(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 119202051 A
(43)申请公布日 2024.12.27

(21)申请号 202411245616.5

(22)申请日 2024.09.05

(71)申请人 中煤科工开采研究院有限公司
地址 101399 北京市顺义区中关村科技园区顺义园临空二路1号

(72)发明人 吕依濛 郭永欣 牟振栋

(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
专利代理师 张星

(51)Int.Cl.
G06F 16/25(2019.01)
G06F 16/22(2019.01)
G06F 16/21(2019.01)

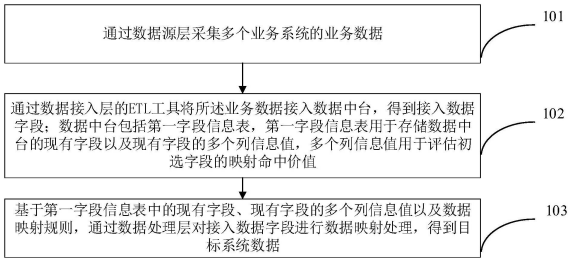
权利要求书3页 说明书9页 附图2页

(54)发明名称

基于数据中台的接口数据自动映射方法以及装置

(57)摘要

本申请提出一种基于数据中台的接口数据自动映射方法以及装置,其中,方法包括:通过数据源层采集多个业务系统的业务数据;通过数据接入层的ETL工具将业务数据接入数据中台,得到接入数据字段;数据中台包括第一字段信息表,所述第一字段信息表用于存储所述数据中台的现有字段以及现有字段的多个列信息值;基于第一字段信息表中的现有字段、现有字段的多个列信息值以及数据映射规则,通过数据处理层对接入数据字段进行数据映射处理,得到目标系统数据。解决现有技术中数据无法实现自动化映射的技术问题。



CN 119202051 A

1.一种基于数据中台的接口数据自动映射方法,其特征在于,包括以下步骤:

通过数据源层采集多个业务系统的业务数据;

通过数据接入层的ETL工具将所述业务数据接入数据中台,得到接入数据字段;所述数据中台包括第一字段信息表,所述第一字段信息表用于存储所述数据中台的现有字段以及所述现有字段的多个列信息值,所述多个列信息值用于评估现有字段的映射命中价值;

基于所述第一字段信息表中的现有字段、现有字段的多个列信息值以及数据映射规则,通过数据处理层对所述接入数据字段进行数据映射处理,得到目标系统数据。

2.根据权利要求1所述的方法,其特征在于,所述基于所述第一字段信息表中的现有字段、现有字段的多个列信息值以及数据映射规则,通过数据处理层对所述接入数据字段进行数据映射处理,得到目标系统数据:

从所述第一字段信息表中的现有字段中,获取与所述接入数据字段相似度满足预设相似度阈值的多个初选字段;

对于所述多个初选字段中的各初选字段,基于所述初选字段的多个列信息值,计算得到多个列信息参数值;

基于所述多个列信息参数值与对应各列信息参数值的预设阈值,得到该初选字段的映射匹配值;

基于所述多个初选字段中各初选字段对应的映射匹配值,从所述多个初选字段中确定与所述接入数据字段匹配的目标映射字段;

将所述目标映射字段作为所述接入数据字段的目标系统数据。

3.根据权利要求2所述的方法,其特征在于,所述从所述第一字段信息表中的现有字段中,获取与所述接入数据字段相似度满足预设相似度阈值的多个初选字段;包括:

通过距离相似度算法计算所述接入数据字段与所述第一字段信息表中各现有字段的相似度,获取与所述接入数据字段相似度满足预设相似度阈值的多个初选字段;

获取所述多个初选字段中各初选字段与所述接入数据字段的距离值L。

4.根据权利要求3所述的方法,其特征在于,所述多个列信息值包括:

字段名,所述字段名用于唯一标识字段;

存储时间,所述存储时间用于记录字段首次抽取到所述数据中台的时间;

读取量,所述读取量用于记录所述数据中台对于字段的读取次数;

读取时间,所述读取时间用于记录字段最近一次读取时的时间;

修改量,所述修改量用于记录所述数据中台对于字段的修改次数;

修改时间,所述修改时间用于记录字段最近一次变更时的修改时间;

人工纠错量,所述人工纠错量用于记录字段在实际联调接口时对于自动接口映射错误的字段进行人工干预的次数;

人工纠错时间,所述人工纠错时间用于记录最近一次所述人工纠错量变更时的指标修改时间;

是否高价值,所述是否高价值用于标识字段是否为高价值数据。

5.根据权利要求4所述的方法,其特征在于,所述基于所述初选字段的多个列信息值,计算得到多个列信息参数值,包括:

基于所述读取量,通过第一公式得到读取量参数y1;所述第一公式表示如下:

y1＝2arccot(read_number(x))/π

其中,read_number表示读取量,x表示初选字段;

基于所述修改量,通过第二公式得到修改量参数y2;所述第二公式表示如下:

y2＝2arctan(modify_number(x))/π

其中,modify_number表示修改量,x表示初选字段;

基于所述人工纠错量,通过第三公式得到人工纠错量参数y3;所述第三公式表示如下:

y3＝2arctan(correct_number(x))/π

其中,correct_number表示人工纠错量,x表示初选字段;

