(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104820700 B
(45)授权公告日 2018.07.20

(21)申请号 201510233801.7

(22)申请日 2015.05.08

(65)同一申请的已公布的文献号
申请公布号 CN 104820700 A

(43)申请公布日 2015.08.05

(73)专利权人 中国南方电网有限责任公司电网技术研究中心
地址 510000 广东省广州市越秀区东风东路水均岗6、8号粤电大厦西塔13-21楼
专利权人 南方电网科学研究院有限责任公司
北京四方继保自动化股份有限公司

(72)发明人 陈浩敏 李鹏 郭晓斌 许爱东 陈波 姚浩 蒋愈勇 张利强 易洋 郭庆武

(74)专利代理机构 广州华进联合专利商标代理有限公司 44224
代理人 周清华

(51)Int.Cl.
G06F 17/30(2006.01)

(56)对比文件
CN 104281130 A,2015.01.14,
CN 104281130 A,2015.01.14,
CN 103399945 A,2013.11.20,
US 2009055429 A1,2009.02.26,
CN 1410915 A,2003.04.16,
CN 103049475 A,2013.04.17,
CN 103810224 A,2014.05.21,
yakcy.非结构化数据的存储与查询.《CSDN博客》.2014,

审查员 李梦诗

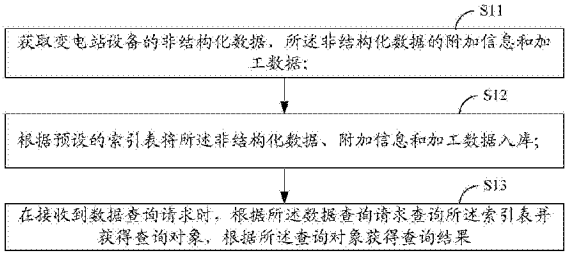
权利要求书1页 说明书6页 附图2页

(54)发明名称

变电站非结构化数据的处理方法

(57)摘要

本发明提供一种变电站非结构化数据的处理方法，包括：获取变电站设备的非结构化数据、所述非结构化数据的附加信息和加工数据；根据预设的表结构将所述非结构化数据、附加信息和加工数据入库；其中，所述表结构的行健包括与所述类型对应的类型掩码、产生时间、数据来源和与所述数据质量对应的质量码，所述表结构的列族包括存储所述非结构化数据的原始数据族、存储所述附加信息的数据描述族和存储所述加工数据的加工数据族；在接收到数据查询请求时，根据所述数据查询请求和所述表结构生成查询对象，根据所述查询对象获得查询结果。本发明的变电站非结构化数据的处理方法，其数据访问处理的速度快，并且便于数据迁移。



CN 104820700 B

1.一种变电站非结构化数据的处理方法，其特征在于，包括如下步骤：

获取变电站设备的非结构化数据、所述非结构化数据的附加信息和加工数据；所述附加信息包括所述非结构化数据的类型、产生时间、数据来源和数据质量，所述加工数据包括对所述非结构化数据及所述附加信息进行处理产生的数据；

根据预设的表结构将所述非结构化数据、附加信息和加工数据入库；其中，所述表结构的行健包括与所述类型对应的类型掩码、产生时间、数据来源和与所述数据质量对应的质量码，所述表结构的列族包括存储所述非结构化数据的原始数据族、存储所述附加信息的数据描述族和存储所述加工数据的加工数据族；

在接收到数据查询请求时，根据所述数据查询请求和所述表结构生成查询对象，根据所述查询对象获得查询结果。

2.根据权利要求1所述的变电站非结构化数据的处理方法，其特征在于，所述根据预设的表结构将所述非结构化数据、附加信息和加工数据入库的步骤包括：

从所述附加信息中获得数据类型、产生时间、数据来源和数据质量，并根据预设的类型掩码、产生时间、数据来源和质量码的排列顺序、预设的字节长度和预设的标识组合生成行健值。

