存储服务器上。

[0081] 如图 1 所示，Map 服务器根据映射规则 1 将数据源 1 的结构化数据分配到存储服务器 1、存储服务器 2、存储服务器 3 上，Map 服务器根据映射规则 2 将数据源 2 的结构化数据分配到存储服务器 1、存储服务器 2、存储服务器 3 上，Map 服务器根据映射规则 3 将数据源 3 的结构化数据分配到存储服务器 1、存储服务器 2、存储服务器 3 上。

[0082] 步骤 15 :存储服务器读取所述结构化数据，存储服务器的重复数据删除模块按重复数据删除方式对结构化数据进行重复数据删除。

[0083] 如图 1 所示，存储服务器 1、存储服务器 2 和存储服务器 3 接收 Map 服务器分配的数据，数据进入重复数据删除模块 1、重复数据删除模块 2 和重复数据删除模块 3。重复数据删除模块 1、重复数据删除模块 2 和重复数据删除模块 3 比对需要存储的数据与磁盘上已持久化的数据来进行重复数据删除。

[0084] 包括步骤 151-154。

[0085] 步骤 151 :重复数据删除模块采用一个 Hash 函数为数据块计算校验和，此 Hash 函数以很高的概率唯一地识别数据。

[0086] 例如，Hash 函数 SHA256 的 Hash 碰撞的概率大约为 2^{-256}。

[0087] 步骤 152 :重复数据删除模块将数据块记录在一张数据表中，所述数据表建立了从所述数据块的校验和到所述数据块的存储位置的映射，以及从所述数据块的校验和到所述数据块的引用次数的映射。

[0088] 步骤 153 :当所述数据块在所述存储服务器中已有数据备份时，仅增加所述数据块的引用次数。

[0089] 步骤 154 :当数据高度冗余时，以备份服务器、备份虚拟机映像或备份源码库的方式成倍减少空间消耗。

[0090] 重复数据删除还可以在文件或者字节层面进行。

[0091] 步骤 16 :所述存储服务器存储重复数据删除完成后的数据。

[0092] 步骤 17 :数据存储的步骤的后期工作包括，根据各个服务器上数据的逻辑关系，提供数据访问接口。

[0093] 如图 3 所示，本实施例的服务器数据备份的步骤涉及源文件的增量备份的备份方法，步骤如下 :

[0094] 步骤 21 :存储服务器扫描源文件的历史备份，决定是否备份及备份方式，包括步骤 211-213。

[0095] 步骤 211 :存储服务器扫描寻找所述源文件的所有历史备份，所述历史备份包括最初的完全备份和最初的完全备份之后的所有增量备份，并将所述所有增量备份存入一个链表。

[0096] 源文件的所有历史备份为在存放备份的目录下列举的所有文件。

[0097] 步骤 212 :存储服务器取出所述链表中的最新一次增量备份，比较最新一次增量备份和源文件的修改日期，如果最新一次增量备份文件的日期较新，则放弃备份，执行步骤 22，否则执行步骤 213。

[0098] 其中，最后一次历史备份可能为完全备份或增量备份。

[0099] 步骤 213 ：备份所述源文件，如果所述源文件以前从未做过备份，则此次执行完全备份。

[0100] 执行完全备份即直接拷贝该文件。

[0101] 步骤 22 ：存储服务器打开源文件的增量备份的索引文件和内容文件，以及源文件的完全备份文件，以备读取。

[0102] 如图 1 中的存储服务器 1、存储服务器 2 和存储服务器 3 打开源文件的所有增量备份的索引文件和内容文件，以及源文件第一次备份时直接拷贝的完全备份文件，以备读取；

