(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111258997 B
(45)授权公告日 2023.11.03

(21)申请号 202010047169.8

(22)申请日 2020.01.16

(65)同一申请的已公布的文献号
申请公布号 CN 111258997 A

(43)申请公布日 2020.06.09

(73)专利权人 浪潮软件股份有限公司
地址 271000 山东省泰安市东岳大街527号浪潮科技园

(72)发明人 胡振升 胡清 王建华

(74)专利代理机构 济南信达专利事务所有限公司 37100
专利代理师 程佩玉

(51)Int.Cl.
G06F 16/215(2019.01)
G06F 16/28(2019.01)

(56)对比文件
CN 109753502 A,2019.05.14
CN 109376196 A,2019.02.22
US 2019370263 A1,2019.12.05
JP H07239792 A,1995.09.12
US 2019392002 A1,2019.12.26
CN 109558400 A,2019.04.02
US 2016203198 A1,2016.07.14
CN 107330028 A,2017.11.07
CN 109299183 A,2019.02.01
梁美红;张男楠;李建;伍东;胡永泉;杨静.一种钻井数据仓库ETL系统的设计.计算机技术与发展.2010,(03),全文.
孟坚,董逸生,王永利.一种基于规则的交互式数据清洗技术.微机发展.2005,(04),全文.

审查员 吴海旋

权利要求书3页 说明书11页 附图3页

(54)发明名称

基于NiFi的数据处理方法和装置

(57)摘要

本发明提供了基于NiFi的数据处理方法和装置,包括:预先部署NiFi组件,并设置用于数据清洗的清洗组件和用于数据转换的转换组件,还包括:将清洗组件和转换组件嵌入在NiFi组件中;利用嵌入后的NiFi组件执行:采集数据源中的待处理数据;调用清洗组件和转换组件对待处理数据进行处理,得到处理后的数据;将处理后的数据存储到数据仓库中。本方案能够去除杂乱的数据。

CN 111258997 B

1.基于NiFi的数据处理方法,其特征在于,预先部署NiFi组件,并设置用于数据清洗的清洗组件和用于数据转换的转换组件,还包括:

将所述清洗组件和所述转换组件嵌入在所述NiFi组件中;

利用嵌入后的所述NiFi组件执行:

采集数据源中的待处理数据;

调用所述清洗组件和所述转换组件对所述待处理数据进行处理,得到处理后的数据;

将所述处理后的数据存储到数据仓库中;

所述调用所述清洗组件和所述转换组件对所述待处理数据进行处理,得到处理后的数据,包括:

S1:确定所述清洗组件的处理优先级是否高于所述转换组件的处理优先级,如果是,执行S2,否则,执行S3;

S2:调用所述清洗组件对所述待处理数据进行数据清洗,将清洗后的数据作为待处理数据,并调用所述转换组件对该待处理数据进行数据转换;

S3:调用所述转换组件对所述待处理数据进行数据转换,将转换后的数据作为待处理数据,并调用所述清洗组件对该待处理数据进行数据清洗;

在所述S1之前,进一步包括:

确定至少一个字段名筛选条件;

所述S2中的调用所述清洗组件对所述待处理数据进行数据清洗,包括:

调用所述清洗组件执行:

D1:确定待清洗集合,其中,所述待清洗集合中包括所述待处理数据中的至少一个第一字段名;

D2:从所述待清洗集合中确定当前第一字段名;

D3:确定所述当前第一字段名是否与所述至少一个字段名筛选条件相匹配,如果是,执行D4,否则,执行D5;

D4:从所述待处理数据中抽取所述第一字段名指示的字段,执行D5;

D5:确定所述当前第一字段名是否为所述待清洗集合中最后一个第一字段名,如果是,结束当前流程,否则,执行D6;

D6:从所述待清洗集合中删除所述当前第一字段名,返回D2;

和/或,

在所述S1之前,进一步包括:

确定至少一个转换条件;

所述S3中的调用所述转换组件对所述待处理数据进行数据转换,包括:

调用所述转换组件执行:

F1:确定待转换集合,其中,所述待转换集合包括所述待处理数据中的至少一个第二字段名;

F2:从所述待转换集合中确定当前第二字段名;

F3:按照所述至少一个转换条件对所述当前第二字段名指示的字段进行转换;

F4:确定所述当前第二字段名是否为所述待转换集合中的最后一个第二字段名,如果是,结束当前流程,否则,执行F5;

F5:从所述待转换集合中删除所述当前第二字段名,返回F2;

所述将所述清洗组件和所述转换组件嵌入在所述NiFi组件中,包括:

确定所述NiFi组件的配置格式;

按照所述配置格式分别对所述清洗组件和所述转换组件进行打包,得到打包后的数据包;

将所述数据包添加到所述NiFi组件中;

运行添加所述数据包后的所述NiFi组件。

2.基于NiFi的数据处理装置,其特征在于,包括:

组件设置模块,用于预先部署NiFi组件,并设置用于数据清洗的清洗组件和用于数据转换的转换组件;

组件嵌入模块,用于将所述组件设置模块设置的所述清洗组件和所述转换组件嵌入在所述NiFi组件中;

数据采集模块,用于利用所述组件嵌入模块嵌入后的所述NiFi组件,采集数据源中的待处理数据;

数据清洗转换模块,用于利用所述组件嵌入模块嵌入后的所述NiFi组件,调用所述清洗组件和所述转换组件对所述数据采集模块采集的所述待处理数据进行处理,得到处理后的数据;

数据存储模块,用于将所述数据清洗转换模块得到的所述处理后的数据存储到数据仓库中;

所述数据清洗转换模块,包括:判断模块、清洗模块和转换模块;

所述判断模块,用于确定所述清洗组件的处理优先级是否高于所述转换组件的处理优先级,如果是,触发所述清洗模块调用所述清洗组件对所述待处理数据进行数据清洗,将清洗后的数据作为待处理数据,并触发所述转换模块调用所述转换组件对所述清洗模块清洗后的该待处理数据进行数据转换;否则,触发所述转换模块调用所述转换组件对所述待处理数据进行数据转换,将转换后的数据作为待处理数据,并触发所述清洗模块调用所述清洗组件对所述转换模块转换的该待处理数据进行数据清洗;

所述判断模块,进一步用于确定至少一个字段名筛选条件;

所述清洗模块,用于调用所述清洗组件执行:

