(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113836131 B
(45)授权公告日 2024.02.02

(21)申请号 202111151699.8
(22)申请日 2021.09.29
(65)同一申请的已公布的文献号
申请公布号 CN 113836131 A
(43)申请公布日 2021.12.24
(73)专利权人 平安科技(深圳)有限公司
地址 518000 广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼
(72)发明人 吴智炜
(74)专利代理机构 深圳市世联合知识产权代理有限公司 44385
专利代理师 杨晖琼
(51)Int.Cl.
G06F 16/215(2019.01)
G06F 9/50(2006.01)
(56)对比文件
CN 107562555 A,2018.01.09
WO 2017162083 A1,2017.09.28
CN 109753496 A,2019.05.14
WO 2021072885 A1,2021.04.22
WO 2021179481 A1,2021.09.16
审查员 吴双

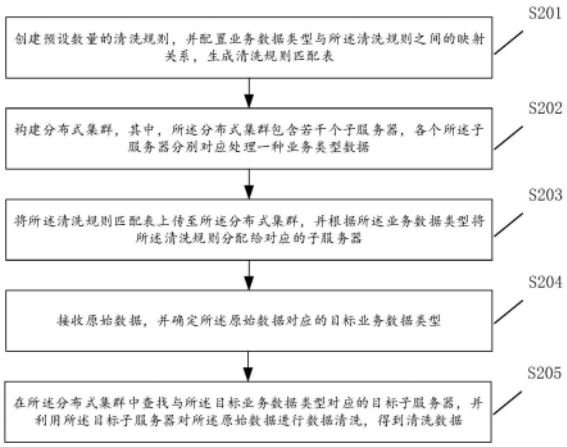
权利要求书2页 说明书13页 附图2页

(54)发明名称

一种大数据清洗方法、装置、计算机设备及存储介质

(57)摘要

本申请公开了一种大数据清洗方法、装置、计算机设备及存储介质，属于大数据技术领域。本申请配置业务数据类型与清洗规则之间的映射关系，生成清洗规则匹配表，构建分布式集群，其中，分布式集群包含若干个子服务器，各个子服务器分别对应处理一种业务类型数据，根据业务数据类型将清洗规则分配给对应的子服务器，确定原始数据对应的目标业务数据类型，查找与目标业务数据类型对应的目标子服务器，并在目标子服务器对原始数据进行数据清洗，得到清洗数据。此外，本申请还涉及区块链技术，原始数据可存储于区块链中。本申请通过构建分布式集群实现不同类型业务数据自动清理，具有较强的通用性和适应性，同时有利于数据清理规则的统一管理。



CN 113836131 B

1.一种大数据清洗方法,其特征在于,包括:

创建预设数量的清洗规则,并配置业务数据类型与所述清洗规则之间的映射关系,生成清洗规则匹配表;

构建分布式集群,其中,所述分布式集群包含若干个子服务器,各个所述子服务器分别对应处理一种业务类型数据;

将所述清洗规则匹配表上传至所述分布式集群,并根据所述业务数据类型将所述清洗规则分配给对应的子服务器;

接收原始数据,并确定所述原始数据对应的目标业务数据类型;

在所述分布式集群中查找与所述目标业务数据类型对应的目标子服务器,并利用所述目标子服务器对所述原始数据进行数据清洗,得到清洗数据;

所述在所述分布式集群中查找与所述目标业务数据类型对应的目标子服务器,并利用所述目标子服务器对所述原始数据进行数据清洗,得到清洗数据的步骤,具体包括:

获取每一个子服务器对应的业务数据标签;

在所述分布式集群中,将所述目标业务数据类型与每一个子服务器对应的业务数据标签一一进行比对;

根据比对结果确定所述原始数据对应的目标子服务器,并通过所述目标子服务器对所述原始数据进行数据清洗,得到清洗数据。

2.如权利要求1所述的数据清洗方法,其特征在于,所述接收原始数据,并确定所述原始数据对应的目标业务数据类型的步骤,具体包括:

接收原始数据,将所述原始数据导入所述分布式集群,并确定所述原始数据中的待清洗字段;

提取所述待清洗字段的关键词,并基于所述关键词确定所述原始数据对应的目标数据类型。

3.如权利要求2所述的数据清洗方法,其特征在于,所述接收原始数据,将所述原始数据导入所述分布式集群,并确定所述原始数据中的待清洗字段的步骤,具体包括:

获取所述原始数据对应的需求文档,其中,所述需求文档中记录了数据清洗的具体要求;

识别所述原始数据的数据结构,得到所述原始数据的结构信息;

基于所述结构信息对原始数据进行分割,得到若干个数据字段;

对每一个所述数据字段进行语义识别,并基于语义识别和所述需求文档得到所述原始数据中的待清洗字段。

4.如权利要求2所述的数据清洗方法,其特征在于,所述提取所述待清洗字段的关键词,并基于所述关键词确定所述原始数据对应的目标数据类型的步骤,具体包括:

对所述待清洗字段进行关键词识别,获取所有待清洗字段的关键词;

整合提取到的关键词,生成所述原始数据的关键词组合;

将所述关键词组合表示的类型确定所述原始数据对应的目标业务数据类型。

5.如权利要求4所述的数据清洗方法,其特征在于,所述整合提取到的关键词,生成所述原始数据的关键词组合的步骤,具体包括:

基于预设的TF-IDF算法计算每一个所述关键词的权重;

对所有关键词的权重进行排序,得到关键词权重序列;

基于所述关键词权重序列组合所述关键词,生成所述原始数据的关键词组合。

6.如权利要求1所述的数据清洗方法,其特征在于,所述通过所述目标子服务器对所述原始数据进行数据清洗,得到清洗数据的步骤,具体包括:

对所述原始数据进行格式化处理,得到格式化数据;

检测所述格式化数据中的重复数据,并对所述重复数据进行清洗,得到去重数据;

检测所述去重数据中的错误数据,并对所述错误数据进行清洗,得到清洗数据。

7.一种大数据清洗装置,其特征在于,包括:

规则配置模块,用于创建预设数量的清洗规则,并配置业务数据类型与所述清洗规则之间的映射关系,生成清洗规则匹配表;

集群构建模块,用于构建分布式集群,其中,所述分布式集群包含若干个子服务器,各个所述子服务器分别对应处理一种业务类型数据;

规则分配模块,用于将所述清洗规则匹配表上传至所述分布式集群,并根据所述业务数据类型将所述清洗规则分配给对应的子服务器;

数据预处理模块,用于接收原始数据,并确定所述原始数据对应的目标业务数据类型;