[0103] 其中，增量备份的索引文件是以索引文件链对增量备份之间的引用信息的记录。增量备份的内容文件是增量备份中的数据。

[0104] 步骤 23 ：根据步骤 22 中打开的增量备份的索引文件，建立索引网络。

[0105] 其中，索引网络是索引文件中的索引文件链形成的网络结构。

[0106] 步骤 24 ：根据步骤 23 中建立的索引网络，从始至末依次将增量备份中各数据块中的数据读出到内存，包括步骤 241 到 243。

[0107] 步骤 241 ：通过索引网络找到所述数据块的依赖块，读取此依赖块。

[0108] 其中，依赖块是指与所述数据块有相似之处或者完全一致的数据块。

[0109] 步骤 242 ：如果步骤 241 中读取的依赖块是不匹配块，则将该依赖块对应的增量备份的数据读出到内存，执行步骤 25。

[0110] 其中，不匹配块是指与所述数据块不完全一致的依赖块，匹配块是指与所述数据块完全一致的依赖块。

[0111] 步骤 243 ：如果步骤 241 中读取的依赖块是匹配块，继续向所述数据块的依赖块追溯，直到遇到不匹配块，执行步骤 242，如果遇到索引文件链中的最开始的匹配块，则执行步骤 25。

[0112] 如果读取的依赖块长度未达到所述数据块的需要，则在索引文件链中搜寻依赖块的下一个数据块继续读取，直到读出的总长度达到所述数据块的要求。这样，就读到了增量备份中各数据块中的内容。

[0113] 步骤 25 ：存储服务器将读出的所述数据块中的数据与源文件中对应位置的数据作比较，对所述数据块生成新的增量备份的索引文件，释放所述数据块所占的内存。

[0114] 存储服务器每读出一个增量备份中的数据块，都和源文件的对应位置的数据作比较，逐块生成新的增量备份的索引文件，并对其中不匹配的部分生成全新的增量备份的内容文件，然后，在读下一个数据块内容之前，先释放这个数据块所占的内存。

[0115] 步骤 26 ：存储服务器释放索引网络所占的内存空间。

[0116] 本实施例针对不同的数据源类型可以指定不同的映射规则，使映射规则最大可能地符合存储需求，提高存储效率。同时，在映射操作前，对需要映射的数据进行预处理，使数据成为最适合其映射规则的结构化数据，大大增加了映射的效率和能力。另外，在重复数据删除时，选择各个服务器单独重复数据删除，减少了不同存储服务器上管理重复数据的开销。

[0117] 以上详细描述了本发明的较佳具体实施例。应当理解，本领域的普通技术人员无

需创造性劳动就可以根据本发明的构思做出诸多修改和变化。因此，凡本技术领域的技术人员依本发明的构思在现有技术的基础上通过逻辑分析、推理或者有限的实验可以得到的技术方案，皆应在由权利要求书所确定的保护范围内。