D1:确定待清洗集合,其中,所述待清洗集合中包括所述待处理数据中的至少一个第一字段名;

D2:从所述待清洗集合中确定当前第一字段名;

D3:确定所述当前第一字段名是否与所述至少一个字段名筛选条件相匹配,如果是,执行D4,否则,执行D5;

D4:从所述待处理数据中抽取所述第一字段名指示的字段,执行D5;

D5:确定所述当前第一字段名是否为所述待清洗集合中最后一个第一字段名,如果是,结束当前流程,否则,执行D6;

D6:从所述待清洗集合中删除所述当前第一字段名,返回D2;

和/或,

所述判断模块,进一步用于确定至少一个转换条件;

所述转换模块，用于调用所述转换组件执行：

F1：确定待转换集合，其中，所述待转换集合包括所述待处理数据中的至少一个第二字段名；

F2：从所述待转换集合中确定当前第二字段名；

F3：按照所述至少一个转换条件对所述当前第二字段名指示的字段进行转换；

F4：确定所述当前第二字段名是否为所述待转换集合中的最后一个第二字段名，如果是，结束当前流程，否则，执行F5；

F5：从所述待转换集合中删除所述当前第二字段名，返回F2；

所述组件嵌入模块，用于确定所述NiFi组件的配置格式；按照所述配置格式分别对所述清洗组件和所述转换组件进行打包，得到打包后的数据包；将所述数据包添加到所述NiFi组件中；运行添加所述数据包后的所述NiFi组件。

3.存储介质，其特征在于，包括：至少一个存储器和至少一个处理器；

所述至少一个存储器，用于存储机器可读程序；

所述至少一个处理器，用于调用所述机器可读程序，执行权利要求1所述的基于NiFi的数据处理方法。

4.计算机可读介质，其特征在于，所述计算机可读介质上存储有计算机指令，所述计算机指令在被处理器执行时，使所述处理器执行权利要求1所述的基于NiFi的数据处理方法。

基于NiFi的数据处理方法和装置

技术领域

[0001] 本发明涉及计算机技术领域,特别涉及基于NiFi的数据处理方法和装置。

背景技术

[0002] 随着科技发达,信息流通,人们之间的交流越来越密切,生活也越来越方便,大数据逐渐成为高科技时代的产物。大数据是如此重要,以至于其数据采集、储存、搜索、共享、分析,乃至可视化地呈现,都成为了当前重要的研究课题。

[0003] 目前,数据仓库中存储的数据通常是面向某一主题的数据的集合,而这些数据通常从多个业务系统中抽取而来,并且通常包含历史数据。这样就避免不了有的数据是错误数据、有的数据相互之间有冲突,从而不利于数据分析。

发明内容

[0004] 本发明实施例提供了基于NiFi的数据处理方法和装置,能够去除杂乱的数据。

[0005] 第一方面,本发明提供了基于NiFi的数据处理方法,预先部署NiFi组件,并设置用于数据清洗的清洗组件和用于数据转换的转换组件,还包括:

[0006] 将所述清洗组件和所述转换组件嵌入在所述NiFi组件中;

[0007] 利用嵌入后的所述NiFi组件执行:

[0008] 采集数据源中的待处理数据;

[0009] 调用所述清洗组件和所述转换组件对所述待处理数据进行处理,得到处理后的数据;

[0010] 将所述处理后的数据存储到数据仓库中。

[0011] 优选地,

[0012] 所述调用所述清洗组件和所述转换组件对所述待处理数据进行处理,得到处理后的数据,包括:

[0013] S1:确定所述清洗组件的处理优先级是否高于所述转换组件的处理优先级,如果是,执行S2,否则,执行S3;

[0014] S2:调用所述清洗组件对所述待处理数据进行数据清洗,将清洗后的数据作为待处理数据,并调用所述转换组件对该待处理数据进行数据转换;

[0015] S3:调用所述转换组件对所述待处理数据进行数据转换,将转换后的数据作为待处理数据,并调用所述清洗组件对该待处理数据进行数据清洗。

[0016] 优选地,

[0017] 在所述S1之前,进一步包括:

[0018] 确定至少一个字段名筛选条件;

[0019] 所述S2中的调用所述清洗组件对所述待处理数据进行数据清洗,包括:

[0020] 调用所述清洗组件执行:

[0021] D1:确定待清洗集合,其中,所述待清洗集合中包括所述待处理数据中的至少一个

第一字段名；

[0022] D2:从所述待清洗集合中确定当前第一字段名；

[0023] D3:确定所述当前第一字段名是否与所述至少一个字段名筛选条件相匹配,如果是,执行D4,否则,执行D5；

[0024] D4:从所述待处理数据中抽取所述第一字段名指示的字段,执行D5；

[0025] D5:确定所述当前第一字段名是否为所述待清洗集合中最后一个第一字段名,如果是,结束当前流程,否则,执行D6；

[0026] D6:从所述待清洗集合中删除所述当前第一字段名,返回D2。

[0027] 优选地，

[0028] 在所述S1之前,进一步包括：

[0029] 确定至少一个转换条件；

[0030] 所述S3中的调用所述转换组件对所述待处理数据进行数据转换,包括：

[0031] 调用所述转换组件执行：

[0032] F1:确定待转换集合,其中,所述待转换集合包括所述待处理数据中的至少一个第二字段名；

[0033] F2:从所述待转换集合中确定当前第二字段名；

[0034] F3:按照所述至少一个转换条件对所述当前第二字段名指示的字段进行转换；

[0035] F4:确定所述当前第二字段名是否为所述待转换集合中的最后一个第二字段名,如果是,结束当前流程,否则,执行F5；

[0036] F5:从所述待转换集合中删除所述当前第二字段名,返回F2。

[0037] 优选地，

[0038] 所述将所述清洗组件和所述转换组件嵌入在所述NiFi组件中,包括：

[0039] 确定所述NiFi组件的配置格式；

[0040] 按照所述配置格式分别对所述清洗组件和所述转换组件进行打包,得到打包后的数据包；

[0041] 将所述数据包添加到所述NiFi组件中；

[0042] 运行添加所述数据包后的所述NiFi组件。

[0043] 第二方面,本发明提供了基于NiFi的数据处理装置,包括：

[0044] 组件设置模块,用于预先部署NiFi组件,并设置用于数据清洗的清洗组件和用于数据转换的转换组件；

[0045] 组件嵌入模块,用于将所述组件设置模块设置的所述清洗组件和所述转换组件嵌入在所述NiFi组件中；

[0046] 数据采集模块,用于利用所述组件嵌入模块嵌入后的所述NiFi组件,采集数据源中的待处理数据；

[0047] 数据清洗转换模块,用于利用所述组件嵌入模块嵌入后的所述NiFi组件,调用所述清洗组件和所述转换组件对所述数据采集模块采集的所述待处理数据进行处理,得到处理后的数据；