数据清洗模块,用于在所述分布式集群中查找与所述目标业务数据类型对应的目标子服务器,并利用所述目标子服务器对所述原始数据进行数据清洗,得到清洗数据;

所述数据清洗模块具体包括:

业务标签获取子模块,用于获取每一个子服务器对应的业务数据标签;

标签比对子模块,用于在所述分布式集群中,将所述目标业务数据类型与每一个子服务器对应的业务数据标签一一进行比对;

数据清洗子模块,用于根据比对结果确定所述原始数据对应的目标子服务器,并通过所述目标子服务器对所述原始数据进行数据清洗,得到清洗数据。

8.一种计算机设备,其特征在于,包括存储器和处理器,所述存储器中存储有计算机可读指令,所述处理器执行所述计算机可读指令时实现如权利要求1至6中任一项所述的大数据清洗方法的步骤。

9.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质上存储有计算机可读指令,所述计算机可读指令被处理器执行时实现如权利要求1至6中任一项所述的大数据清洗方法的步骤。

一种大数据清洗方法、装置、计算机设备及存储介质

技术领域

[0001] 本申请属于大数据技术领域,具体涉及一种大数据清洗方法、装置、计算机设备及存储介质。

背景技术

[0002] 数据清洗是指发现并纠正数据文件中可识别的错误的最后一道程序,包括检查数据一致性,处理无效值和缺失值等。与问卷审核不同,录入后的数据清理一般是由计算机而不是人工完成。

[0003] 目前,针对数据化清洗场景,不同的业务部门的不同或不同的业务场景所产生数据的数据量、数据有效周期或者数据管理规则可能完全不同,现有的数据清洗方案针对不同业务需求通常需要单独开发对应的数据清理规则,但上述清洗方案会在数据清理规则开发阶段耗费较大的人力和物理,且对于一些共用的数据清理规则无法实现复用,导致了开发资源浪费,同时也不利于数据清理规则的管理。

发明内容

[0004] 本申请实施例的目的在于提出一种大数据清洗方法、装置、计算机设备及存储介质,以解决现有的大数据清洗方案中存在的某些共用的数据清理规则无法实现复用,导致的数据清理规则开发资源浪费,且数据清理规则不易管理的技术问题。

[0005] 为了解决上述技术问题,本申请实施例提供一种大数据清洗方法,采用了如下所述的技术方案:

[0006] 一种大数据清洗方法,包括:

[0007] 创建预设数量的清洗规则,并配置业务数据类型与所述清洗规则之间的映射关系,生成清洗规则匹配表;

[0008] 构建分布式集群,其中,所述分布式集群包含若干个子服务器,各个所述子服务器分别对应处理一种业务类型数据;

[0009] 将所述清洗规则匹配表上传至所述分布式集群,并根据所述业务数据类型将所述清洗规则分配给对应的子服务器;

[0010] 接收原始数据,并确定所述原始数据对应的目标业务数据类型;

[0011] 在所述分布式集群中查找与所述目标业务数据类型对应的目标子服务器,并利用所述目标子服务器对所述原始数据进行数据清洗,得到清洗数据。

[0012] 进一步地,所述接收原始数据,并确定所述原始数据对应的目标业务数据类型的步骤,具体包括:

[0013] 接收原始数据,将所述原始数据导入所述分布式集群,并确定所述原始数据中的待清洗字段;

[0014] 提取所述待清洗字段的关键词,并基于所述关键词确定所述原始数据对应的目标数据类型。

[0015] 进一步地,所述接收原始数据,将所述原始数据导入所述分布式集群,并确定所述原始数据中的待清洗字段的步骤,具体包括:

[0016] 获取所述原始数据对应的需求文档,其中,所述需求文档中记录了数据清洗的具体要求;

[0017] 识别所述原始数据的数据结构,得到所述原始数据的结构信息;

[0018] 基于所述结构信息对原始数据进行分割,得到若干个数据字段;

[0019] 对每一个所述数据字段进行语义识别,并基于语义识别和所述需求文档得到所述原始数据中的待清洗字段。

[0020] 进一步地,所述提取所述待清洗字段的关键词,并基于所述关键词确定所述原始数据对应的目标数据类型的步骤,具体包括:

[0021] 对所述待清洗字段进行关键词识别,获取所有待清洗字段的关键词;

[0022] 整合提取到的关键词,生成所述原始数据的关键词组合;

[0023] 将所述关键词组合表示的类型确定所述原始数据对应的目标业务数据类型。

[0024] 进一步地,所述整合提取到的关键词,生成所述原始数据的关键词组合的步骤,具体包括:

[0025] 基于预设的TF-IDF算法计算每一个所述关键词的权重;

[0026] 对所有关键词的权重进行排序,得到关键词权重序列;

[0027] 基于所述关键词权重序列组合所述关键词,生成所述原始数据的关键词组合。

[0028] 进一步地,所述在所述分布式集群中查找与所述目标业务数据类型对应的目标子服务器,并利用所述目标子服务器对所述原始数据进行数据清洗,得到清洗数据的步骤,具体包括:

[0029] 获取每一个子服务器对应的业务数据标签;

[0030] 在所述分布式集群中,将所述目标业务数据类型与每一个子服务器对应的业务数据标签一一进行比对;

[0031] 根据比对结果确定所述原始数据对应的目标子服务器,并通过所述目标子服务器对所述原始数据进行数据清洗,得到清洗数据。

[0032] 进一步地,所述通过所述目标子服务器对所述原始数据进行数据清洗,得到清洗数据的步骤,具体包括:

[0033] 对所述原始数据进行格式化处理,得到格式化数据;

[0034] 检测所述格式化数据中的重复数据,并对所述重复数据进行清洗,得到去重数据;

[0035] 检测所述去重数据中的错误数据,并对所述错误数据进行清洗,得到清洗数据。

[0036] 为了解决上述技术问题,本申请实施例还提供一种大数据清洗装置,采用了如下所述的技术方案:

[0037] 一种大数据清洗装置,包括:

[0038] 规则配置模块,用于创建预设数量的清洗规则,并配置业务数据类型与所述清洗规则之间的映射关系,生成清洗规则匹配表;

[0039] 集群构建模块,用于构建分布式集群,其中,所述分布式集群包含若干个子服务器,各个所述子服务器分别对应处理一种业务类型数据;