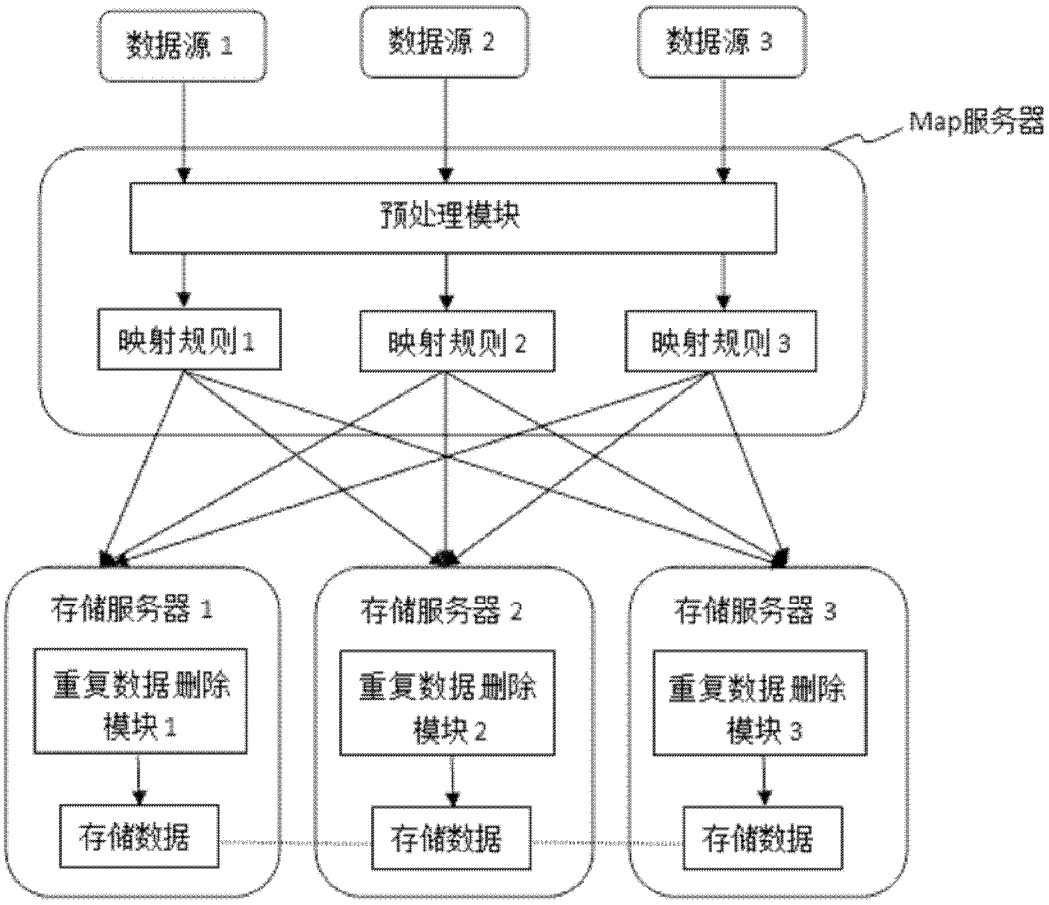

数据源 1
数据源 2
数据源 3
Map服务器
预处理模块
映射规则 1
映射规则 2
映射规则 3
存储服务器 1
存储服务器 2
存储服务器 3
重复数据删除模块 1
重复数据删除模块 2
重复数据删除模块 3
存储数据
存储数据
存储数据