基于所述存储时间和所述读取时间,通过第四公式得到读取时间参数t1;所述第四公式表示如下:

t1＝2arccot(read_time(x)-time(x))/π

其中,read_time表示读取时间,time表示存储时间,x表示初选字段;

基于所述存储时间和所述修改时间,通过第五公式得到修改时间参数t2;所述第五公式表示如下:

t2＝2arctan(modify_time(x)-time(x))/π

其中,modify_time表示修改时间,time表示存储时间,x表示初选字段;

基于所述存储时间和所述人工纠错时间,通过第六公式得到人工纠错时间参数t3;所述第六公式表示如下:

t3＝2arctan(correct_time(x)-time(x))/π

其中,modify_time表示人工纠错时间,time表示存储时间,x表示初选字段;

基于所述是否高价值,通过第七公式得到高价值参数m;所述第七公式表示如下:

m＝is_high

其中,is_high表示是否高价值;

基于所述距离值L,通过第八公式得到距离参数s;所述第八公式表示如下:

s＝(threshold-L)/threshold

其中,threshold表示预设相似度阈值。

6.根据权利要求2或5所述的方法,其特征在于,所述基于所述多个列信息参数值与对应各列信息参数值的预设阈值,得到该初选字段的映射匹配值;包括:

获取所述多个列信息参数值与对应各列信息参数值的预设阈值的加权平均值;

将所述加权平均值作为该初选字段的映射匹配值。

7.根据权利要求1所述的方法,其特征在于,所述基于所述多个初选字段中各初选字段对应的映射匹配值,从所述多个初选字段中确定与所述接入数据字段匹配的目标映射字段;包括:

将所述多个初选字段中映射匹配值最大的初选字段确定为与所述接入数据字段匹配的目标映射字段。

8.一种基于数据中台的接口数据自动映射装置,其特征在于,包括:

数据获取模块,用于通过数据源层采集多个业务系统的业务数据;

数据接入模块,用于通过数据接入层的ETL工具将所述业务数据接入数据中台,得到接入数据字段;所述数据中台包括第一字段信息表,所述第一字段信息表用于存储所述数据

中台的现有字段以及所述现有字段的多个列信息值,所述多个列信息值用于评估现有字段的映射命中价值;

数据映射模块,用于基于所述第一字段信息表中的现有字段、现有字段的多个列信息值以及数据映射规则,通过数据处理层对所述接入数据字段进行数据映射处理,得到目标系统数据。

9.一种电子设备,其特征在于,包括:处理器,以及与所述处理器通信连接的存储器;

所述存储器存储计算机执行指令;

所述处理器执行所述存储器存储的计算机执行指令,以实现如权利要求1-7中任一项所述的方法。

10.一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中存储有计算机执行指令,所述计算机执行指令被处理器执行时用于实现如权利要求1-7中任一项所述的方法。

基于数据中台的接口数据自动映射方法以及装置

技术领域

[0001] 本申请涉及数据处理技术领域,尤其涉及一种基于数据中台的接口数据自动映射方法、装置、电子设备及存储介质。

背景技术

[0002] 煤矿行业存在一数多源、数据分散重复、数据质量有待提升,无法满足业务流程协同、各项业务数据同步的业务需求,以及随着信息化建设的迅速发展,业务系统会逐渐增多。迫切需要建立统一数据标准、打破信息壁垒、贯通数据共享、提升数据质量及提升数据应用价值的平台系统。

[0003] 数据映射是将现实世界中的数据映射到计算机系统中的过程。这个过程可以通过特定的软件或工具实现,将不同来源、不同格式和不同结构的数据整合到同一个平台上进行分析和处理。现有的数据映射方法一般是数据源端与平台端进行协商,确定统一标准的数据接口实现数据映射,由于数据源数量和种类的不断增多,数据映射的工作量比较大,效率低。由此,需要寻求一种自动化的数据映射方法,以解决上述问题。

发明内容

[0004] 本申请旨在至少在一定程度上解决相关技术中的技术问题之一。

[0005] 为此,本申请的第一个目的在于提出一种基于数据中台的接口数据自动映射方法,以实现接口数据的自动映射,解决相关技术中无法实现自动化的数据映射的问题。

[0006] 本申请的第二个目的在于提出一种基于数据中台的接口数据自动映射装置。

[0007] 本申请的第三个目的在于提出一种电子设备。

[0008] 本申请的第四个目的在于提出一种计算机可读存储介质。

[0009] 本申请的第五个目的在于提出一种计算机程序产品。

[0010] 为达上述目的,本申请第一方面实施例提出了一种基于数据中台的接口数据自动映射方法,包括:

[0011] 通过数据源层采集多个业务系统的业务数据;

[0012] 通过数据接入层的ETL工具将所述业务数据接入数据中台,得到接入数据字段;所述数据中台包括第一字段信息表,所述第一字段信息表用于存储所述数据中台的现有字段以及所述现有字段的多个列信息值,所述多个列信息值用于评估现有字段的映射命中价值;

[0013] 基于所述第一字段信息表中的现有字段、现有字段的多个列信息值以及数据映射规则,通过数据处理层对所述接入数据字段进行数据映射处理,得到目标系统数据。

[0014] 为达上述目的,本申请第二方面实施例提出了一种基于数据中台的接口数据自动映射装置,包括:

[0015] 数据获取模块,用于通过数据源层采集多个业务系统的业务数据;