3.根据权利要求1或2所述的变电站非结构化数据的处理方法，其特征在于，所述非结构化数据的类型包括日志数据、告警数据、录波数据、音频数据或视频数据；所述类型掩码为预设的与所述日志数据、告警数据、录波数据、音频数据或视频数据对应的类型标识。

4.根据权利要求1或2所述的变电站非结构化数据的处理方法，其特征在于，所述数据来源包括区域、厂站或设备。

5.根据权利要求2所述的变电站非结构化数据的处理方法，其特征在于，所述根据预设的表结构将所述非结构化数据、附加信息和加工数据入库的步骤包括：

根据预设的表结构将所述非结构化数据、附加信息和加工数据存储在HBase数据库中。

6.根据权利要求1所述的变电站非结构化数据的处理方法，其特征在于，在预设的基于Hadoop的并行运算框架中根据所述数据查询请求和所述表结构生成查询对象，根据所述查询对象获得查询结果。

7.根据权利要求6所述的变电站非结构化数据的处理方法，其特征在于，所述在预设的基于Hadoop的并行运算框架中根据所述数据查询请求和所述表结构生成查询对象，根据所述查询对象获得查询结果的步骤包括：

根据所述数据查询请求从HBase数据库中读取非结构化数据表获得数据源；

调用Hadoop分布式系统中的并行模型MapReduce，其中，所述并行模型MapReduce包括Map作业模块和Reduce作业模块；

将所述数据源输入至Map作业模块，通过所述Map作业模块筛选出满足条件的非结构化数据，并对筛选出满足条件的非结构化数据执行预设的数据处理方法；

将Map作业模块中执行预设的数据处理方法后的结果输入至Reduce作业模块，通过所述Reduce作业模块进行汇总，将汇总结果输出后，并写入所述HBase数据库中的非结构化数据表中的加工数据族字段。

变电站非结构化数据的处理方法

技术领域

[0001] 本发明涉及数据处理技术领域，特别是涉及一种变电站非结构化数据的处理的方法。

背景技术

[0002] 为满足社会日益增长的用电需求，电网企业不断壮大，变电设备成倍增长，变电站设备运维管理利用集中采集、智能分析、智能控制、智能展示等技术实现变电站内运行监控、设备运维管理、环境管理等功能，为变电站运行人员、管理人员、检修人员提供“一站式”的决策支持服务。变电站设备运维管理是一个综合的监控管理系统，其采集的信息多而全，其中不但包含传统结构化的数据，如状态监测数据等，而且包含非结构化数据，比如录波、视频等，这类数据的特点是无法用统一的结构来表示，通常以文件的形式存放。为了对其进行统一的管理，在传统关系库中对其的处理通常以压缩块的方式存放，所以对此类数据的存取需要进行压缩和解压，进而造成在传统关系库中对此类数据的处理和访问的功能局限访问效率低下。

[0003] 由于非结构化数据的容量相较前者来说更大，随着时间推移，所占用的磁盘空间将会变得非常庞大，可扩展性也是亟需解决的问题。采用商业库有一些应对此类问题的折衷方案，例如存储近几年的数据，更早的数据导出以文件形式存放等。此类方式，对于数据做长周期的、复杂的统计分析来说，需要进行备份数据的重新载入，数据迁移代价比较大。

发明内容

[0004] 基于此，本发明提供一种变电站非结构化数据的处理方法，其数据访问处理的速度快，并且便于数据迁移。

[0005] 一种变电站非结构化数据的处理方法，包括如下步骤：

[0006] 获取变电站设备的非结构化数据、所述非结构化数据的附加信息和加工数据；所述附加信息包括所述非结构化数据的类型、产生时间、数据来源和数据质量；

[0007] 根据预设的表结构将所述非结构化数据、附加信息和加工数据入库；其中，所述表结构的行健包括与所述类型对应的类型掩码、产生时间、数据来源和与所述数据质量对应的质量码，所述表结构的列族包括存储所述非结构化数据的原始数据族、存储所述附加信息的数据描述族和存储所述加工数据的加工数据族；

