(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110914818 B
(45)授权公告日 2024.04.12

(21)申请号 201880037759.3
(22)申请日 2018.06.05
(65)同一申请的已公布的文献号
申请公布号 CN 110914818 A
(43)申请公布日 2020.03.24
(30)优先权数据
62/516,396 2017.06.07 US
15/997,278 2018.06.04 US
(85)PCT国际申请进入国家阶段日
2019.12.06
(86)PCT国际申请的申请数据
PCT/US2018/035976 2018.06.05
(87)PCT国际申请的公布数据
WO2018/226637 EN 2018.12.13

(73)专利权人 起元技术有限责任公司
地址 美国马萨诸塞州
(72)发明人 R·A·爱泼斯坦 M·帕尔默
(74)专利代理机构 北京林达刘知识产权代理事务所(普通合伙) 11277
专利代理师 刘新宇

(51)Int.Cl.
G06F 16/25(2019.01)
G06F 9/445(2018.01)
G06F 9/451(2018.01)
G06F 8/34(2018.01)
(56)对比文件
JP 2017500646 A,2017.01.05
US 2015242093 A1,2015.08.27
审查员 徐霞

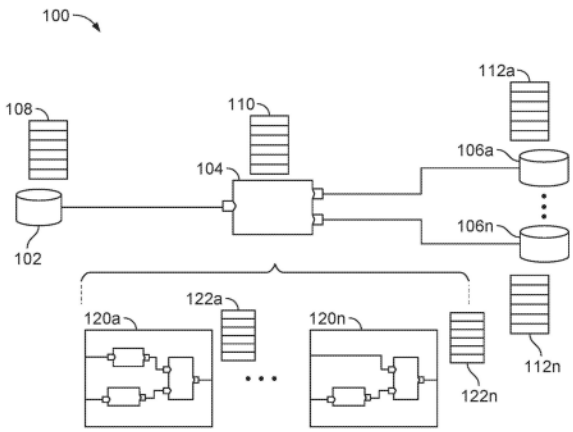
权利要求书5页 说明书9页 附图15页

(54)发明名称

数据流图配置

(57)摘要

一种用于配置第一计算机可执行程序的方法,所述方法包括通过用户界面接收指示数据源和数据目标的信息;通过所述用户界面,接收过程的特性,所述特性包括所述过程的类型和与所述过程相关联的相应特性的值。所述方法包括基于所接收到的信息,自动地为所述第一计算机可执行程序的相应参数分配值,以使所述第一计算机可执行程序在被执行时从所述数据源接收数据并将所述数据输出到所述数据目标。所述方法包括基于所接收到的所述过程的特性,自动地配置所述第一计算机可执行程序以引用第二计算机可执行程序。所述配置包括基于所述过程的类型标识所述第二计算机可执行程序;以及基于与所述过程相关联的所述相应特性的值,为所述第二计算机可执行程序的相应参数分配值。



CN 110914818 B

1.一种用于自动配置第一计算机可执行程序的方法,所述方法包括:

通过用户界面,接收指示数据源和数据目标的信息;

通过所述用户界面,接收过程的特性,所述特性包括所述过程的类型和与所述过程相关联的相应特性的值;

基于所接收到的指示所述数据源和所述数据目标的信息,自动地为所述第一计算机可执行程序的相应参数分配值,以使所述第一计算机可执行程序在被执行时从所述数据源接收数据并将数据输出到所述数据目标,其中,分配至所述第一计算机可执行程序的参数的值定义所述第一计算机可执行程序在被执行时的行为;以及

基于所接收到的所述过程的特性,自动地配置所述第一计算机可执行程序以引用第二计算机可执行程序,所述配置包括:

基于所述过程的类型标识所述第二计算机可执行程序;以及

基于与所述过程相关联的所述相应特性的值,为所述第二计算机可执行程序的相应参数分配值,其中,分配至所述第二计算机可执行程序的参数的值定义所述第二计算机可执行程序在被执行时的行为。

2.如权利要求1所述的方法,其中,所述第一计算机可执行程序包括第一数据流图,所述第一数据流图包括表示数据处理组件的一个或多个节点和表示所述数据处理组件之间的数据流的一个或多个链路。

3.如权利要求2所述的方法,其中,所述第二计算机可执行程序包括第二数据流图,所述第二数据流图包括表示数据处理组件的一个或多个节点和表示所述数据处理组件之间的数据流的一个或多个链路。