[0016] 数据接入模块,用于通过数据接入层的ETL工具将所述业务数据接入数据中台,得

到接入数据字段;所述数据中台包括第一字段信息表,所述第一字段信息表用于存储所述数据中台的现有字段以及所述现有字段的多个列信息值,所述多个列信息值用于评估现有字段的映射命中价值;

[0017] 数据映射模块,用于基于所述第一字段信息表中的现有字段、现有字段的多个列信息值以及数据映射规则,通过数据处理层对所述接入数据字段进行数据映射处理,得到目标系统数据。

[0018] 为达上述目的,本申请第三方面实施例提出了一种电子设备,包括:处理器,以及与所述处理器通信连接的存储器;所述存储器存储计算机执行指令;所述处理器执行所述存储器存储的计算机执行指令,以实现第一方面所述的方法。

[0019] 为达上述目的,本申请第四方面实施例提出了一种计算机可读存储介质,所述计算机可读存储介质中存储有计算机执行指令,所述计算机执行指令被处理器执行时用于实现第一方面所述的方法。

[0020] 为达上述目的,本申请第五方面实施例提出了一种计算机程序产品,包括计算机程序,该计算机程序被处理器执行时实现第一方面所述的方法。

[0021] 本申请提供的基于数据中台的接口数据自动映射方法、装置、电子设备及存储介质:在获取接入数据字段之后,基于第一字段信息表中的现有字段、现有字段的多个列信息值以及数据映射规则,对接入数据字段进行数据映射得到目标系统数据,实现将源系统数据自动映射为目标系统数据,以提升数据接入效率。

[0022] 本申请附加的方面和优点将在下面的描述中部分给出,部分将从下面的描述中变得明显,或通过本申请的实践了解到。

附图说明

[0023] 本申请上述的和/或附加的方面和优点从下面结合附图对实施例的描述中将变得明显和容易理解,其中:

[0024] 图1为本申请实施例所提供的一种基于数据中台的接口数据自动映射方法的流程示意图;

[0025] 图2为本申请实施例所提供的另一种基于数据中台的接口数据自动映射方法的流程示意图;

[0026] 图3为本申请实施例所提供的一种基于数据中台的接口数据自动映射装置的框图;

[0027] 图4为本申请实施例所提供的一种电子设备的框图。

具体实施方式

[0028] 下面详细描述本申请的实施例,所述实施例的示例在附图中示出,其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的,旨在用于解释本申请,而不能理解为对本申请的限制。

[0029] 术语解释:

[0030] ETL,是英文Extract-Transform-Load的缩写,用来描述将数据从来源端经过抽取(extract)、转换(transform)、加载(load)至目的端的过程。

[0031] 下面参考附图描述本申请实施例的基于数据中台的接口数据自动映射方法、装置及设备。

[0032] 图1为本申请实施例所提供的一种基于数据中台的接口数据自动映射方法的流程示意图。

[0033] 需要说明的是,本申请实施例的基于数据中台的接口数据自动映射方法的执行主体为本申请实施例的基于数据中台的接口数据自动映射装置,该基于数据中台的接口数据自动映射装置可被配置于电子设备中,以使该电子设备可以执行基于数据中台的接口数据自动映射功能。

[0034] 如图1所示,该基于数据中台的接口数据自动映射方法包括以下步骤:

[0035] 步骤101,通过数据源层采集多个业务系统的业务数据。

[0036] 需要说明的是,基于数据中台的跨系统接口可以有效地整合多个业务系统的数据,实现数据的自动流转和共享,对于数据接口流经的系统,分为:

[0037] 1、源系统,数据接口服务的发起方。

[0038] 2、数据中台系统,数据接口服务的处理方。

[0039] 3、目标系统,数据接口服务的接收方。

[0040] 作为一种实现方式,通过接口标准化,对于接入数据中台的数据接口,统一按照统一的API规范,使用RESTful API描述接口规范,生成API文档和接口代码模板。

[0041] 在本实施例中,从数据中台架构角度,数据中台系统的平台处理数据接口服务包括数据源层、数据接入层、数据处理层、数据服务层和应用层,其中:

[0042] 数据源层用于采集包含多个业务系统的数据源,包括结构化数据和非结构化数据,如数据库、文件系统等。此模块完成接口对接系统数据源的对接。

[0043] 步骤102,通过数据接入层的ETL工具将所述业务数据接入数据中台,得到接入数据字段;数据中台包括第一字段信息表,第一字段信息表用于存储数据中台的现有字段以及现有字段的多个列信息值,多个列信息值用于评估现有字段的映射命中价值。

[0044] 在一些实施例中,数据接入层用于通过ETL工具将不同系统的数据接入数据中台,解决数仓由原始数据向报表数据逐步转化加工的问题。提供了拖曳式的图形化操作界面,对于传统的脚本结构提供了新的层次化、可视化的表达方式,能够更为清晰的显示数据加工的脉络关系,真正实现“零”编码,消除手工编码,降低出错概率,降低用工成本。此模块具体完成接口对接系统数据的抽取、转换和加载。

[0045] 数据处理层用于对接入的数据进行清洗、转换、映射等处理,其中,数据映射处理包括定义数据映射规则,将不同业务系统的数据格式转换为数据中台的标准格式,将源系统数据根据数据映射规则翻译成目标系统的数据。