[0008] 在接收到数据查询请求时，根据所述数据查询请求查询所述表结构并获得查询对象，根据所述查询对象获得查询结果。

[0009] 本发明的变电站非结构化数据的处理方法，表结构的行健包括类型掩码、产生时间、数据来源和质量码，列族包括原始数据族、数据描述族和加工数据族，因此在数据格式定义阶段有效的将原始数据、描述数据进行了数据模式的设计，将非结构化数据进行统一的管理并纳入同一个处理框架，克服了传统技术中采用文件系统存储处理非结构化数据在可扩展性、统一管理性方面的不足。该方法具备可扩展性，数据规模理论上无限制，可以存

储电网运行产生的长周期非结构化数据，其数据访问处理的速度快，并且便于数据迁移。

附图说明

[0010] 图1为本发明变电站非结构化数据的处理方法在一实施例中的流程示意图。

[0011] 图2为行健的示意图。

[0012] 图3为并行计算框架的示意图。

具体实施方式

[0013] 下面结合实施例及附图对本发明作进一步详细说明，但本发明的实施方式不限于此。

[0014] 如图1所示，是本发明一种变电站非结构化数据的处理方法的流程示意图，包括如下步骤：

[0015] S11、获取变电站设备的非结构化数据、所述非结构化数据的附加信息和加工数据；所述附加信息包括所述非结构化数据的类型、产生时间、数据来源和数据质量；

[0016] S12、根据预设的表结构将所述非结构化数据、附加信息和加工数据入库；其中，所述表结构的行健包括与所述类型对应的类型掩码、产生时间、数据来源和与所述数据质量对应的质量码，所述表结构的列族包括存储所述非结构化数据的原始数据族、存储所述附加信息的数据描述族和存储所述加工数据的加工数据族；

[0017] S13、在接收到数据查询请求时，根据所述数据查询请求查询所述表结构并获得查询对象，根据所述查询对象获得查询结果；

[0018] 本实施例的变电站非结构化数据的处理方法中设计的表结构，其行健包括类型掩码、产生时间、数据来源和质量码，其列族包括原始数据族、数据描述族和加工数据族，本实施例的方法在数据格式定义阶段有效的将原始数据、描述数据进行了数据模式的设计，将非结构化数据进行统一的管理并纳入同一个处理框架，克服了传统技术中采用文件系统存储处理非结构化数据在可扩展性、统一管理性方面的不足。该方法具备可扩展性，数据规模理论上无限制，可以存储电网运行产生的长周期非结构化数据，其数据访问处理的速度快，并且便于数据迁移。

[0019] 对于步骤S11、获取变电站设备的非结构化数据、所述非结构化数据的附加信息和加工数据；所述附加信息包括所述非结构化数据的类型、产生时间、数据来源和数据质量；

[0020] 非结构化数据是指变电站设备中产生的包括日志数据、告警数据、录波数据、音频数据、视频数据等原始数据，附加信息是指该变电站设备在产生非结构化数据时附加的数据，加工数据是指根据不同用户设定的数据处理方法对非结构化数据进行处理后得到的二次加工数据；其中，非结构化数据及其附加信息是由变电站设备产生，而加工数据是在对非结构化数据及其附加信息进行一定的处理上产生的，在数据初始阶段不一定产生有加工数据。

[0021] 对于步骤S12、根据预设的表结构将所述非结构化数据、附加信息和加工数据入库；其中，所述表结构的行健包括与所述类型对应的类型掩码、产生时间、数据来源和与所述数据质量对应的质量码，所述表结构的列族包括存储所述非结构化数据的原始数据族、存储所述附加信息的数据描述族和存储所述加工数据的加工数据族；