[0048] 数据存储模块,用于将所述数据清洗转换模块得到的所述处理后的数据存储到数据仓库中。

[0049] 优选地，

[0050] 所述数据清洗转换模块，包括：判断模块、清洗模块和转换模块；

[0051] 所述判断模块，用于确定所述清洗组件的处理优先级是否高于所述转换组件的处理优先级，如果是，触发所述清洗模块调用所述清洗组件对所述待处理数据进行数据清洗，将清洗后的数据作为待处理数据，并触发所述转换模块调用所述转换组件对所述清洗模块清洗后的该待处理数据进行数据转换；否则，触发所述转换模块调用所述转换组件对所述待处理数据进行数据转换，将转换后的数据作为待处理数据，并触发所述清洗模块调用所述清洗组件对所述转化模块转换的该待处理数据进行数据清洗。

[0052] 优选地，

[0053] 所述判断模块，进一步用于确定至少一个字段名筛选条件；

[0054] 所述清洗模块，用于调用所述清洗组件执行：

[0055] D1：确定待清洗集合，其中，所述待清洗集合中包括所述待处理数据中的至少一个第一字段名；

[0056] D2：从所述待清洗集合中确定当前第一字段名；

[0057] D3：确定所述当前第一字段名是否与所述至少一个字段名筛选条件相匹配，如果是，执行D4，否则，执行D5；

[0058] D4：从所述待处理数据中抽取所述第一字段名指示的字段，执行D5；

[0059] D5：确定所述当前第一字段名是否为所述待清洗集合中最后一个第一字段名，如果是，结束当前流程，否则，执行D6；

[0060] D6：从所述待清洗集合中删除所述当前第一字段名，返回D2。

[0061] 优选地，

[0062] 所述判断模块，进一步用于确定至少一个转换条件；

[0063] 所述转换模块，用于调用所述转换组件执行：

[0064] F1：确定待转换集合，其中，所述待转换集合包括所述待处理数据中的至少一个第二字段名；

[0065] F2：从所述待转换集合中确定当前第二字段名；

[0066] F3：按照所述至少一个转换条件对所述当前第二字段名指示的字段进行转换；

[0067] F4：确定所述当前第二字段名是否为所述待转换集合中的最后一个第二字段名，如果是，结束当前流程，否则，执行F5；

[0068] F5：从所述待转换集合中删除所述当前第二字段名，返回F2。

[0069] 优选地，

[0070] 所述组件嵌入模块，用于确定所述NiFi组件的配置格式；按照所述配置格式分别对所述清洗组件和所述转换组件进行打包，得到打包后的数据包；将所述数据包添加到所述NiFi组件中；运行添加所述数据包后的所述NiFi组件。

[0071] 第三方面，本发明还提供了存储介质，包括：至少一个存储器和至少一个处理器；

[0072] 所述至少一个存储器，用于存储机器可读程序；

[0073] 所述至少一个处理器，用于调用所述机器可读程序，执行上述任一项中所述的基于NiFi的数据处理方法。

[0074] 第四方面，本发明还提供了计算机可读介质，所述计算机可读介质上存储有计算

机指令，所述计算机指令在被处理器执行时，使所述处理器执行上述任一项中所述的基于NiFi的数据处理方法。

[0075] 本发明实施例提供了NiFi的数据处理方法和装置，在NiFi组件中嵌入用于进行数据清洗的清洗组件和用于数据转换的转换组件，可以在数据处理过程中，由NiFi组件负责从数据源中抽取数据，然后调用清洗组件和转换组件来对抽取的数据进行相应数据清洗、转换操作，最后再将处理后的数据入库，实现获取去除杂乱后的数据的目的。并且，由于NiFi组件为开源的程序，因此，通过在其内部设置相应的数据清洗、转换程序，即可满足数据的处理需求，无需开发人员编写大量数据处理相关的代码，从而降低获取去除杂乱后的数据的难度。

附图说明

[0076] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施例，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。

[0077] 图1是本发明一实施例提供的基于NiFi的数据处理方法的流程图；

[0078] 图2是本发明另一实施例提供的基于NiFi的数据处理方法的流程图；

[0079] 图3是本发明一实施例提供的基于NiFi的数据处理装置的结构示意图；

[0080] 图4是本发明另一实施例提供的基于NiFi的数据处理装置的结构示意图。

具体实施方式

[0081] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例，基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动的前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0082] 如图1所示，本发明实施例提供了基于NiFi的数据处理方法，包括：

[0083] 步骤101：预先部署NiFi组件；

[0084] 步骤102：设置用于数据清洗的清洗组件和用于数据转换的转换组件；

[0085] 步骤103：将所述清洗组件和所述转换组件嵌入在所述NiFi组件中；

[0086] 步骤104：利用嵌入后的所述NiFi组件执行：采集数据源中的待处理数据；

[0087] 步骤105：调用所述清洗组件和所述转换组件对所述待处理数据进行处理，得到处理后的数据；

[0088] 步骤106：将所述处理后的数据存储到数据仓库中。

[0089] 在本发明实施例中，在NiFi组件中嵌入用于进行数据清洗的清洗组件和用于数据转换的转换组件，可以在数据处理过程中，由NiFi组件负责从数据源中抽取数据，然后调用清洗组件和转换组件来对抽取的数据进行相应数据清洗、转换操作，最后再将处理后的数据入库，实现获取去除杂乱后的数据的目的。并且，由于NiFi组件为开源的程序，因此，通过在其内部设置相应的数据清洗、转换程序，即可满足数据的处理需求，无需开发人员编写大量数据处理相关的代码，从而降低获取去除杂乱后的数据的难度。