[0046] 数据服务层用于提供统一的数据访问接口。

[0047] 应用层用于各业务系统通过数据中台提供的接口实现数据共享和交互。

[0048] 通过数据中台实现跨系统、跨技术体系的数据交互与共享,减少重复工作,打通数据孤岛。通过数据中台交互接口部分的复用降低数据交互难度、规范数据标准、提高数据质量、加强敏感数据的统一管控,解决不同系统之间数据格式不一致的问题,实现数据的自动转换和对接。

[0049] 在一些实施例中,接入各业务系统的数据,得到接入数据字段;包括:通过ETL工具

将各业务系统的数据接入数据中台,得到接入数据字段。

[0050] 需要说明的是,数据中台经ETL处理后的字段,会在数据中台中存储该字段相应的一些列信息值,这些列信息值用于后续步骤计算接入数据字段与数据中台中的现有字段的相关性,从而实现自动的字段映射,列信息值的具体内容将在后面步骤进行描述。

[0051] 步骤103,基于第一字段信息表中的现有字段、现有字段的多个列信息值以及数据映射规则,通过数据处理层对接入数据字段进行数据映射处理,得到目标系统数据。

[0052] 在一些实施例中,基于所述第一字段信息表中的现有字段、现有字段的多个列信息值以及数据映射规则,通过数据处理层对所述接入数据字段进行数据映射处理,得到目标系统数据的实现方式,包括如下步骤:

[0053] 步骤201,从所述第一字段信息表中的现有字段中,获取与所述接入数据字段相似度满足预设相似度阈值的多个初选字段。

[0054] 在一些实施例中,获取与接入数据字段相似度满足预设相似度阈值的多个初选字段的方法;包括:通过距离相似度算法计算接入数据字段与第一字段信息表中各现有字段的相似度,获取与接入数据字段相似度满足预设相似度阈值的多个初选字段;获取多个初选字段中各初选字段与接入数据字段的距离值L。

[0055] 示例性的,利用Levenshtein距离相似度算法,计算第一字段信息表中现有字段中与接入数据字段相似的多个初选字段。

[0056] 示例性的,设定预设相似度阈值threshold,如果通过距离相似度算法得到的距离小于或等于预定义的threshold,则认为当前的两个字段名相似,并将该距离值记为L。

[0057] 本步骤的目的是通过相似判断,快速匹配与接入数据字段相似的多个初选字段,以便后续进一步筛选,得到目标映射字段。

[0058] 步骤202,对于多个初选字段中的各初选字段,基于初选字段的多个列信息值,计算得到多个列信息参数值。

[0059] 多个列信息值用于评估现有字段的映射命中价值

[0060] 在一些实施例中,多个列信息值包括字段名、存储时间、读取量、读取时间、修改量、修改时间、人工纠错量、人工纠错时间和是否高价值,其中:

[0061] 字段名(name_id),字段名用于唯一标识字段,记录字段id,为该字段在数据中台中的唯一id标识;

[0062] 存储时间(time),存储时间用于记录字段首次抽取到数据中台的时间;

[0063] 读取量(read_number),读取量用于记录数据中台对于字段的读取次数,默认值为0,每读取一次,读取值+1,该数值只增不减;

[0064] 读取时间(read_time),读取时间用于记录字段最近一次读取时的时间;

[0065] 修改量(modify_number),修改量用于记录数据中台对于字段的修改次数,默认值为0,每读取一次,读取值+1,该数值只增不减;

[0066] 修改时间(modify_time),修改时间用于记录字段最近一次变更时的修改时间;

[0067] 人工纠错量(correct_number),人工纠错量用于记录字段在实际联调接口时对于自动接口映射错误的字段进行人工干预的次数,默认值为0,每修改一次,读取值+1,该数值只增不减;

[0068] 人工纠错时间(correct_time),人工纠错时间用于记录最近一次人工纠错量变更

时的指标修改时间;

[0069] 是否高价值(is_high),是否高价值用于标识字段是否为高价值数据,0表示低价值字段,1表示高价值字段,默认为0。

[0070] 本实施例中,多个列信息参数值包括读取量参数y1、修改量参数y2、人工纠错量参数y3、读取时间参数t1、修改时间参数t2、人工纠错时间参数t3、高价值参数m和距离参数s。

[0071] 例如,一个相似字段为x,则多个列信息参数值的计算方法分别如下:

[0072] 读取量参数y1:2arccot(read_number(x))/π;

[0073] 修改量参数y2:2arctan(modify_number(x))/π;

[0074] 人工纠错量参数y3:2arctan(correct_number(x))/π;

[0075] 读取时间参数t1:2arccot(read_time(x)-time(x))/π;

[0076] 修改时间参数t2:2arctan(modify_time(x)-time(x))/π;

[0077] 人工纠错时间参数t3:2arctan(correct_time(x)-time(x))/π;

[0078] 高价值参数m:is_high;