[0022] 为了对非结构化数据进行有效集中管理，需要将其进行统一的存储模式设计；HBase中对于数据形态没有严格的定义，数据记录可能包含不同的列、不确定的大小。存储数据使用四维坐标系统：行健、列族、列限定符和时间版本。只有行键是一种从行的方向有效筛选数据集提高命中准确率和查询效率的元素，对其设计基于非结构数据预期的访问模式来建模，因此本实施例根据非结构化数据的特点对行键进行特殊设计。

[0023] 非结构化数据的类型，可以为巡维数据中的日志数据、告警数据、录波数据、音频数据、视频数据等；在表结构的行健中，可采用预设的标识作为类型掩码字段值；

[0024] 非结构化数据的产生时间，可为电力系统产生此非结构化数据的时间戳，该时间一般由产生该数据的设备附加在数据上；在表结构的行健中，存储到库中这一字段时可采用UNIX时间戳；

[0025] 非结构化数据的数据来源，可为标识此数据的来源，最低到设备级别，其中可包含三个子字段：区域、厂站或设备；在表结构的行健中，可采用预设的标识作为数据来源字段值；

[0026] 非结构化数据的数据质量，可从数据的合法性(包括好、无效、未定义、可疑)、故障、旧数据、操作员闭锁等方面来描述数据的质量，在表结构的行健中，可采用掩码的数据结构来定义质量码，用预设的标识作为数据质量字段值。

[0027] 在列的设计上，采用三个列族：原始数据族、数据描述族和加工数据族；同一列族的数据在物理上存储在同一个存储区域下；此列族的设计考虑将来针对非结构化数据的预期处理场景。

[0028] 原始数据族存储非结构化数据的本体内容，按照字节流的方式存储。此列族为非结构化数据原始内容，作为数据的导出及自定义数据分析挖掘方法的应用的输入。该列被单独划分为一列族，是因为非结构化数据一般数据容量较大，对于多数查询场合用户一般更关心加工出来的二次熟数据和其描述数据；而本体内容则更多被数据分析挖掘用户使用。如果原始数据列与其他列划分为一个列族，由于列族内的数据物理上在一起，对于只查询描述数据和熟数据场合，效率将非常低下，将其独立出来有助于提高查询和分析的效率。

[0029] 数据描述族中，则是对此非结构化数据的附加信息进行记录，包括非结构化数据内容的格式描述文件，数据的大小等。即使相同类型的非结构化数据，其内容的格式也是不同的，对应的格式描述文件保证了在解析文件内容时能生成解析器对象，从而保证在并行处理海量非结构化数据的方法的普适性。

[0030] 加工数据族：存储对非结构化数据的二次加工数据，由于各种定制的处理方法对非结构化数据的处理流程不同，其输出产生的结果也不同。而列族内的列是稀疏和可定制的，所有处理结果的输出可存储于此列族，这点保证了方法的可扩展性。

[0031] 在一较佳实施例中，所述根据预设的表结构将所述非结构化数据、附加信息和加工数据入库的步骤包括：

[0032] 从所述附加信息中获得数据类型、产生时间、数据来源和数据质量，并根据预设的类型掩码、产生时间、数据来源和质量码的排列顺序、预设的字节长度和预设的标识组合生成行健值；

[0033] 在本实施例中，行键设计采用组合各种固定长度的字段形成总的键，使得主键具有多字段索引能力。采用固定长度分割而不是分隔符对各个字段进行分割，是因为采用任

何的分隔符都可能会跟索引字段中的值重复，很可能造成数据解析的错误。而采用固定长度的字段则语义明确，有利于后续的查询与解析。

[0034] 如图2所示，是本实施例中行健的示意图，该行健包括由数据类型、产生时间、数据来源和数据质量按顺序组合拼接构成的主键，其中类型掩码1字节、产生时间8字节、数据来源24字节、质量码1字节；各字段排列顺序和字节大小可根据实际情况而设定。、接着，将所述行健值存储在所述表结构中的其中一行行健中，将与所述附加信息对应的非结构化数据存储在与该行行健对应的所述原始数据族中、将所述附加信息存储在与该行行健对应的所述数据描述族中，并将与所述附加信息对应的加工数据存储在与该行行健对应的所述加工数据族中。