[0040] 规则分配模块,用于将所述清洗规则匹配表上传至所述分布式集群,并根据所述

业务数据类型将所述清洗规则分配给对应的子服务器；

[0041] 数据预处理模块，用于接收原始数据，并确定所述原始数据对应的目标业务数据类型；

[0042] 数据清洗模块，用于在所述分布式集群中查找与所述目标业务数据类型对应的目标子服务器，并利用所述目标子服务器对所述原始数据进行数据清洗，得到清洗数据。

[0043] 为了解决上述技术问题，本申请实施例还提供一种计算机设备，采用了如下所述的技术方案：

[0044] 一种计算机设备，包括存储器和处理器，所述存储器中存储有计算机可读指令，所述处理器执行所述计算机可读指令时实现如上述所述的大数据清洗方法的步骤。

[0045] 为了解决上述技术问题，本申请实施例还提供一种计算机可读存储介质，采用了如下所述的技术方案：

[0046] 一种计算机可读存储介质，所述计算机可读存储介质上存储有计算机可读指令，所述计算机可读指令被处理器执行时实现如上述所述的大数据清洗方法的步骤。

[0047] 与现有技术相比，本申请实施例主要有以下有益效果：

[0048] 本申请公开了一种大数据清洗方法、装置、计算机设备及存储介质，属于大数据技术领域。本申请通过构建分布式集群，并根据业务数据类型为分布式集群中的各个子服务器配置相应的数据清洗规则，其中，每一个子服务器配置一种业务数据类型对应的数据清洗规则。当需要进行业务数据清洗时，服务器通过识别待清洗业务数据的数据类型，并根据待清洗业务数据的数据类型将待清洗业务数据分配给分布式集群中对应的子服务器进行处理。本申请通过构建分布式集群以实现不同类型业务数据自动清理，具有较强的通用性和适应性，能够有效降低数据清理规则开发阶段的耗费，提高公用数据清理规则的复用率，同时有利于数据清理规则的统一管理。

附图说明

[0049] 为了更清楚地说明本申请中的方案，下面将对本申请实施例描述中所需要使用的附图作一个简单介绍，显而易见地，下面描述中的附图是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0050] 图1示出了本申请可以应用于其中的示例性系统架构图；

[0051] 图2示出了根据本申请的大数据清洗方法的一个实施例的流程图；

[0052] 图3示出了根据本申请的大数据清洗装置的一个实施例的结构示意图；

[0053] 图4示出了根据本申请的计算机设备的一个实施例的结构示意图。

具体实施方式

[0054] 除非另有定义，本文所使用的所有的技术和科学术语与属于本申请的技术领域的技术人员通常理解的含义相同；本文中在申请的说明书中所使用的术语只是为了描述具体的实施例的目的，不是旨在于限制本申请；本申请的说明书和权利要求书及上述附图说明中的术语“包括”和“具有”以及它们的任何变形，意图在于覆盖不排他的包含。本申请的说明书和权利要求书或上述附图中的术语“第一”、“第二”等是用于区别不同对象，而不是用于描述特定顺序。

[0055] 在本文中提及“实施例”意味着,结合实施例描述的特定特征、结构或特性可以包含在本申请的至少一个实施例中。在说明书中的各个位置出现该短语并不一定均是指相同的实施例,也不是与其它实施例互斥的独立的或备选的实施例。本领域技术人员显式地和隐式地理解的是,本文所描述的实施例可以与其它实施例相结合。

[0056] 为了使本技术领域的人员更好地理解本申请方案,下面将结合附图,对本申请实施例中的技术方案进行清楚、完整地描述。

[0057] 如图1所示,系统架构100可以包括终端设备101、102、103,网络104和服务器105。网络104用以在终端设备101、102、103和服务器105之间提供通信链路的介质。网络104可以包括各种连接类型,例如有线、无线通信链路或者光纤电缆等等。

[0058] 用户可以使用终端设备101、102、103通过网络104与服务器105交互,以接收或发送消息等。终端设备101、102、103上可以安装有各种通讯客户端应用,例如网页浏览器应用、购物类应用、搜索类应用、即时通信工具、邮箱客户端、社交平台软件等。

[0059] 终端设备101、102、103可以是具有显示屏并且支持网页浏览的各种电子设备,包括但不限于智能手机、平板电脑、电子书阅读器、MP3播放器(Moving Picture ExpertsGroup Audio Layer III,动态影像专家压缩标准音频层面3)、MP4(Moving PictureExperts Group Audio Layer IV,动态影像专家压缩标准音频层面4)播放器、膝上型便携计算机和台式计算机等等。

[0060] 服务器105可以是提供各种服务的服务器,例如对终端设备101、102、103上显示的页面提供支持的后台服务器,服务器可以是独立的服务器,也可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、内容分发网络(Content Delivery Network,CDN)、以及大数据和人工智能平台等基础云计算服务的云服务器。

[0061] 需要说明的是,本申请实施例所提供的大数据清洗方法一般由服务器执行,相应地,大数据清洗装置一般设置于服务器中。

[0062] 应该理解,图1中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要,可以具有任意数目的终端设备、网络和服务器。

[0063] 继续参考图2,示出了根据本申请的大数据清洗方法的一个实施例的流程图。本申请实施例可以基于人工智能技术对相关的数据进行获取和处理。其中,人工智能(Artificial Intelligence,AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能,感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。

[0064] 人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、机器人技术、生物识别技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。所述的大数据清洗方法,包括以下步骤:

[0065] S201,创建预设数量的清洗规则,并配置业务数据类型与所述清洗规则之间的映射关系,生成清洗规则匹配表。

[0066] 具体的,本申请在构建分布式集群之前,先通过服务器创建预设数量的清洗规则,其中,清洗规则包括格式化规则、去重规则、纠正规则等等。然后根据业务场景需求定义业

务数据类型与清洗规则之间的映射关系,例如,对文本数据的清洗至少需要进行分词、格式化、去重等操作,因此针对文本数据至少需要配置相应的分词规则、格式化规则和去重规则,最后整合所有业务数据类型与清洗规则之间的映射关系,生成清洗规则匹配表。

[0067] 在本实施例中,通过创建清洗规则,以及配置业务数据类型与清洗规则之间的映射关系,生成清洗规则匹配表,以便于数据清理规则的统一管理。

[0068] S202,构建分布式集群,其中,所述分布式集群包含若干个子服务器,各个所述子服务器分别对应处理一种业务类型数据。

[0069] 其中,本申请中的分布式集群可以基于Spark集群架构搭建,Spark集群架构包括开源分布式存储系统Tachyon、开源分布式资源管理框架Mesos、资源管理器YARN、大规模并行查询引擎BlinkDB等,其中Tachyon是基于内存的分布式文件系统,便于各任务共享数据且降低计算过程中JVM的负载;Mesos是一个集群管理器,使用程序协调服务Zookeeper实现集群容错性;BlinkDB是一个大规模并行查询引擎,允许通过权衡数据精度来提升查询响应时间。