[0079] 距离参数s:(threshold-L)/threshold。

[0080] 通过以上公式,可以得到各初选字段的多个列信息参数值,从而在后续步骤计算各初选字段的映射匹配值。

[0081] 步骤203,基于多个列信息参数值与对应各列信息参数值的预设阈值,得到该初选字段的映射匹配值。

[0082] 在本一些实施例中,上述多个列信息参数值的预设阈值分别为p1,p2,p3,p4,p5,p6,p7,p8。

[0083] 需要说明的是,多个列信息参数值的预设阈值设置有默认值,且是能够随时进行调整的。

[0084] 在一些实施例中,基于多个列信息参数值与对应各列信息参数值的预设阈值,得到该初选字段的映射匹配值;包括:获取多个列信息参数值与对应各列信息参数值的预设阈值的加权平均值;将加权平均值作为该初选字段的映射匹配值。

[0085] 也就是说,通过如下公式计算各初选字段的映射匹配值,该公式表示如下:

[0086] V＝(p1*y1+p2*y2+p3*y3+p4*t1+p5*t2+p6*t3+p7*m+p8*L)/(p1+p2+p3+p4+p5+p6+p7+p8)

[0087] 其中,V为取值在(0,1]的值。

[0088] 由此,通过计算各初选字段的V值,得到初选字段的映射匹配值。

[0089] 步骤204,基于多个初选字段中各初选字段对应的映射匹配值,从多个初选字段中确定与接入数据字段匹配的目标映射字段。

[0090] 在一些实现方式中,基于多个初选字段中各初选字段对应的映射匹配值,从多个初选字段中确定与接入数据字段匹配的目标映射字段;包括:将多个初选字段中映射匹配值最大的初选字段确定为与接入数据字段匹配的目标映射字段。

[0091] 可以理解为,几个初选字段即相似字段的V值越大,表示映射命中价值越高,由此,可以将V值最大的字段,作为目标映射字段。

[0092] 需要说明的是,在获取目标映射字段之后,还设有人工纠错环节,若发现映射错误,可以进行人工纠错,从而修改人工纠错量,以提升数据映射的准确性。

[0093] 本申请实施例的基于数据中台的接口数据自动映射方法,在获取接入数据字段之后,基于第一字段信息表中的现有字段、现有字段的多个列信息值以及数据映射规则,对接入数据字段进行数据映射得到目标系统数据,实现将源系统数据自动映射为目标系统数据,以提升数据接入效率。进一步的,通过字段之间的相似度获取多个初选字段,再通过初选字段的多个列信息值计算该初选字段的映射命中价值,从而进一步从多个初选字段中确定目标映射字段,以提高数据映射的精确度和效率。

[0094] 为了实现上述实施例,本申请还提出一种基于数据中台的接口数据自动映射装置。图3为本申请实施例提供的一种基于数据中台的接口数据自动映射装置的框图。如图3所示,该基于数据中台的接口数据自动映射装置可以包括:数据获取模块301、数据接入模块302和数据映射模块303。

[0095] 其中,数据获取模块301,用于通过数据源层采集多个业务系统的业务数据;

[0096] 数据接入模块302,用于通过数据接入层的ETL工具将所述业务数据接入数据中台,得到接入数据字段;所述数据中台包括第一字段信息表,所述第一字段信息表用于存储所述数据中台的现有字段以及所述现有字段的多个列信息值,所述多个列信息值用于评估现有字段的映射命中价值;

[0097] 数据映射模块303,用于基于所述第一字段信息表中的现有字段、现有字段的多个列信息值以及数据映射规则,通过数据处理层对所述接入数据字段进行数据映射处理,得到目标系统数据。

[0098] 在一些实现方式中,数据映射模块303具体用于:

[0099] 从所述第一字段信息表中的现有字段中,获取与所述接入数据字段相似度满足预设相似度阈值的多个初选字段;

[0100] 对于所述多个初选字段中的各初选字段,基于所述初选字段的多个列信息值,计算得到多个列信息参数值;

[0101] 基于所述多个列信息参数值与对应各列信息参数值的预设阈值,得到该初选字段的映射匹配值;;

[0102] 基于所述多个初选字段中各初选字段对应的映射匹配值,从所述多个初选字段中确定与所述接入数据字段匹配的目标映射字段;

[0103] 将所述目标映射字段作为所述接入数据字段的目标系统数据。

[0104] 在一些实现方式中,数据映射模块303在从所述第一字段信息表中的现有字段中,获取与所述接入数据字段相似度满足预设相似度阈值的多个初选字段时,用于:

[0105] 通过距离相似度算法计算所述接入数据字段与所述第一字段信息表中各现有字段的相似度,获取与所述接入数据字段相似度满足预设相似度阈值的多个初选字段;

[0106] 获取所述多个初选字段中各初选字段与所述接入数据字段的距离值L。

[0107] 在一些实现方式中,多个列信息值包括:

[0108] 字段名,字段名用于唯一标识接入数据字段;

[0109] 存储时间,存储时间用于记录接入数据字段首次抽取到数据中台的时间;

[0110] 读取量,读取量用于记录数据中台对于接入数据字段的读取次数;

[0111] 读取时间,读取时间用于记录接入数据字段最近一次读取时的时间;

[0112] 修改量,修改量用于记录数据中台对于接入数据字段的修改次数;

[0113] 修改时间,修改时间用于记录接入数据字段最近一次变更时的修改时间;

[0114] 人工纠错量,人工纠错量用于记录接入数据字段在实际联调接口时对于自动接口映射错误的字段进行人工干预的次数;

[0115] 人工纠错时间,人工纠错时间用于记录最近一次人工纠错量变更时的指标修改时间;