[0035] 在一较佳实施例中，所述根据预设的表结构将所述非结构化数据、附加信息和加工数据入库的步骤包括：

[0036] 根据预设的表结构将所述非结构化数据、附加信息和加工数据存储在HBase数据库中。

[0037] 对于步骤S13、在接收到数据查询请求时，根据所述数据查询请求查询所述表结构并获得查询对象，根据所述查询对象获得查询结果；

[0038] 在获得数据查询请求时，查询所述表结构中与所述数据查询请求对应的行健，获取与所述对应的行健同一行的列族中存储的非结构化数据、附加信息和加工数据，得到查询结果。

[0039] 在一较佳实施例中，在预设的基于Hadoop的并行运算框架中根据所述数据查询请求查询所述表结构并获得查询对象，根据所述查询对象获得查询结果；

[0040] 所述在预设的基于Hadoop的并行运算框架中根据所述数据查询请求查询所述表结构并获得查询对象，根据所述查询对象获得查询结果的步骤包括：

[0041] 根据所述数据查询请求从所述HBase数据库中读取非结构化数据表获得数据源；

[0042] 调用Hadoop分布式系统中的并行模型MapReduce，其中，所述并行模型MapReduce包括Map作业模块和Reduce作业模块；

[0043] 将所述数据源输入至Map作业模块，通过所述Map作业模块筛选出满足条件的非结构化数据，并对筛选出满足条件的非结构化数据执行预设的数据处理方法；

[0044] 将Map作业模块中执行预设的数据处理方法后的结果输入至Reduce作业模块，通过所述Reduce作业模块进行汇总，将汇总结果输出后，并写入所述HBase数据库中的非结构化数据表中的加工数据族字段。

[0045] Hadoop，即分布式系统基础架构；Hadoop系统中，MapReduce的输入为存储于HDFS上的文件，文件的格式可为文本数据、键值对文本数据、二进制数据。本发明结合前文所设计的数据模式，结合MapReduce的工作机制，本实施例采用预设的基于原生Hadoop系统与HBase的并行计算框架，如图3所示，是该并行计算框架的示意图；

[0046] 在MapReduce的处理流程中，在数据准备阶段将非结构化数据表作为数据源，将表中由用户需要处理的非结构化数据作为MapReduce任务的输入。非结构化数据的范围确定，需要用户前文所述的行键定义元素(非结构化类型、日期、数据来源等)定义查询条件，形成自定义扫描对象，筛选满足条件的非结构化数据作为后续并行处理算法的输入。

[0047] Map处理过程中，对筛选出的非结构化数据的内容执行用户自定义算法。一个数据

区域上面执行一个Map任务。因为非结构化数据内容格式不固定，在进行处理时需要同时接入数据描述族的对应的格式描述文件，生成相应的解析器对象，对数据内容进行解析处理。Map任务执行算法的主要部分，一般包括数据的解析与处理。Map的算法是可定制的，用户只需继承Mapper接口，即可在函数体中实现算法实体部分。

[0048] 在Reduce阶段，直接接受来自Map的输出，对非结构化数据分析统计的结果进行汇总，并将结果根据主键，写回该条非结构化数据的二次加工族中字段。如果Reduce阶段的任务比较简单，没有汇总的需求，可将写回表的功能移至Mapper类。

[0049] 本实施例中的分布式处理框架，节省了集中式数据库处理进行数据迁移、网络交换、临时空间的代价，提高了海量非结构化数据的处理能力和效率，具有较强的适用性和经济性。