[0070] 具体的,服务器基于Spark集群架构搭建用于实现数据清洗的分布式集群,其中,分布式集群包含若干个子服务器,各个子服务器分别对应处理一种业务类型数据,本申请通过构建分布式集群不仅能实现不同类型业务数据自动清理,具有较强的通用性和适应性,且使用分布式集群能够响应大规模数据的即时处理需求。

[0071] S203,将所述清洗规则匹配表上传至所述分布式集群,并根据所述业务数据类型将所述清洗规则分配给对应的子服务器。

[0072] 具体的,服务器在完成分布式集群的基础搭建之后,将清洗规则匹配表上传至分布式集群,并按照业务数据类型将清洗规则匹配表上对应的清洗规则分配给对应的子服务器。例如,在本申请一种具体的实施例中,将文本数据对应的清洗规则分配给子服务器A,将数值数据对应的清洗规则分配给子服务器B。在本申请一种更具体的实施例中,业务数据为保单业务数据,将保单业务数据中的文本数据对应的清洗规则分配给子服务器A1,将保单业务数据中的数值数据对应的清洗规则分配给子服务器B1。

[0073] 在本实施例中,通过构建分布式集群,并根据业务数据类型为分布式集群中的各个子服务器配置相应的数据清洗规则,以实现不同类型业务数据清理,具有较强的通用性和适应性。

[0074] S204,接收原始数据,并确定所述原始数据对应的目标业务数据类型。

[0075] 具体的,当存在数据清洗需求时,服务器接收数据清洗指令,并接收客户端上传的原始数据和需求文档,其中,需求文档中记录了数据清洗的具体要求。服务器将客户端上传的原始数据导入分布式集群,并预先确定原始数据对应的目标数据类型,以及在分布式集群中查找与业务数据类型对应的子服务器,最后将原始数据导入该子服务器进行数据清洗。

[0076] 在本实施例中,大数据清洗方法运行于其上的电子设备(例如图1所示的服务器)可以通过有线连接方式或者无线连接方式接收数据清洗指令。需要指出的是,上述无线连接方式可以包括但不限于3G/4G连接、WiFi连接、蓝牙连接、WiMAX连接、Zigbee连接、UWB(ultra wideband)连接、以及其他现在已知或将来开发的无线连接方式。

[0077] S205,在所述分布式集群中查找与所述目标业务数据类型对应的目标子服务器,

并利用所述目标子服务器对所述原始数据进行数据清洗,得到清洗数据。

[0078] 具体的,服务器在根据业务数据类型将清洗规则分配给对应的子服务器之后,会为每一个子服务器生成对应的业务数据标签,例如,在上述实施例中,子服务器A1对应的业务数据标签为“保单业务数据-文本数据”,子服务器B1对应的业务数据标签为“保单业务数据-数值数据”。在进行业务数据清洗时,服务器在确定原始数据对应的目标业务数据类型后,将目标业务数据类型与每一个子服务器对应的业务数据标签一一进行比对,当目标业务数据类型与其中一个子服务器对应的业务数据标签匹配时,将该子服务器作为目标子服务器,并将原始数据输入到目标子服务器进行进行数据清洗,得到清洗数据。

[0079] 在上述实施例中,本申请通过构建分布式集群,并根据业务数据类型为分布式集群中的各个子服务器配置相应的数据清洗规则,其中,每一个子服务器配置一种业务数据类型对应的数据清洗规则。当需要进行业务数据清洗时,服务器通过识别待清洗业务数据的数据类型,并根据待清洗业务数据的数据类型将待清洗业务数据分配给分布式集群中对应的子服务器进行处理。本申请通过构建分布式集群以实现不同类型业务数据自动清理,具有较强的通用性和适应性,能够有效降低数据清理规则开发阶段的耗费,提高公用数据清理规则的复用率,同时有利于数据清理规则的统一管理。

[0080] 进一步地,所述接收原始数据,并确定所述原始数据对应的目标业务数据类型的步骤,具体包括:

[0081] 接收原始数据,将所述原始数据导入所述分布式集群,并确定所述原始数据中的待清洗字段;

[0082] 提取所述待清洗字段的关键词,并基于所述关键词确定所述原始数据对应的目标数据类型。

[0083] 具体的,服务器在接收原始数据后,将原始数据导入分布式集群,在分布式集群中先对原始数据进行字段分割,在分割出了的字段中根据需求文档确定原始数据中的待清洗字段,通过提取待清洗字段的关键词,并对提取的关键词进行语义分析,以确定原始数据对应的目标数据类型。

[0084] 进一步地,所述接收原始数据,将所述原始数据导入所述分布式集群,并确定所述原始数据中的待清洗字段的步骤,具体包括:

[0085] 获取所述原始数据对应的需求文档,其中,所述需求文档中记录了数据清洗的具体要求;

[0086] 识别所述原始数据的数据结构,得到所述原始数据的结构信息;

[0087] 基于所述结构信息对原始数据进行分割,得到若干个数据字段;

[0088] 对每一个所述数据字段进行语义识别,并基于语义识别和所述需求文档得到所述原始数据中的待清洗字段。

[0089] 具体的,服务器识别原始数据的数据结构,得到原始数据的结构信息,基于结构信息对原始数据进行字段分割,得到多个数据字段,例如,一种具体的原始数据的数据结构为多段落结构,根据多段落结构分布对原始数据进行字段分割,得到多个数据字段,其中,对原始数据进行字段分割后得到的数据字段包括待清洗字段和无需清洗的数据字段。最后对每一个数据字段进行语义识别,并基于语义识别结果和需求文档确定原始数据中的待清洗字段。

[0090] 在上述实施例中,本申请通过获取原始数据的结构信息,并依据原始数据的结构信息对原始数据进行分割,将原始数据分割成多个标准数据字段,以便于后续进行语义识别获取原始数据中的待清洗字段。

[0091] 进一步地,所述提取所述待清洗字段的关键词,并基于所述关键词确定所述原始数据对应的目标数据类型的步骤,具体包括:

[0092] 对所述待清洗字段进行关键词识别,获取所有待清洗字段的关键词;

[0093] 整合提取到的关键词,生成所述原始数据的关键词组合;

[0094] 将所述关键词组合表示的类型确定所述原始数据对应的目标业务数据类型。

[0095] 具体的,服务器对待清洗字段进行关键词识别,获取所有待清洗字段的关键词,其中,关键词识别可以采用OCR字段扫描实现。服务器在完成关键词提取后,并计算每一个关键词的权重,基于计算的权重整合关键词,生成原始数据的关键词组合,基于关键词组合确定原始数据对应的目标业务数据类型。

[0096] 在上述实施例中,本申请通过用OCR字段扫描获取所有待清洗字段的关键词,并通过计算关键词的权重,以及根据原始数据中关键词的权重来确定原始数据对应的目标业务数据类型。