[0116] 是否高价值,是否高价值用于标识接入数据字段是否为高价值数据。

[0117] 在一些实现方式中,数据映射模块303在基于初选字段的多个列信息值,计算得到多个列信息参数值时,用于:

[0118] 基于读取量,通过第一公式得到读取量参数y1;第一公式表示如下:

[0119] $y1=2\operatorname{arccot}(read\_number(x))/\pi$

[0120] 其中,read_number表示读取量,x表示初选字段;

[0121] 基于修改量,通过第二公式得到修改量参数y2;第二公式表示如下:

[0122] $y2=2\arctan(modify\_number(x))/\pi$

[0123] 其中,modify_number表示修改量,x表示初选字段;

[0124] 基于人工纠错量,通过第三公式得到人工纠错量参数y3;第三公式表示如下:

[0125] $y3=2\arctan(correct\_number(x))/\pi$

[0126] 其中,correct_number表示人工纠错量,x表示初选字段;

[0127] 基于存储时间和读取时间,通过第四公式得到读取时间参数t1;第四公式表示如下:

[0128] $t1=2\operatorname{arccot}(read\_time(x)-time(x))/\pi$

[0129] 其中,read_time表示读取时间,time表示存储时间,x表示初选字段;

[0130] 基于存储时间和修改时间,通过第五公式得到修改时间参数t2;第五公式表示如下:

[0131] $t2=2\arctan(modify\_time(x)-time(x))/\pi$

[0132] 其中,modify_time表示修改时间,time表示存储时间,x表示初选字段;

[0133] 基于存储时间和人工纠错时间,通过第六公式得到人工纠错时间参数t3;第六公式表示如下:

[0134] $t3=2\arctan(correct\_time(x)-time(x))/\pi$

[0135] 其中,modify_time表示人工纠错时间,time表示存储时间,x表示初选字段;

[0136] 基于是否高价值,通过第七公式得到高价值参数m;第八公式表示如下:

[0137] $m=is\_high$

[0138] 其中,is_high表示是否高价值;

[0139] 基于距离值L,通过第八公式得到距离参数s;第八公式表示如下:

[0140] $s=(threshold-L)/threshold$

[0141] 其中,threshold表示预设相似度阈值。

[0142] 在一些实现方式中,数据映射模块303在基于所述多个列信息参数值与对应各列信息参数值的预设阈值,得到该初选字段的映射匹配值时,具体用于:

[0143] 获取所述多个列信息参数值与对应各列信息参数值的预设阈值的加权平均值;

[0144] 将所述加权平均值作为该初选字段的映射匹配值。

[0145] 在一些实现方式中,数据映射模块303在基于所述多个初选字段中各初选字段对应的映射匹配值,从所述多个初选字段中确定与所述接入数据字段匹配的目标映射字段时,用于:

[0146] 将多个初选字段中映射匹配值最大的初选字段确定为与接入数据字段匹配的目标映射字段。

[0147] 需要说明的是,前述对基于数据中台的接口数据自动映射方法实施例的解释说明也适用于该实施例的基于数据中台的接口数据自动映射装置,此处不再赘述。

[0148] 为了实现上述实施例,本申请还提出一种电子设备。请参见图4,图4是本申请实施例提供的电子设备的框图。如图4所示,电子设备400包括:处理器401,以及与处理器401通信连接的存储器402;存储器402存储计算机执行指令;处理器401执行存储器存储的计算机执行指令,以实现执行前述实施例所提供的方法。

[0149] 为了实现上述实施例,本申请还提出一种计算机可读存储介质,计算机可读存储介质中存储有计算机执行指令,所述计算机执行指令被处理器执行时用于实现前述实施例所提供的方法。

[0150] 为了实现上述实施例,本申请还提出一种计算机程序产品,包括计算机程序,该计算机程序被处理器执行时实现前述实施例所提供的方法。

[0151] 本申请中所涉及的用户个人信息的收集、存储、使用、加工、传输、提供和公开等处理,均符合相关法律法规的规定,且不违背公序良俗。

[0152] 需要说明的是,来自用户的个人信息应当被收集用于合法且合理的用途,并且不在这些合法使用之外共享或出售。此外,应在收到用户知情同意后进行此类采集/共享,包括但不限于在用户使用该功能前,通知用户阅读用户协议/用户通知,并签署包括授权相关用户信息的协议/授权。此外,还需采取任何必要步骤,保卫和保障对此类个人信息数据的访问,并确保有权访问个人信息数据的其他人遵守其隐私政策和流程。

[0153] 本申请预期可提供用户选择性阻止使用或访问个人信息数据的实施方案。即本公开预期可提供硬件和/或软件,以防止或阻止对此类个人信息数据的访问。一旦不再需要个人信息数据,通过限制数据收集和删除数据可最小化风险。此外,在适用时,对此类个人信息去除个人标识,以保护用户的隐私。

[0154] 在前述各实施例描述中,参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本申请的至少一个实施例或示例中。在本说明书中,对上述术语的示意性表述不必须针对的是相同的实施例或示例。而且,描述的具体特征、结构、材料或者特点可以在任一个或多个实施例或示例中以合适的方式结合。此外,在不相互矛盾的情况下,本领域的技术人员可以将本说明书中描述的不同实施例或示例以及不同实施例或示例的特征进行结合和组合。