[0050] 接下来再通过一具体实施例详细阐述本发明方法的实施过程。

[0051] (1)定义录波数据的存储模式

[0052] a)定义如图2所示的行健；

[0053] 行键在库中存储的格式为字节数组，因此在变电站设备中，行键的生成与解析均需要根据此格式定义执行。

[0054] b)利用与变电站设备的接口，将非结构化数据表的列族定义中加入原始数据族、数据描述族和加工数据族；

[0055] (2)导入录波数据

[0056] 由录波文件入库代理调用入库接口：

[0057] Bool Upload(int Type,long time,long AreaID,long StationID,long DeviceID,byte QcodeMask)；

[0058] 其中Type为非结构化数据的类型，这里为录波文件对应的类型值；Time为Unix时间戳；Areaid为所属行政变电巡维中心ID；Stationid为变电站ID；DeviceID为设备ID；QcodeMask为质量码；

[0059] 通过此接口所有产生的录波文件入库。

[0060] (3)录波文件的批处理

[0061] a)扫描准备

[0062] 通步骤3类似通过设置起始行键、包含原始数据族和数据描述族，准备扫描对象，选择感兴趣的录波文件进行处理。

[0063] b)Mapper方法实现

[0064] 本实施例中，通过预设的Mapper方法将录波文件的内容进行解析，并根据用户需求，抽取某个电压的时间序列进行小波变换，并将分解后前八阶的系数均值进行传递。

[0065] Map作业模块的输入为[k1,v1]，其中k1为行键的类型，v1为扫描的检索结果。Mapper方法过程如下：

[0066] i).得到原始数据族的文件列，即录波文件内容列；

[0067] ii).得到数据描述族的格式规范列，及录波文件格式列；

[0068] iii).根据格式规范生成录波文件解析器，并取出指定电压通道时间序列；

[0069] iv).对此时间序列进行小波分解，将前八阶系数数组交给输出；

[0070] 而Mapper的输出为<k2,v2>，k2为行键，v2为执行上述方法解析录波文件后生成的

小波系数数组。

[0071] c)Reducer方法实现

[0072] Reducer方法的目的是将分解后前八阶的系数均值存储于加工处理列族的分解系数列中。Reducer作业模块的输入为[k2,v2];同上步,方法是将结果输出到加工数据列族中的分解系数列中。

[0073] 4)查询录波文件

[0074] 通过生成查询的起始行键,并在扫描中配置查询结果应包含的列族生成新的查询对象扫描:

[0075] StartKey=GenKey(Type time,AreaID,long StationidStationID,long DeviceidDeviceID,byte QcodemaskQcodeMask);

[0076] EndKey=GenKey(Type,time,AreaID,StationID,DeviceID,byte QcodeMask);

[0077] 通过扫描获得结果集并对结果集进行遍历。用户可以通过只选择原始数据族导出感兴趣的录波数据,也可以根据录波文件的描述列族数据描述族对录波文件进行统计,还可以通过加工数据族对录波分析结果进行分析和挖掘。

[0078] 以上所述实施例仅表达了本发明的几种实施方式,其描述较为具体和详细,但并不能因此而理解为对本发明专利范围的限制。应当指出的是,对于本领域的普通技术人员来说,在不脱离本发明构思的前提下,还可以做出若干变形和改进,这些都属于本发明的保护范围。因此,本发明专利的保护范围应以所附权利要求为准。