[0097] 进一步地,所述整合提取到的关键词,生成所述原始数据的关键词组合的步骤,具体包括:

[0098] 基于预设的TF-IDF算法计算每一个所述关键词的权重;

[0099] 对所有关键词的权重进行排序,得到关键词权重序列;

[0100] 基于所述关键词权重序列组合所述关键词,生成所述原始数据的关键词组合。

[0101] 具体的,服务器基于预设的TF-IDF算法计算每一个关键词的权重,对所有关键词的权重进行降序排列,得到关键词权重序列,根据需求文档的要求在关键词权重序列中选择排名靠前的关键词,组合排名靠前的关键词,得到原始数据的关键词组合。

[0102] 其中,TF-IDF(term frequency-inverse document frequency)是一种用于资讯检索与资讯探勘的常用加权技术。TF-IDF是一种统计方法,用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加,但同时会随着它在语料库中出现的频率成反比下降。TF-IDF加权的各种形式常被搜寻引擎应用,作为文件与用户查询之间相关程度的度量或评级。除了TF-IDF以外,因特网上的搜寻引擎还会使用基于连结分析的评级方法,以确定文件在搜寻结果中出现的顺序。

[0103] 其中,基于预设的TF-IDF算法计算每一个关键词的权重,具体包括:

[0104] 计算关键词的词频,以及计算关键词的逆文档频率;

[0105] 基于关键词的词频和关键词的逆文档频率,计算关键词的第一分词权重。

[0106] 其中,计算关键词的词频,以及计算关键词的逆文档频率,具体包括:

[0107] 确定关键词所在的待清洗字段,得到目标字段;

[0108] 统计关键词在目标字段中的出现次数,得到第一分词数,以及统计所有关键词在各个待清洗字段中的出现次数总和,得到第二分词数;

[0109] 基于第一分词数和第二分词数,计算关键词的词频;

[0110] 统计目标字段的数量,得到第一文档数量,以及统计待清洗字段的总数,得到第二

文档数量；

[0111] 基于第一文档数量和第二文档数量，计算关键词的逆文档频率。

[0112] 具体的，词频TF的计算公式如下：

[0113] $$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}}$$

[0114] 其中，$tf_{i,j}$为关键词t_i的词频，$n_{i,j}$为关键词t_i在某个待清洗字段d_j中的出现次数，$\sum_k n_{k,j}$为k个关键词在所有待清洗字段中的出现次数总和。

[0115] 逆文本频率IDF的计算公式如下：

[0116] $$idf_{i,j}=\log\frac{|D|}{|\{j:t_i\in d_j\}|}$$

[0117] 其中，$idf_{i,j}$为关键词t_i的逆文本频率指数，|D|为待清洗字段的总数，$|\{j:t_i\in d_j\}|$包含关键词t_i的待清洗字段的数量。

[0118] 在本实施例中，通过计算关键词的权重，以及对关键词的权重进行降序排列，并选择排序结果中排名靠前的关键词进行组合，得到关键词组合，通过关键词权重计算和排序，将原始数据中比较重要的关键词组合在一起，以便更精准地确定原始数据对应的目标业务数据类型。

[0119] 进一步地，所述在所述分布式集群中查找与所述目标业务数据类型对应的目标子服务器，并利用所述目标子服务器对所述原始数据进行数据清洗，得到清洗数据的步骤，具体包括：

[0120] 获取每一个子服务器对应的业务数据标签；

[0121] 在所述分布式集群中，将所述目标业务数据类型与每一个子服务器对应的业务数据标签一一进行比对；

[0122] 根据比对结果确定所述原始数据对应的目标子服务器，并通过所述目标子服务器对所述原始数据进行数据清洗，得到清洗数据。

[0123] 具体的，服务器在根据业务数据类型将清洗规则分配给对应的子服务器之后，会为每一个子服务器生成对应的业务数据标签，在进行业务数据清洗时，服务器在确定原始数据对应的目标业务数据类型，将目标业务数据类型与每一个子服务器对应的业务数据标签一一进行比对，当目标业务数据类型与其中一个子服务器对应的业务数据标签匹配时，将该子服务器作为目标子服务器，并将原始数据输入到目标子服务器进行进行数据清洗，得到清洗数据。

[0124] 在本实施例中，在进行业务数据清洗时，服务器先将目标业务数据类型与每一个子服务器对应的业务数据标签一一进行比对，并根据比对结果将原始数据输入到对应的子服务器进行进行数据清洗。

[0125] 进一步地，所述通过所述目标子服务器对所述原始数据进行数据清洗，得到清洗数据的步骤，具体包括：

[0126] 对所述原始数据进行格式化处理，得到格式化数据；

[0127] 检测所述格式化数据中的重复数据，并对所述重复数据进行清洗，得到去重数据；

[0128] 检测所述去重数据中的错误数据,并对所述错误数据进行清洗,得到清洗数据。

[0129] 具体的,服务器通过调用存储在目标子服务器中的数据格式化规则对原始数据进行格式化处理,使得原始数据统一规整为符合要求的规范数据,然后调用数据去重规则对原始数据中的重复数据进行去重处理,去除多余重复数据,最后检索去重数据中存在的错误数据,并调用数据纠正规则对错误数据进行错误内容清洗,生成最终的清洗数据。

[0130] 在本实施例中,服务器根据需求文档上的清洗要求选择对应的数据清洗规则,并通过数据清洗规则对原始数据进行清洗,得到清洗数据。

[0131] 本申请公开了一种大数据清洗方法,属于大数据技术领域。本申请通过构建分布式集群,并根据业务数据类型为分布式集群中的各个子服务器配置相应的数据清洗规则,其中,每一个子服务器配置一种业务数据类型对应的数据清洗规则。当需要进行业务数据清洗时,服务器通过识别待清洗业务数据的数据类型,并根据待清洗业务数据的数据类型将待清洗业务数据分配给分布式集群中对应的子服务器进行处理。本申请通过构建分布式集群以实现不同类型业务数据自动清理,具有较强的通用性和适应性,能够有效降低数据清理规则开发阶段的耗费,提高公用数据清理规则的复用率,同时有利于数据清理规则的统一管理。

[0132] 需要强调的是,为进一步保证上述原始数据的私密和安全性,上述原始数据还可以存储于一区块链的节点中。

[0133] 本申请所指区块链是分布式数据存储、点对点传输、共识机制、加密算法等计算机技术的新型应用模式。区块链(Blockchain),本质上是一个去中心化的数据库,是一串使用密码学方法相关联产生的数据块,每一个数据块中包含了一批次网络交易的信息,用于验证其信息的有效性(防伪)和生成下一个区块。区块链可以包括区块链底层平台、平台产品服务层以及应用服务层等。