[0155] 此外,术语“第一”、“第二”仅用于描述目的,而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此,限定有“第一”、“第二”的特征可以明示或者隐含地包括至少一个该特征。在本申请的描述中,“多个”的含义是至少两个,例如两个,三个等,除非另有明确具体的限定。

[0156] 流程图中或在此以其他方式描述的任何过程或方法描述可以被理解为,表示包括

一个或更多个用于实现定制逻辑功能或过程的步骤的可执行指令的代码的模块、片段或部分,并且本申请的优选实施方式的范围包括另外的实现,其中可以不按所示出或讨论的顺序,包括根据所涉及的功能按基本同时的方式或按相反的顺序,来执行功能,这应被本申请的实施例所属技术领域的技术人员所理解。

[0157] 在流程图中表示或在此以其他方式描述的逻辑和/或步骤,例如,可以被认为是用于实现逻辑功能的可执行指令的定序列表,可以具体实现在任何计算机可读介质中,以供指令执行系统、装置或设备(如基于计算机的系统、包括处理器的系统或其他可以从指令执行系统、装置或设备取指令并执行指令的系统)使用,或结合这些指令执行系统、装置或设备而使用。就本说明书而言,"计算机可读介质"可以是任何可以包含、存储、通信、传播或传输程序以供指令执行系统、装置或设备或结合这些指令执行系统、装置或设备而使用的装置。计算机可读介质的更具体的示例(非穷尽性列表)包括以下:具有一个或多个布线的电连接部(电子装置),便携式计算机盘盒(磁装置),随机存取存储器(RAM),只读存储器(ROM),可擦除可编辑只读存储器(EPROM或闪速存储器),光纤装置,以及便携式光盘只读存储器(CDROM)。另外,计算机可读介质甚至可以是可在其上打印所述程序的纸或其他合适的介质,因为可以例如通过对纸或其他介质进行光学扫描,接着进行编辑、解译或必要时以其他合适方式进行处理来以电子方式获得所述程序,然后将其存储在计算机存储器中。

[0158] 应当理解,本申请的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中,多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。如,如果用硬件来实现和在另一实施方式中一样,可用本领域公知的下列技术中的任一项或他们的组合来实现:具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路,具有合适的组合逻辑门电路的专用集成电路,可编程门阵列(PGA),现场可编程门阵列(FPGA)等。

[0159] 本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通过程序来指令相关的硬件完成,所述的程序可以存储于一种计算机可读存储介质中,该程序在执行时,包括方法实施例的步骤之一或其组合。

[0160] 此外,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模块中,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模块中。上述集成的模块既可以采用硬件的形式实现,也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时,也可以存储在一个计算机可读取存储介质中。

[0161] 上述提到的存储介质可以是只读存储器,磁盘或光盘等。尽管上面已经示出和描述了本申请的实施例,可以理解的是,上述实施例是示例性的,不能理解为对本申请的限制,本领域的普通技术人员在本申请的范围内可以对上述实施例进行变化、修改、替换和变型。

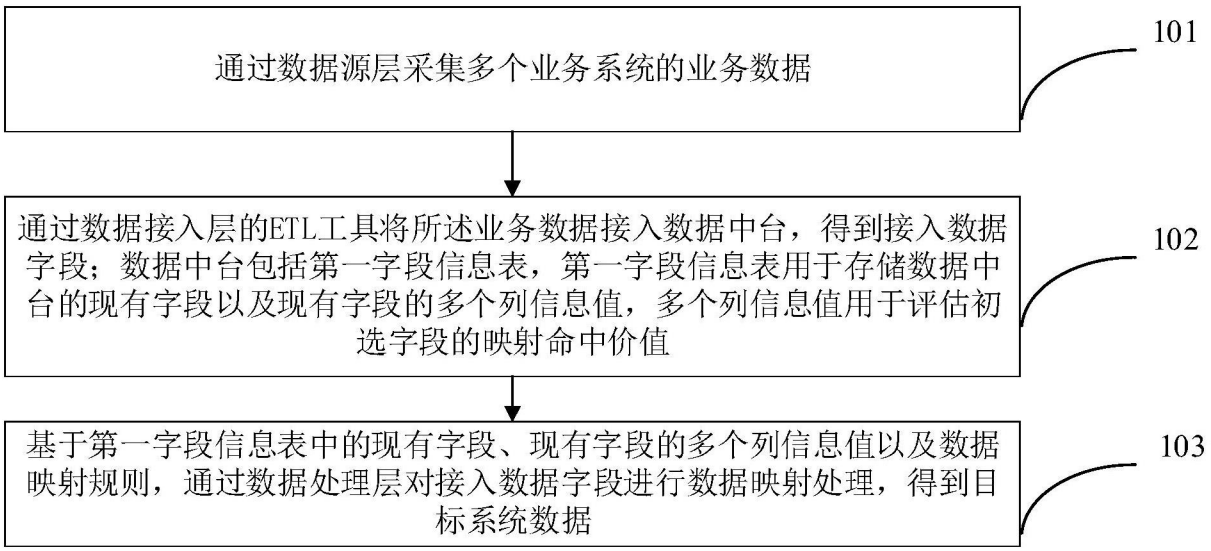

通过数据源层采集多个业务系统的业务数据
101
通过数据接入层的ETL工具将所述业务数据接入数据中台，得到接入数据字段；数据中台包括第一字段信息表，第一字段信息表用于存储数据中台的现有字段以及现有字段的多个列信息值，多个列信息值用于评估初选字段的映射命中价值
102
基于第一字段信息表中的现有字段、现有字段的多个列信息值以及数据映射规则，通过数据处理层对接入数据字段进行数据映射处理，得到目标系统数据
103