[0090] 在本发明一实施例中，所述调用所述清洗组件和所述转换组件对所述待处理数据进行处理，得到处理后的数据，包括：

[0091] S1：确定所述清洗组件的处理优先级是否高于所述转换组件的处理优先级，如果是，执行S2，否则，执行S3；

[0092] S2：调用所述清洗组件对所述待处理数据进行数据清洗，将清洗后的数据作为待处理数据，并调用所述转换组件对该待处理数据进行数据转换；

[0093] S3：调用所述转换组件对所述待处理数据进行数据转换，将转换后的数据作为待处理数据，并调用所述清洗组件对该待处理数据进行数据清洗。

[0094] 在本发明实施例中，嵌入在NiFi组件中的清洗组件和转换组件对于数据处理的优先级可能不同，因此，在NiFi组件抽取到数据后，如果清洗组件高于转换组件的处理优先级，则清洗组件优先对NiFi组件抽取的数据进行数据清洗处理，然后再由转换组件对数据进行数据转换处理。反之，则转换组件优先对NiFi组件抽取的数据进行数据转换处理，然后再由清洗组件进行数据清洗处理。

[0095] 在本发明一实施例中，在所述S1之前，进一步包括：

[0096] 确定至少一个字段名筛选条件；

[0097] 所述S2中的调用所述清洗组件对所述待处理数据进行数据清洗，包括：

[0098] 调用所述清洗组件执行：

[0099] D1：确定待清洗集合，其中，所述待清洗集合中包括所述待处理数据中的至少一个第一字段名；

[0100] D2：从所述待清洗集合中确定当前第一字段名；

[0101] D3：确定所述当前第一字段名是否与所述至少一个字段名筛选条件相匹配，如果是，执行D4，否则，执行D5；

[0102] D4：从所述待处理数据中抽取所述第一字段名指示的字段，执行D5；

[0103] D5：确定所述当前第一字段名是否为所述待清洗集合中最后一个第一字段名，如果是，结束当前流程，否则，执行D6；

[0104] D6：从所述待清洗集合中删除所述当前第一字段名，返回D2。

[0105] 在本发明实施例中，字段名筛选条件可以包括：用于校验待处理数据中的第一字段名所指示的字段为非空数据的非空校验、第一字段名的前缀包括所要筛选的字段信息且第一字段名指示的字段为非空数据的前缀校验(非)、第一字段名的后缀包括所要筛选的字段信息且第一字段名指示的字段为非空数据的后缀校验(非)、第一字段名的长度达到一定要求的最大长度校验和最小长度校验中的至少一个，但不限于此。在对数据进行数据清洗时，由于第一字段名可能在满足某一字段名筛选条件时不满足其他的字段名筛选条件，为了避免筛选出无用的数据，需要待处理数据中的每一个第一字段名符合所有的字段名筛选条件时，才能对符合条件的第一字段名指示的字段进行抽取操作。

[0106] 具体地，根据需求还可以对待处理数据中的字段名进行空数据校验，即校验待处理数据中的第一字段名所指示的字段中不包含任何数据，然后对不包含任何数据的第一字段名进行记录，以便确定记录的第一字段名中不包含任何数据的原因。

[0107] 在本发明一实施例中，在所述S1之前，进一步包括：

[0108] 确定至少一个转换条件；

[0109] 所述S3中的调用所述转换组件对所述待处理数据进行数据转换，包括：

[0110] 调用所述转换组件执行：

[0111] F1：确定待转换集合，其中，所述待转换集合包括所述待处理数据中的至少一个第二字段名；

[0112] F2：从所述待转换集合中确定当前第二字段名；

[0113] F3：按照所述至少一个转换条件对所述当前第二字段名指示的字段进行转换；

[0114] F4：确定所述当前第二字段名是否为所述待转换集合中的最后一个第二字段名，如果是，结束当前流程，否则，执行F5；

[0115] F5：从所述待转换集合中删除所述当前第二字段名，返回F2。

[0116] 在本发明实施例中，数据的转换条件可以包括：日期格式转换、按照字典进行不同语言转换的字典转换、普通值替换、数据中的控制替换为指定字符的空值替换、正则替换、schema指定字段转换和schema大小写转换中的至少一个，但不限于此。在对数据进行转换处理时，由于待处理数据中的第二字段名可能在满足某一转换条件时不满足其他的转换条件，为了避免筛选出的数据非所需的，因此需要将每一个第二字段名按照所有的转换条件进行相应的转换处理。

[0117] 在本发明一实施例中，所述将所述清洗组件和所述转换组件嵌入在所述NiFi组件中，包括：

[0118] 确定所述NiFi组件的配置格式；

[0119] 按照所述配置格式分别对所述清洗组件和所述转换组件进行打包，得到打包后的数据包；

[0120] 将所述数据包添加到所述NiFi组件中；

[0121] 运行添加所述数据包后的所述NiFi组件。

[0122] 在本发明实施例中，由于NiFi组件的文件有其专属的格式，因此为了使得NiFi能够调用即将嵌入的清洗组件和转换组件，需要将清洗组件和转换组件打包成与NiFi组件的配置格式相匹配的数据包，然后再将数据包放置到NiFi组件可以调用的位置处，对嵌入数据包后的NiFi组件进行重启，在需要的数据进行清洗和转换时，NiFi组件即可调用被嵌入的组件进行相应操作。

[0123] 如图2所示，为了更加清楚地说明本发明的技术方案及优点，下面以清洗组件的处理优先级高于转换组件的处理优先级为例，对本发明实施例提供的基于NiFi的数据处理方法进行详细说明，具体可以包括以下步骤：

[0124] 步骤201：预先部署NiFi组件。

[0125] 步骤202：设置用于数据清洗的清洗组件和用于数据转换的转换组件。

[0126] 步骤203：确定NiFi组件的配置格式。

[0127] 具体地，不同的组件有自己专属的格式，因此为了使NiFi能够调用嵌入其内部的组件，需要依据NiFi组件文件的格式对待嵌入的组件进行处理。

[0128] 步骤204：按照配置格式分别对清洗组件和转换组件进行打包，得到打包后的数据包。

[0129] 步骤205：将数据包添加到NiFi组件中。

[0130] 步骤206：运行添加数据包后的NiFi组件。

[0131] 举例来说，当NiFi组件的配置格式为avro格式，则需要将能够进行数据清洗处理的清洗组件和能够对数据进行转换处理的转换组件，打包成与avro格式相匹配的数据包。然后将数据包添加到NiFi组件中，为了便于NiFi组件能够识别并调用该数据包，需要对添加数据包后的NiFi组件进行重启操作。