[0134] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程,是可以通过计算机可读指令来指令相关的硬件来完成,该计算机可读指令可存储于一计算机可读取存储介质中,该计算机可读指令在执行时,可包括如上述各方法的实施例的流程。其中,前述的存储介质可为磁碟、光盘、只读存储记忆体(Read-Only Memory,ROM)等非易失性存储介质,或随机存储记忆体(Random Access Memory,RAM)等。

[0135] 应该理解的是,虽然附图的流程图中的各个步骤按照箭头的指示依次显示,但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明,这些步骤的执行并没有严格的顺序限制,其可以以其他的顺序执行。而且,附图的流程图中的至少一部分步骤可以包括多个子步骤或者多个阶段,这些子步骤或者阶段并不必然是在同一时刻执行完成,而是可以在不同的时刻执行,其执行顺序也不必然是依次进行,而是可以与其他步骤或者其他步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

[0136] 进一步参考图3,作为对上述图2所示方法的实现,本申请提供了一种大数据清洗装置的一个实施例,该装置实施例与图2所示的方法实施例相对应,该装置具体可以应用于各种电子设备中。

[0137] 如图3所示,本实施例所述的大数据清洗装置包括:

[0138] 规则配置模块301,用于创建预设数量的清洗规则,并配置业务数据类型与所述清洗规则之间的映射关系,生成清洗规则匹配表;

[0139] 集群构建模块302,用于构建分布式集群,其中,所述分布式集群包含若干个子服务器,各个所述子服务器分别对应处理一种业务类型数据;

[0140] 规则分配模块303,用于将所述清洗规则匹配表上传至所述分布式集群,并根据所述业务数据类型将所述清洗规则分配给对应的子服务器;

[0141] 数据预处理模块304,用于接收原始数据,并确定所述原始数据对应的目标业务数据类型;

[0142] 数据清洗模块305,用于在所述分布式集群中查找与所述目标业务数据类型对应的目标子服务器,并利用所述目标子服务器对所述原始数据进行数据清洗,得到清洗数据。

[0143] 进一步地,所述数据预处理模块304具体包括:

[0144] 字段识别子模块,用于接收原始数据,将所述原始数据导入所述分布式集群,并确定所述原始数据中的待清洗字段;

[0145] 关键词识别子模块,用于提取所述待清洗字段的关键词,并基于所述关键词确定所述原始数据对应的目标数据类型。

[0146] 进一步地,所述字段识别子模块具体包括:

[0147] 需求文档获取单元,用于获取所述原始数据对应的需求文档,其中,所述需求文档中记录了数据清洗的具体要求;

[0148] 结构识别单元,用于识别所述原始数据的数据结构,得到所述原始数据的结构信息;

[0149] 字段分割单元,用于基于所述结构信息对原始数据进行分割,得到若干个数据字段;

[0150] 语义识别单元,用于对每一个所述数据字段进行语义识别,并基于语义识别和所述需求文档得到所述原始数据中的待清洗字段。

[0151] 进一步地,所述关键词识别子模块具体包括:

[0152] 关键词识别单元,用于对所述待清洗字段进行关键词识别,获取所有待清洗字段的关键词;

[0153] 关键词组合单元,用于整合提取到的关键词,生成所述原始数据的关键词组合;

[0154] 业务类型判断单元,用于将所述关键词组合表示的类型确定所述原始数据对应的目标业务数据类型。

[0155] 进一步地,所述关键词组合单元具体包括:

[0156] 权重计算子单元,用于基于预设的TF-IDF算法计算每一个所述关键词的权重;

[0157] 权重排序子单元,用于对所有关键词的权重进行排序,得到关键词权重序列;

[0158] 关键词组合子单元,用于基于所述关键词权重序列组合所述关键词,生成所述原始数据的关键词组合。

[0159] 进一步地,所述数据清洗模块305具体包括:

[0160] 业务标签获取子模块,用于获取每一个子服务器对应的业务数据标签;

[0161] 标签比对子模块,用于在所述分布式集群中,将所述目标业务数据类型与每一个子服务器对应的业务数据标签一一进行比对;

[0162] 数据清洗子模块,用于根据比对结果确定所述原始数据对应的目标子服务器,并通过所述目标子服务器对所述原始数据进行数据清洗,得到清洗数据。

[0163] 进一步地,所述数据清洗子模块具体包括:

[0164] 格式化单元,用于对所述原始数据进行格式化处理,得到格式化数据;

[0165] 第一清洗单元,用于检测所述格式化数据中的重复数据,并对所述重复数据进行清洗,得到去重数据;

[0166] 第二清洗单元,用于检测所述去重数据中的错误数据,并对所述错误数据进行清洗,得到清洗数据。

[0167] 本申请公开了一种大数据清洗装置,属于大数据技术领域。本申请通过构建分布式集群,并根据业务数据类型为分布式集群中的各个子服务器配置相应的数据清洗规则,其中,每一个子服务器配置一种业务数据类型对应的数据清洗规则。当需要进行业务数据清洗时,服务器通过识别待清洗业务数据的数据类型,并根据待清洗业务数据的数据类型将待清洗业务数据分配给分布式集群中对应的子服务器进行处理。本申请通过构建分布式集群以实现不同类型业务数据自动清理,具有较强的通用性和适应性,能够有效降低数据清理规则开发阶段的耗费,提高公用数据清理规则的复用率,同时有利于数据清理规则的统一管理。

[0168] 为解决上述技术问题,本申请实施例还提供计算机设备。具体请参阅图4,图4为本实施例计算机设备基本结构框图。

[0169] 所述计算机设备4包括通过系统总线相互通信连接存储器41、处理器42、网络接口43。需要指出的是,图中仅示出了具有组件41-43的计算机设备4,但是应理解的是,并不要求实施所有示出的组件,可以替代的实施更多或者更少的组件。其中,本技术领域技术人员可以理解,这里的计算机设备是一种能够按照事先设定或存储的指令,自动进行数值计算和/或信息处理的设备,其硬件包括但不限于微处理器、专用集成电路(Application Specific Integrated Circuit,ASIC)、可编程门阵列(Field—Programmable Gate Array,FPGA)、数字处理器(Digital Signal Processor,DSP)、嵌入式设备等。

[0170] 所述计算机设备可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。所述计算机设备可以与用户通过键盘、鼠标、遥控器、触摸板或声控设备等方式进行人机交互。