图 1

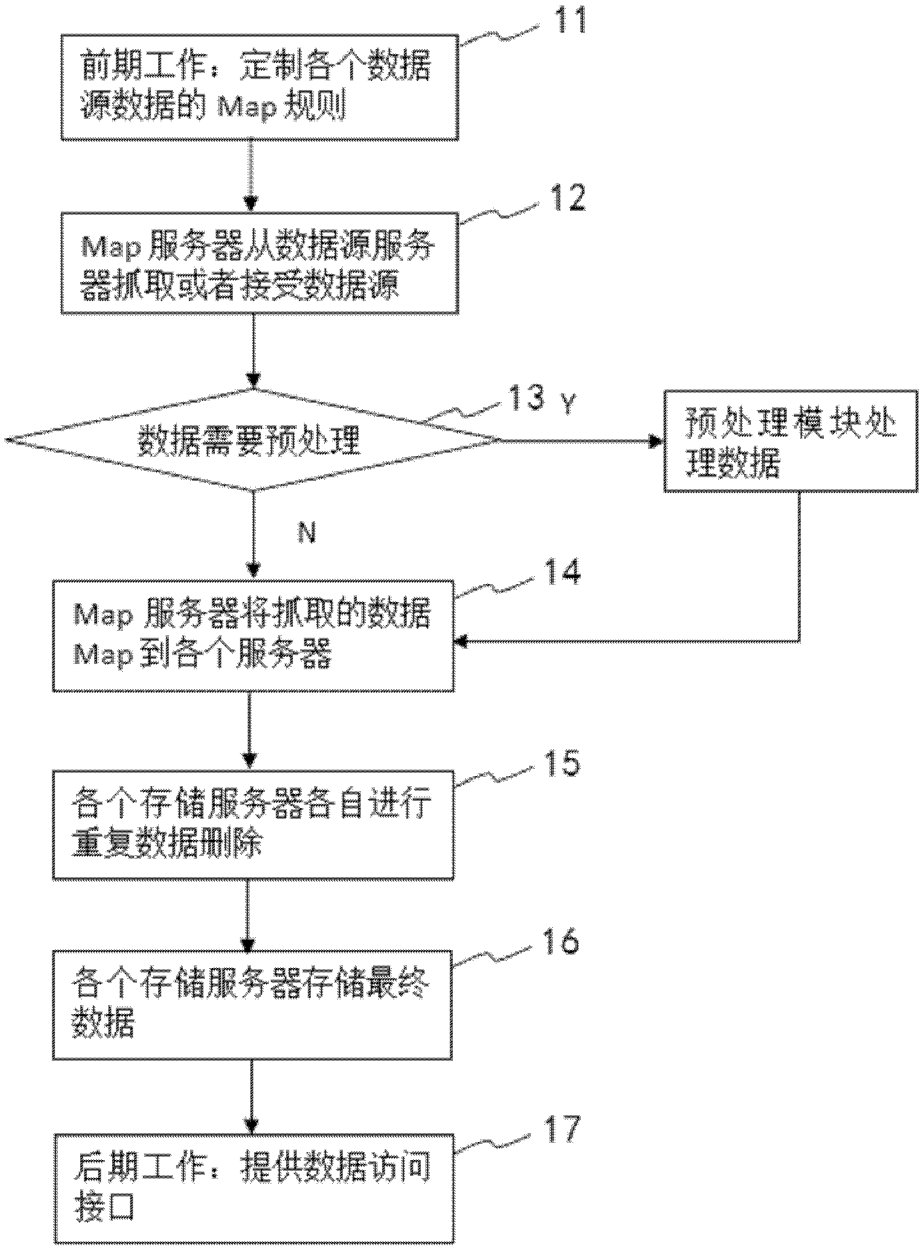

11
前期工作：定制各个数据源数据的 Map 规则
12
Map 服务器从数据源服务器抓取或者接受数据源
13
Y
数据需要预处理
预处理模块处理数据
N
14
Map 服务器将抓取的数据 Map 到各个服务器
15
各个存储服务器各自进行重复数据删除
16
各个存储服务器存储最终数据
17
后期工作：提供数据访问接口

图 2

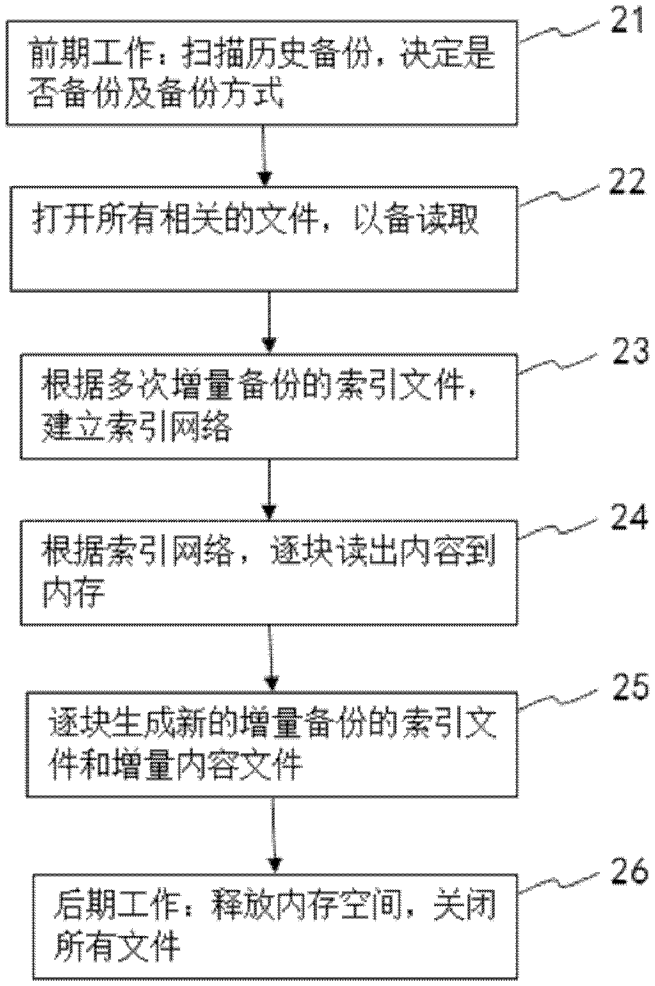

前期工作：扫描历史备份，决定是否备份及备份方式
21
打开所有相关的文件，以备读取
22
根据多次增量备份的索引文件，建立索引网络
23
根据索引网络，逐块读出内容到内存
24
逐块生成新的增量备份的索引文件和增量内容文件
25
后期工作：释放内存空间，关闭所有文件
26

图 3