[0132] 步骤207：确定至少一个字段名筛选条件和至少一个转换条件。

[0133] 具体地，如需从NiFi组件抽取的数据中获取有用的数据，开发人员可以根据需求设置数据清洗过程中对数据进行筛选的字段名筛选条件，以及便于不同业务系统中数据的格式统一的转换条件。

[0134] 比如，字段名筛选条件“非空校验”和转换条件“大写转换小写”。

[0135] 步骤208：利用嵌入后的NiFi组件执行：采集数据源中的待处理数据。

[0136] 具体地，根据需求NiFi组件可以从不同的数据源中采集数据，其中，数据源可以包括：FTP/SFPT、关系型数据库、消息队列以及物联网中的至少一个，但不限于此。

[0137] 步骤209：调用清洗组件确定待清洗集合，其中，待清洗集合中包括待处理数据中的至少一个第一字段名。

[0138] 具体地，NiFi组件在从关系型数据库中采集到字段名为“iGetJDBC”和“abcd”，其中，字段名为“iGetJDBC”所指示的字段内容为“iGetJDBC 1.8.0”和字段名“abcd”所指示的字段内容为“IS_PARENT”。设置由字段名为“iGetJDBC”和“abcd”组成的待清洗集合。

[0139] 步骤210：从待清洗集合中确定当前第一字段名。

[0140] 具体地，在待清洗集合中可以根据需求从中确定任意一个当前第一字段名，也可以根据字段名的字母、数字、字符长度等条件确定。例如，从待清洗集合中随机确定当前第一字段名“iGetJDBC”。

[0141] 步骤211：确定当前第一字段名是否与各个字段名筛选条件相匹配，如果是，执行步骤212，否则，执行步骤213。

[0142] 具体地，由于字段名筛选条件为“非空校验”，所以只需确认当前第一字段名“iGetJDBC”所指示的字段是否为空数据，即是否存在数据，由于当前第一字段名“iGetJDBC”所指示的字段为“iGetJDBC 1.8.0”，不是空数据，并且字段名筛选条件仅有字段名筛选条件“非空校验”一个条件，因此，当前第一字段名“iGetJDBC”校验通过，可以抽取其指示的字段为“iGetJDBC1.8.0”。

[0143] 步骤212：从待处理数据中抽取第一字段名指示的字段，执行步骤213。

[0144] 步骤213：确定当前第一字段名是否为待清洗集合中最后一个第一字段名，如果是，执行步骤215，否则，执行步骤214。

[0145] 步骤214：从待清洗集合中删除当前第一字段名，返回步骤210。

[0146] 具体地，由于当前第一字段名“iGetJDBC”不是待清洗结合中的最后一个第一字段名，在待清洗集合中还存在一个第一字段名“abcd”，因此，为了防止对已筛选过得字段名多次清洗筛选，可以从待清洗集合中删除当前第一字段名为“iGetJDBC”，并将待清洗集合中的第一字段名“abcd”作为当前第一字段名，基于清洗条件进行校验，直至待清洗集合中不存在未筛选的第一字段名为止。

[0147] 步骤215：将数据清洗后的数据作为待处理数据，执行步骤216。

[0148] 具体地，在待清洗集合中所有的第一字段名均被筛选完毕后，可以将经过清洗条

件筛选后的第一字段名和其指示的字段作为待处理数据，进行数据转换处理。

[0149] 步骤216：调用转换组件确定待转换集合，其中，待转换集合包括待处理数据中的至少一个第二字段名。

[0150] 具体地，由于经过筛选后的第一字段名“iGetJDBC”和“abcd”均符合清洗条件，因此，可以将其作为第二字段名形成待转换集合。

[0151] 步骤217：从待转换集合中确定当前第二字段名。

[0152] 具体地，在待转换集合中可以根据需求从中确认任意一个当前第二字段名，也可以根据字段名的字母、数字、字符长度等条件确定。例如，从待转换集合中随机确定当前第二字段名“abcd”。

[0153] 步骤218：按照各个转换条件对当前第二字段名指示的字段进行转换。

[0154] 具体地，转换条件为“大写转换小写”，且述当前第二字段名指示的字段“IS_PARENT”中包括大写字符，因此，需要依据转换条件转换为“is_parent”，由于转换条件仅有“大写转换小写”一个，因此，当前第二字段名指示的字段的数据转换处理操作即为完成。

[0155] 步骤219：确定当前第二字段名是否为待转换集合中的最后一个第二字段名，如果是，执行步骤221，否则，执行步骤220。

[0156] 步骤220：从待转换集合中删除当前第二字段名，返回步骤217。

[0157] 具体地，由于当前第二字段名“abcd”不是待转换集合中的最后一个第二字段名，因此，需要将其从待转换集合中删除，防止对待转换集合中的第二字段名指示的字段进行多次转换操作。并将待转换集合中的第二字段名“iGetJDBC”作为当前第二字段名，基于转换条件对其指示的字段进行数据转换，直至待转换集合中的所有第二字段名指示的字段均完成数据转换操作为止。

[0158] 步骤221：将处理后的数据存储到数据仓库中。

[0159] 具体地，在基于需求对所需要的字段名指示的字段进行数据清洗、转换操作后，NiFi组件可以对处理后的数据进行存储，完成数据处理操作。

[0160] 需要说明的是，NiFi组件可以是单例也可以是以集群的形式存在，当NiFi组件以集群的形式存在时，该集群中存在主NiFi组件和副NiFi组件，主NiFi组件可以对所有的副NiFi组件需要进行调用、管理。

[0161] 如图3所示，本发明实施例提供了基于NiFi的数据处理装置，包括：

[0162] 组件设置模块301，用于预先部署NiFi组件，并设置用于数据清洗的清洗组件和用于数据转换的转换组件；

[0163] 组件嵌入模块302，用于将所述组件设置模块301设置的所述清洗组件和所述转换组件嵌入在所述NiFi组件中；

[0164] 数据采集模块303，用于利用所述组件嵌入模块302嵌入后的所述NiFi组件，采集数据源中的待处理数据；

[0165] 数据清洗转换模块304，用于利用所述组件嵌入模块302嵌入后的所述NiFi组件，调用所述清洗组件和所述转换组件对所述数据采集模块303采集的所述待处理数据进行处理，得到处理后的数据；