[0171] 所述存储器41至少包括一种类型的可读存储介质,所述可读存储介质包括闪存、硬盘、多媒体卡、卡型存储器(例如,SD或DX存储器等)、随机访问存储器(RAM)、静态随机访问存储器(SRAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、可编程只读存储器(PROM)、磁性存储器、磁盘、光盘等。在一些实施例中,所述存储器41可以是所述计算机设备4的内部存储单元,例如该计算机设备4的硬盘或内存。在另一些实施例中,所述存储器41也可以是所述计算机设备4的外部存储设备,例如该计算机设备4上配备的插接式硬盘,智能存储卡(Smart Media Card,SMC),安全数字(Secure Digital,SD)卡,闪存卡(Flash Card)等。当然,所述存储器41还可以既包括所述计算机设备4的内部存储单元也包括其外部存储设备。本实施例中,所述存储器41通常用于存储安装于所述计算机设备4的操作系统和各类应用软件,例如大数据清洗方法的计算机可读指令等。此外,所述存储器41还可以用于暂时地存储已经输出或者将要输出的各类数据。

[0172] 所述处理器42在一些实施例中可以是中央处理器(Central Processing Unit,CPU)、控制器、微控制器、微处理器、或其他数据处理芯片。该处理器42通常用于控制所述计

算机设备4的总体操作。本实施例中,所述处理器42用于运行所述存储器41中存储的计算机可读指令或者处理数据,例如运行所述大数据清洗方法的计算机可读指令。

[0173] 所述网络接口43可包括无线网络接口或有线网络接口,该网络接口43通常用于在所述计算机设备4与其他电子设备之间建立通信连接。

[0174] 本申请公开了一种计算机设备,属于大数据技术领域。本申请通过构建分布式集群,并根据业务数据类型为分布式集群中的各个子服务器配置相应的数据清洗规则,其中,每一个子服务器配置一种业务数据类型对应的数据清洗规则。当需要进行业务数据清洗时,服务器通过识别待清洗业务数据的数据类型,并根据待清洗业务数据的数据类型将待清洗业务数据分配给分布式集群中对应的子服务器进行处理。本申请通过构建分布式集群以实现不同类型业务数据自动清理,具有较强的通用性和适应性,能够有效降低数据清理规则开发阶段的耗费,提高公用数据清理规则的复用率,同时有利于数据清理规则的统一管理。

[0175] 本申请还提供了另一种实施方式,即提供一种计算机可读存储介质,所述计算机可读存储介质存储有计算机可读指令,所述计算机可读指令可被至少一个处理器执行,以使所述至少一个处理器执行如上述的大数据清洗方法的步骤。

[0176] 本申请公开了一种存储介质,属于大数据技术领域。本申请通过构建分布式集群,并根据业务数据类型为分布式集群中的各个子服务器配置相应的数据清洗规则,其中,每一个子服务器配置一种业务数据类型对应的数据清洗规则。当需要进行业务数据清洗时,服务器通过识别待清洗业务数据的数据类型,并根据待清洗业务数据的数据类型将待清洗业务数据分配给分布式集群中对应的子服务器进行处理。本申请通过构建分布式集群以实现不同类型业务数据自动清理,具有较强的通用性和适应性,能够有效降低数据清理规则开发阶段的耗费,提高公用数据清理规则的复用率,同时有利于数据清理规则的统一管理。

[0177] 通过以上的实施方式的描述,本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现,当然也可以通过硬件,但很多情况下前者是更佳的实施方式。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(如ROM/RAM、磁碟、光盘)中,包括若干指令用以使得一台终端设备(可以是手机,计算机,服务器,空调器,或者网络设备等)执行本申请各个实施例所述的方法。

[0178] 本申请可用于众多通用或专用的计算机系统环境或配置中。例如:个人计算机、服务器计算机、手持设备或便携式设备、平板型设备、多处理器系统、基于微处理器的系统、置顶盒、可编程的消费电子设备、网络PC、小型计算机、大型计算机、包括以上任何系统或设备的分布式计算环境等等。本申请可以在由计算机执行的计算机可执行指令的一般上下文中描述,例如程序模块。一般地,程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等等。也可以在分布式计算环境中实践本申请,在这些分布式计算环境中,由通过通信网络而被连接的远程处理设备来执行任务。在分布式计算环境中,程序模块可以位于包括存储设备在内的本地和远程计算机存储介质中。

[0179] 显然,以上所描述的实施例仅仅是本申请一部分实施例,而不是全部的实施例,附图中给出了本申请的较佳实施例,但并不限制本申请的专利范围。本申请可以以许多不同的形式来实现,相反地,提供这些实施例的目的是使对本申请的公开内容的理解更加透彻

全面。尽管参照前述实施例对本申请进行了详细的说明,对于本领域的技术人员来而言,其依然可以对前述各具体实施方式所记载的技术方案进行修改,或者对其中部分技术特征进行等效替换。凡是利用本申请说明书及附图内容所做的等效结构,直接或间接运用在其他相关的技术领域,均同理在本申请专利保护范围之内。