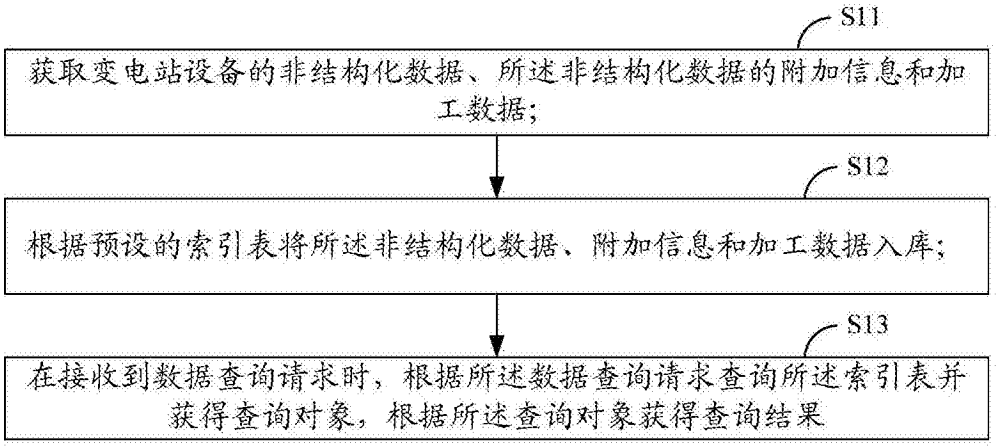

S11
获取变电站设备的非结构化数据、所述非结构化数据的附加信息和加工数据；
S12
根据预设的索引表将所述非结构化数据、附加信息和加工数据入库；
S13
在接收到数据查询请求时，根据所述数据查询请求查询所述索引表并获得查询对象，根据所述查询对象获得查询结果

图1


类型掩码
（1字节）
产生时间
（8字节）
数据来源（24字节）
区域
（8字节）
厂站
（8字节）
设备
（8字节）
质量码
（1字节）

图2

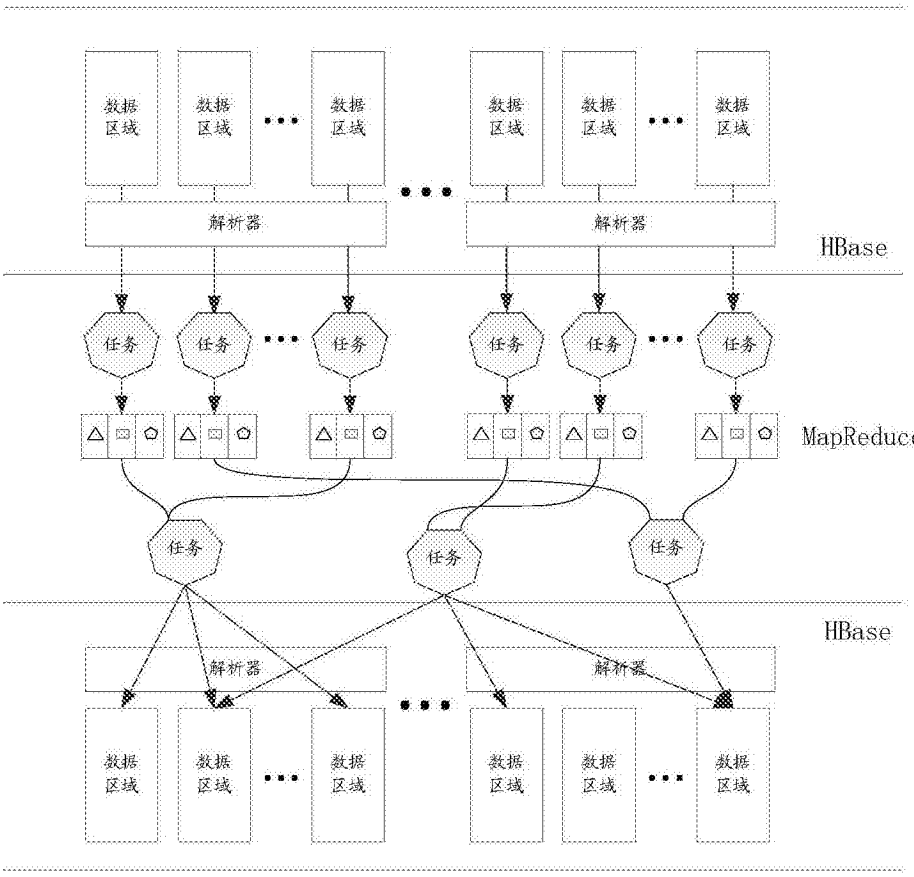


图3