[0166] 数据存储模块305，用于将所述数据清洗转换模块304得到的所述处理后的数据存储到数据仓库中。

[0167] 在本发明实施例中，通过组件嵌入模块在组价设置模块部署的NiFi组件中，嵌入组件设置模块设置的用于进行数据清洗的清洗组件和用于数据转换的转换组件，可以在数据处理过程中，通过数据采集模块由NiFi组件负责从数据源中抽取数据，然后通过数据清洗转换模块利用清洗组件和转换组件来对抽取的数据进行相应数据清洗、转换操作，最后再由数据存储模块利用NiFi将处理后的数据入库，实现获取去除杂乱后的数据的目的。并且，由于NiFi组件为开源的程序，因此，通过在其内部设置相应的数据清洗、转换程序，即可满足数据的处理需求，无需开发人员编写大量数据处理相关的代码，从而降低获取去除杂乱后的数据的难度。

[0168] 如图4所示，在本发明实施例中，所述数据清洗转换模块303，包括：判断模块3031、清洗模块3032和转换模块3033；

[0169] 所述判断模块3031，用于确定所述清洗组件的处理优先级是否高于所述转换组件的处理优先级，如果是，触发所述清洗模块3032调用所述清洗组件对所述待处理数据进行数据清洗，将清洗后的数据作为待处理数据，并触发所述转换模块3033调用所述转换组件对所述清洗模块清洗后的该待处理数据进行数据转换；否则，触发所述转换模块3033调用所述转换组件对所述待处理数据进行数据转换，将转换后的数据作为待处理数据，并触发所述清洗模块3032调用所述清洗组件对所述转化模块转换的该待处理数据进行数据清洗。

[0170] 在本发明实施例中，所述判断模块，进一步用于确定至少一个字段名筛选条件；

[0171] 所述清洗模块，用于调用所述清洗组件执行：

[0172] D1：确定待清洗集合，其中，所述待清洗集合中包括所述待处理数据中的至少一个第一字段名；

[0173] D2：从所述待清洗集合中确定当前第一字段名；

[0174] D3：确定所述当前第一字段名是否与所述至少一个字段名筛选条件相匹配，如果是，执行D4，否则，执行D5；

[0175] D4：从所述待处理数据中抽取所述第一字段名指示的字段，执行D5；

[0176] D5：确定所述当前第一字段名是否为所述待清洗集合中最后一个第一字段名，如果是，结束当前流程，否则，执行D6；

[0177] D6：从所述待清洗集合中删除所述当前第一字段名，返回D2。

[0178] 在本发明实施例中，所述判断模块，进一步用于确定至少一个转换条件；

[0179] 所述转换模块，用于调用所述转换组件执行：

[0180] F1：确定待转换集合，其中，所述待转换集合包括所述待处理数据中的至少一个第二字段名；

[0181] F2：从所述待转换集合中确定当前第二字段名；

[0182] F3：按照所述至少一个转换条件对所述当前第二字段名指示的字段进行转换；

[0183] F4：确定所述当前第二字段名是否为所述待转换集合中的最后一个第二字段名，如果是，结束当前流程，否则，执行F5；

[0184] F5：从所述待转换集合中删除所述当前第二字段名，返回F2。

[0185] 在本发明实施例中，所述组件嵌入模块，用于确定所述NiFi组件的配置格式；按照所述配置格式分别对所述清洗组件和所述转换组件进行打包，得到打包后的数据包；将所述数据包添加到所述NiFi组件中；运行添加所述数据包后的所述NiFi组件。

[0186] 可以理解的是，本发明实施例示意的结构并不构成对基于NiFi的数据处理装置的具体限定。在本发明的另一些实施例中，基于NiFi的数据处理装置可以包括比图示更多或者更少的部件，或者组合某些部件，或者拆分某些部件，或者不同的部件布置。图示的部件可以以硬件、软件或者软件和硬件的组合来实现。

[0187] 上述装置内的各单元之间的信息交互、执行过程等内容，由于与本发明方法实施例基于同一构思，具体内容可参见本发明方法实施例中的叙述，此处不再赘述。

[0188] 本发明实施例还提供了存储介质，包括：至少一个存储器和至少一个处理器；

[0189] 所述至少一个存储器，用于存储机器可读程序；

[0190] 所述至少一个处理器，用于调用所述机器可读程序，执行上述任一实施例中所述的基于NiFi的数据处理方法。

[0191] 本发明实施例还提供了计算机可读介质，所述计算机可读介质上存储有计算机指令，所述计算机指令在被处理器执行时，使所述处理器执行上述任一实施例中所述的基于NiFi的数据处理方法。

[0192] 具体地，可以提供配有存储介质的系统或者装置，在该存储介质上存储着实现上述实施例中任一实施例的功能的软件程序代码，且使该系统或者装置的计算机(或CPU或MPU)读出并执行存储在存储介质中的程序代码。

[0193] 在这种情况下，从存储介质读取的程序代码本身可实现上述实施例中任何一项实施例的功能，因此程序代码和存储程序代码的存储介质构成了本发明的一部分。

[0194] 用于提供程序代码的存储介质实施例包括软盘、硬盘、磁光盘、光盘(如CD-ROM、CD-R、CD-RW、DVD-ROM、DVD-RAM、DVD-RW、DVD+RW)、磁带、非易失性存储卡和ROM。可选择地，可以由通信网络从服务器计算机上下载程序代码。

[0195] 此外，应该清楚的是，不仅可以通过执行计算机所读出的程序代码，而且可以通过基于程序代码的指令使计算机上操作的操作系统等来完成部分或者全部的实际操作，从而实现上述实施例中任意一项实施例的功能。

[0196] 此外，可以理解的是，将由存储介质读出的程序代码写到插入计算机内的扩展板中所设置的存储器中或者写到与计算机相连接的扩展单元中设置的存储器中，随后基于程序代码的指令使安装在扩展板或者扩展单元上的CPU等来执行部分和全部实际操作，从而实现上述实施例中任一实施例的功能。

[0197] 本发明各个实施例至少具有如下有益效果：

[0198] 本方案从数据采集、清洗、转换，该方法易于实现，且开发成本低、灵活多样，高可扩展，具有广泛的适用场景。

[0199] 需要说明的是，上述各流程和各系统结构图中不是所有的步骤和模块都是必须的，可以根据实际的需要忽略某些步骤或模块。各步骤的执行顺序不是固定的，可以根据需要进行调整。上述各实施例中描述的系统结构可以是物理结构，也可以是逻辑结构，即，有些模块可能由同一物理实体实现，或者，有些模块可能分由多个物理实体实现，或者，可以由多个独立设备中的某些部件共同实现。