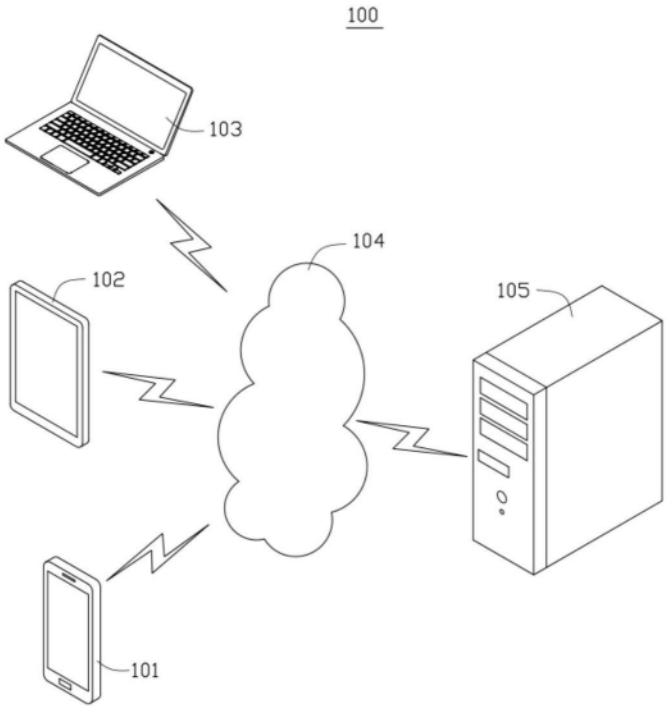

100
103
104
102
105
101

图1

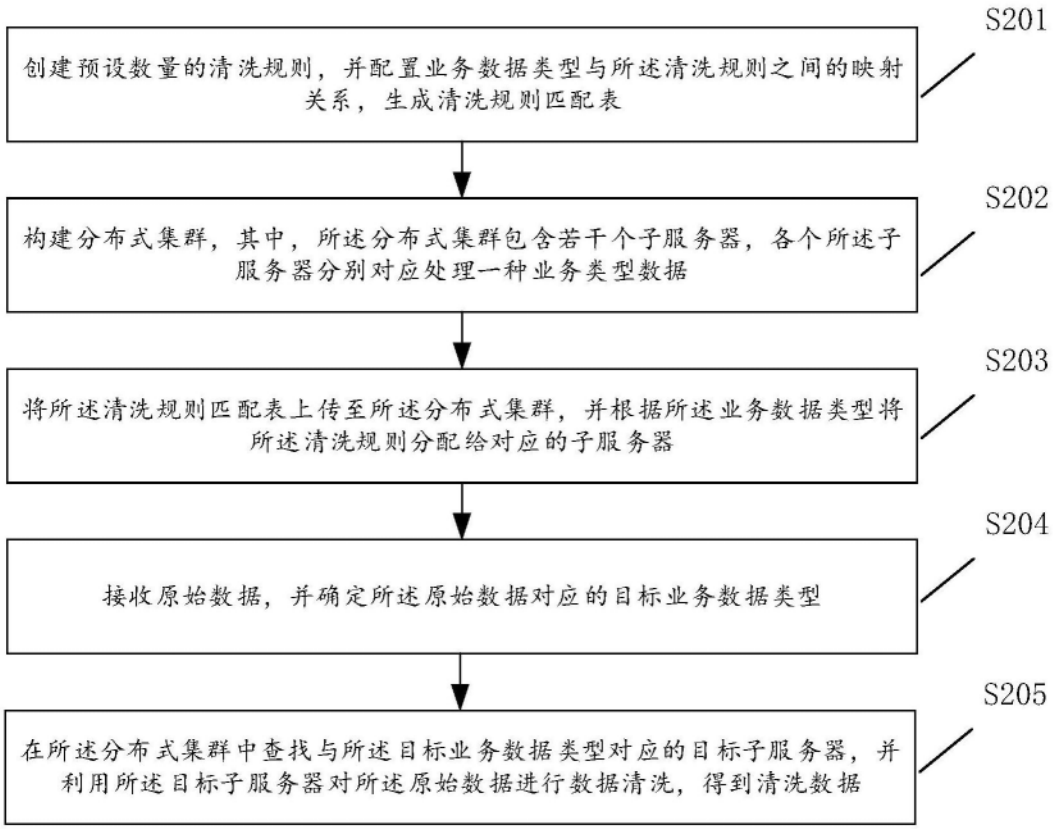

S201
创建预设数量的清洗规则，并配置业务数据类型与所述清洗规则之间的映射关系，生成清洗规则匹配表
S202
构建分布式集群，其中，所述分布式集群包含若干个子服务器，各个所述子服务器分别对应处理一种业务类型数据
S203
将所述清洗规则匹配表上传至所述分布式集群，并根据所述业务数据类型将所述清洗规则分配给对应的子服务器
S204
接收原始数据，并确定所述原始数据对应的目标业务数据类型
S205
在所述分布式集群中查找与所述目标业务数据类型对应的目标子服务器，并利用所述目标子服务器对所述原始数据进行数据清洗，得到清洗数据

图2

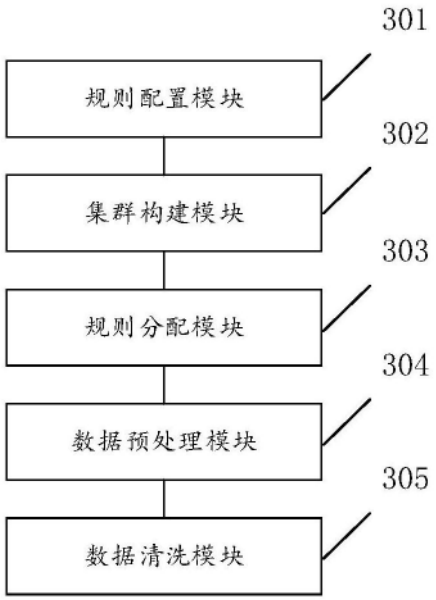

301
规则配置模块
302
集群构建模块
303
规则分配模块
304
数据预处理模块
305
数据清洗模块

图3

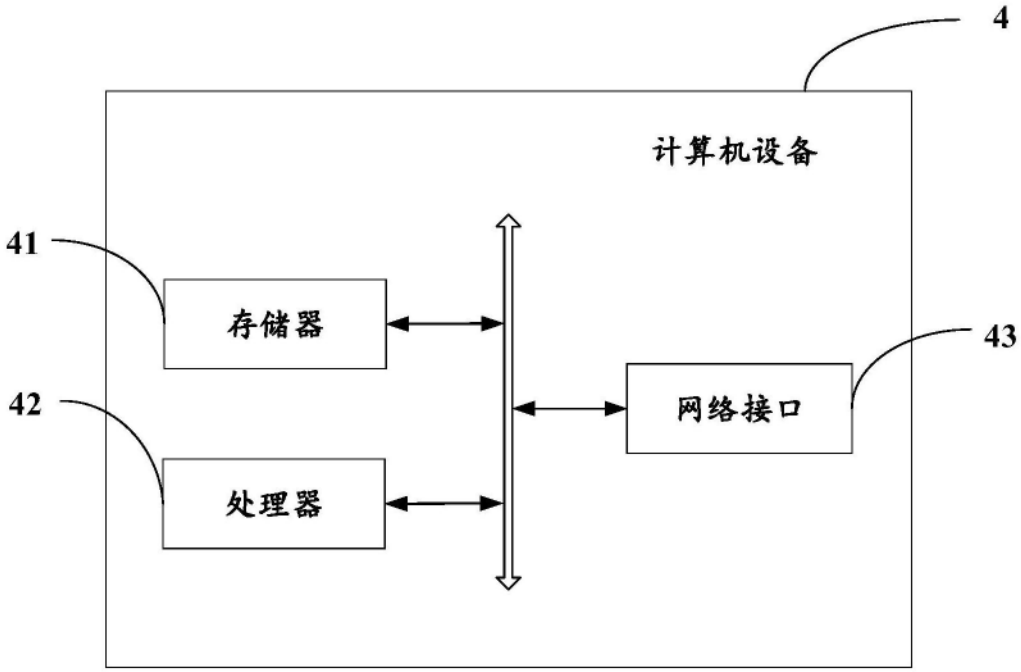

4
计算机设备
41
存储器
42
处理器
43
网络接口

图4