图1

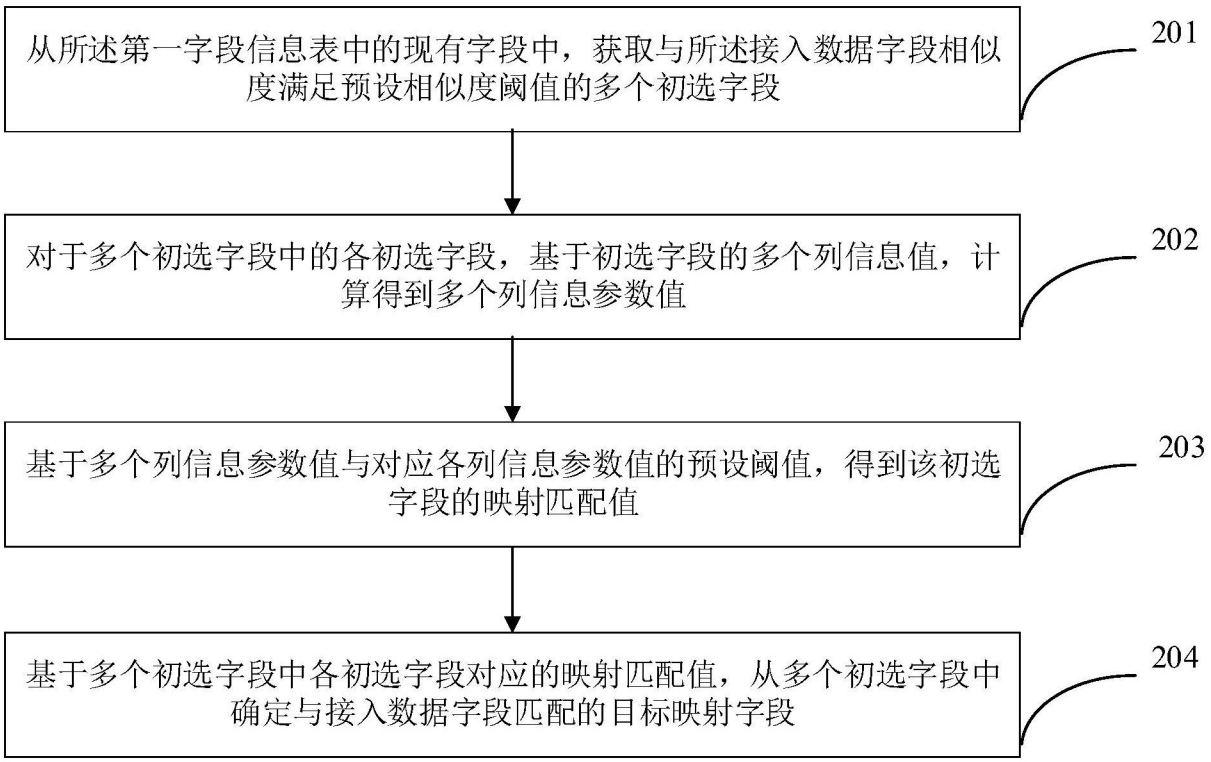

从所述第一字段信息表中的现有字段中，获取与所述接入数据字段相似度满足预设相似度阈值的多个初选字段
201
对于多个初选字段中的各初选字段，基于初选字段的多个列信息值，计算得到多个列信息参数值
202
基于多个列信息参数值与对应各列信息参数值的预设阈值，得到该初选字段的映射匹配值
203
基于多个初选字段中各初选字段对应的映射匹配值，从多个初选字段中确定与接入数据字段匹配的目标映射字段
204

图2

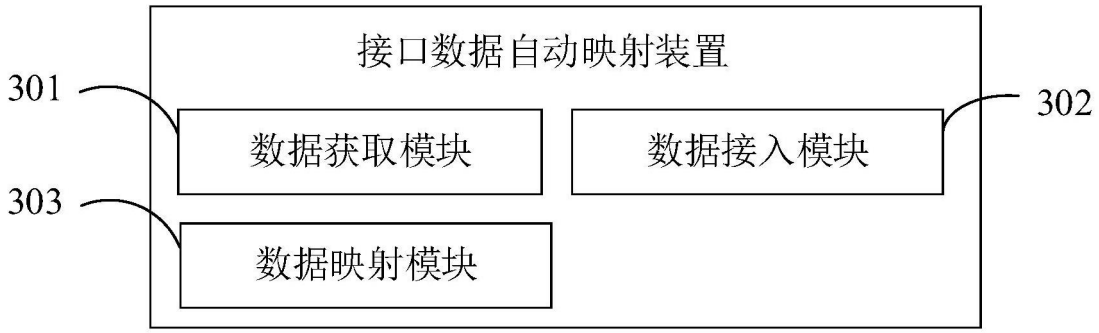

接口数据自动映射装置
301
302
303
数据获取模块
数据接入模块
数据映射模块

图3


电子设备
401
402
400
处理器
存储器

图4