[0200] 以上各实施例中，硬件单元可以通过机械方式或电气方式实现。例如，一个硬件单元可以包括永久性专用的电路或逻辑(如专门的处理器，FPGA或ASIC)来完成相应操作。硬件单元还可以包括可编程逻辑或电路(如通用处理器或其它可编程处理器)，可以由软件进

行临时的设置以完成相应操作。具体的实现方式(机械方式、或专用的永久性电路、或者临时设置的电路)可以基于成本和时间上的考虑来确定。

[0201] 上文通过附图和优选实施例对本发明进行了详细展示和说明,然而本发明不限于这些已揭示的实施例,基与上述多个实施例本领域技术人员可以知晓,可以组合上述不同实施例中的代码审核手段得到本发明更多的实施例,这些实施例也在本发明的保护范围之内。

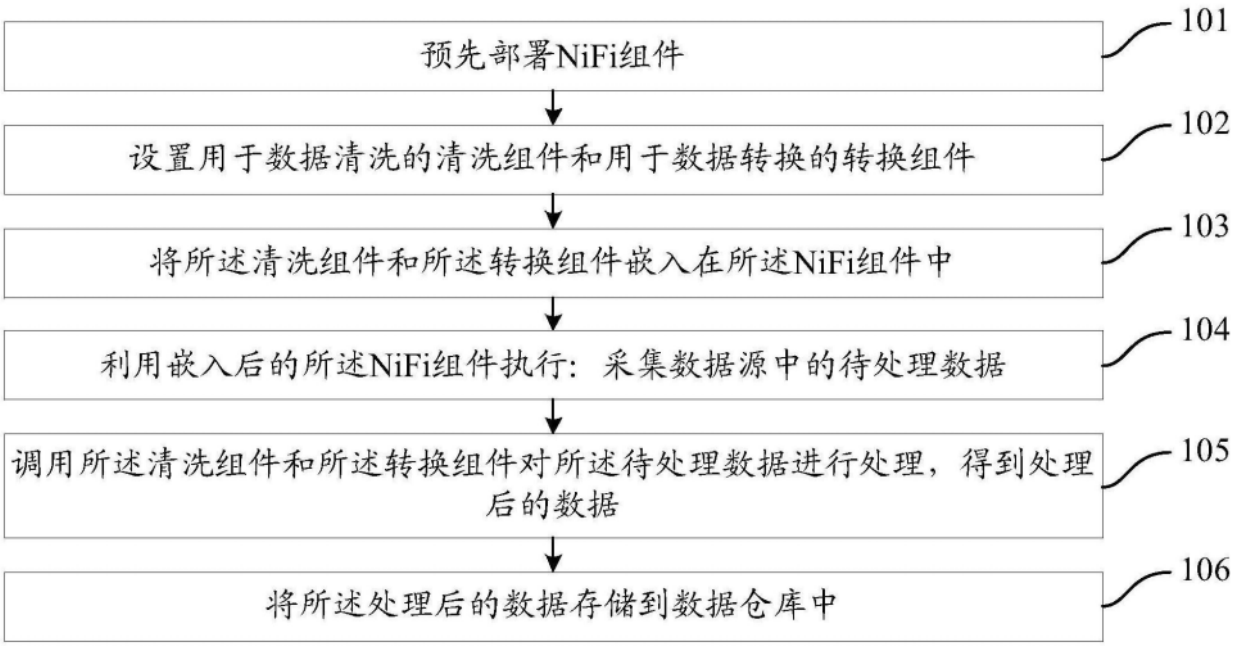
预先部署NiFi组件
101
设置用于数据清洗的清洗组件和用于数据转换的转换组件
102
将所述清洗组件和所述转换组件嵌入在所述NiFi组件中
103
利用嵌入后的所述NiFi组件执行：采集数据源中的待处理数据
104
调用所述清洗组件和所述转换组件对所述待处理数据进行处理，得到处理后的数据
105
将所述处理后的数据存储到数据仓库中
106

图1

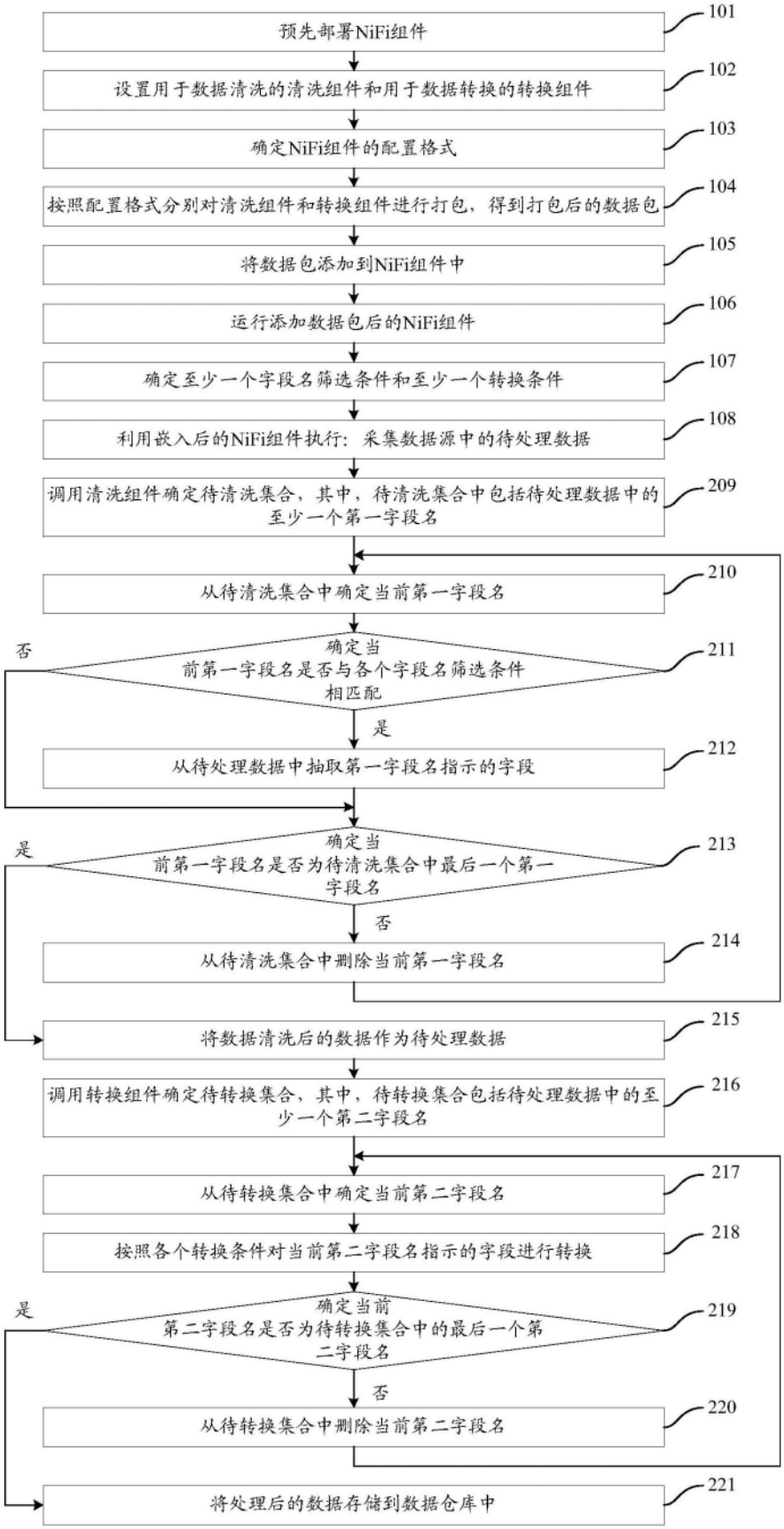
预先部署NiFi组件
101
设置用于数据清洗的清洗组件和用于数据转换的转换组件
102
确定NiFi组件的配置格式
103
按照配置格式分别对清洗组件和转换组件进行打包，得到打包后的数据包
104
将数据包添加到NiFi组件中
105
运行添加数据包后的NiFi组件
106
确定至少一个字段名筛选条件和至少一个转换条件
107
利用嵌入后的NiFi组件执行：采集数据源中的待处理数据
108
调用清洗组件确定待清洗集合，其中，待清洗集合中包括待处理数据中的至少一个第一字段名
209
从待清洗集合中确定当前第一字段名
210
否
确定当前第一字段名是否与各个字段名筛选条件相匹配
211
是
从待处理数据中抽取第一字段名指示的字段
212
是
确定当前第一字段名是否为待清洗集合中最后一个第一字段名
213
否
从待清洗集合中删除当前第一字段名
214
将数据清洗后的数据作为待处理数据
215
调用转换组件确定待转换集合，其中，待转换集合包括待处理数据中的至少一个第二字段名
216
从待转换集合中确定当前第二字段名
217
按照各个转换条件对当前第二字段名指示的字段进行转换
218
是
确定当前第二字段名是否为待转换集合中的最后一个第二字段名
219
否
从待转换集合中删除当前第二字段名
220
将处理后的数据存储到数据仓库中
221

图2

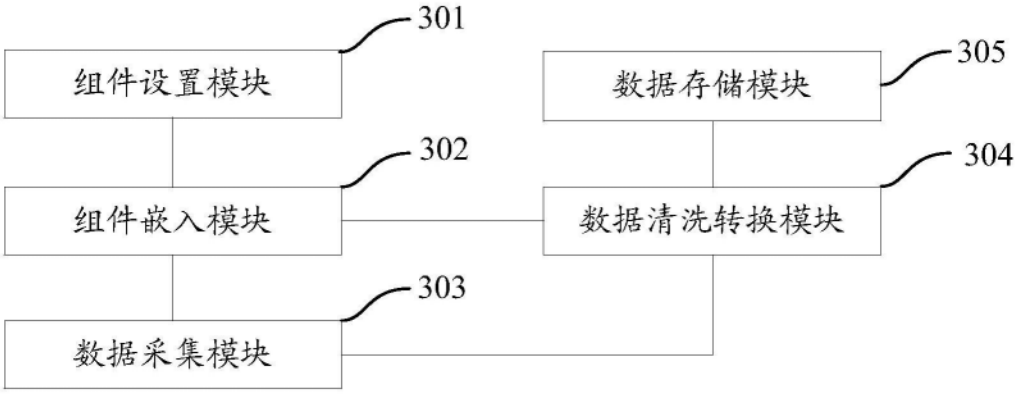
301
组件设置模块
305
数据存储模块
302
304
组件嵌入模块
数据清洗转换模块
303
数据采集模块

图3

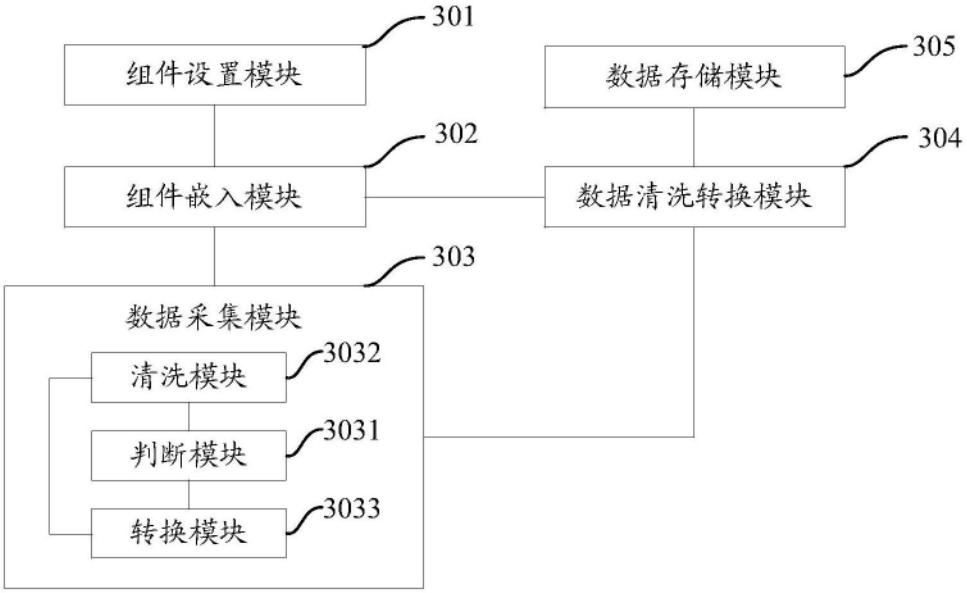
301
组件设置模块
305
数据存储模块
302
304
组件嵌入模块
数据清洗转换模块
303
数据采集模块
3032
清洗模块
3031
判断模块
3033
转换模块

图4