4.如权利要求1至3中任一项所述的方法,其中,指示所述数据源的所述信息标识所述数据源的类型。

5.如权利要求4所述的方法,其中,接收指示所述数据源的所述信息包括:接收从一个或多个类型的集中对所述数据源的类型的选择。

6.如权利要求5所述的方法,其中,所述一个或多个类型的集包括文件、表、HDFS数据集、HIVE表、和队列中的一个或多个。

7.如权利要求1至3中任一项所述的方法,其中,指示所述数据源的所述信息包括指示所述数据源的位置的信息。

8.如权利要求1至3中任一项所述的方法,其中,指示所述数据目标的所述信息标识所述数据目标的类型。

9.如权利要求8所述的方法,其中,接收指示所述数据目标的所述信息包括:接收从一个或多个类型的集中对所述数据目标的类型的选择。

10.如权利要求9所述的方法,其中,所述一个或多个类型的集包括文件、表、HDFS数据集、HIVE表、和队列中的一个或多个。

11.如权利要求1至3中任一项所述的方法,其中,指示所述数据目标的所述信息包括指示多个数据目标的信息。

12.如权利要求1至3中任一项所述的方法,其中,指示所述数据目标的所述信息包括指示所述数据目标的位置的信息。

13.如权利要求1至3中任一项所述的方法,其中,所述过程的所述特性包括多个过程中

的每个过程的特性。

14.如权利要求1至3中任一项所述的方法,包括基于所述过程的类型确定与所述过程相关联的所述特性。

15.如权利要求1至3中任一项所述的方法,其中,接收过程的特性包括:接收从一个或多个过程的集中对特定过程的选择。

16.如权利要求15所述的方法,其中,所述集中的所述过程包括映射过程、过滤过程、去重过程、聚合过程、复制过程、规范化过程、划分和排序过程、改变数据捕获过程、以及合并过程中的一个或多个。

17.如权利要求3所述的方法,包括在所述第一数据流图的运行时自动地配置所述第一数据流图以引用所述第二数据流图。

18.如权利要求3所述的方法,其中,自动地配置所述第一数据流图以引用所述第二数据流图包括配置所述第一数据流图以在所述第一数据流图的执行期间执行所述第二数据流图。

19.如权利要求1所述的方法,还包括在所述用户界面中向用户显示第一配置界面,以提供用于从所述用户界面中显示的一组多个过程类型中选择过程的类型的输入,所述多个过程类型各自与相应的第二计算机可执行程序相对应,各第二计算机可执行程序能够被配置为接收数据记录、处理数据记录并输出处理后的数据记录,以及

其中,接收过程的特性包括接收从所述用户界面中显示的所述一组多个过程类型中对过程的类型的选择。

20.如权利要求19所述的方法,还包括响应于接收到对所述过程的类型的选择,在所述用户界面中显示第二配置界面,其中,所述第二配置界面用于与关联于所选择的类型的过程的一个或多个特性有关的信息,所述第二配置界面是基于所选择的类型的过程的,以及

接收过程的特性包括接收与关联于所选择的类型的过程的一个或多个特性有关的所述信息。

21.如权利要求20所述的方法,其中,基于所述过程的类型标识所述第二计算机可执行程序包括自动地标识与所选择的类型的过程相对应的特定的第二计算机可执行程序。

22.如权利要求21所述的方法,其中,为所述第二计算机可执行程序的相应参数分配值包括基于接收到的与所述过程的一个或多个特性有关的信息,自动地确定所述特定的第二计算机可执行程序的一个或多个参数中的各参数的值。

23.如权利要求22所述的方法,其中,为所述第二计算机可执行程序的相应参数分配值包括自动地分配所述特定的第二计算机可执行程序的各参数的确定值,以在接收数据记录、处理数据记录和输出处理后的数据记录的方面对所述第二计算机可执行程序进行配置,以及

当所述第一计算机可执行程序被执行时,将确定值向所述特定的第二计算机可执行程序的各参数的分配使得所述特定的第二计算机可执行程序根据由接收到的信息指示的过程的特性而做出行为。

24.一种非暂态计算机可读介质,所述非暂态计算机可读介质存储用于使计算系统自动配置第一计算机可执行程序的指令,所述指令包括用于使所述计算系统执行以下操作的指令:

通过用户界面,接收指示数据源和数据目标的信息;

通过所述用户界面,接收过程的特性,所述特性包括所述过程的类型和与所述过程相关联的相应特性的值;

基于所接收到的指示所述数据源和所述数据目标的信息,自动地为所述第一计算机可执行程序的相应参数分配值,以使所述第一计算机可执行程序在被执行时从所述数据源接收数据并将数据输出到所述数据目标,其中,分配至所述第一计算机可执行程序的参数的值定义所述第一计算机可执行程序在被执行时的行为;以及

基于所接收到的所述过程的特性,自动地配置所述第一计算机可执行程序以引用第二计算机可执行程序,所述配置包括:

基于所述过程的类型标识所述第二计算机可执行程序;以及

基于与所述过程相关联的所述相应特性的值,为所述第二计算机可执行程序的相应参数分配值,其中,分配至所述第二计算机可执行程序的参数的值定义所述第二计算机可执行程序在被执行时的行为。

25.如权利要求24所述的非暂态计算机可读介质,其中,所述第一计算机可执行程序包括第一数据流图,所述第一数据流图包括表示数据处理组件的一个或多个节点和表示所述数据处理组件之间的数据流的一个或多个链路。

26.如权利要求25所述的非暂态计算机可读介质,其中,所述第二计算机可执行程序包括第二数据流图,所述第二数据流图包括表示数据处理组件的一个或多个节点和表示所述数据处理组件之间的数据流的一个或多个链路。

27.如权利要求24所述的非暂态计算机可读介质,其中,指示所述数据源的所述信息标识所述数据源的类型。

28.如权利要求27所述的非暂态计算机可读介质,其中,接收指示所述数据源的所述信息包括:接收从一个或多个类型的集中对所述数据源的类型的选择。

29.如权利要求24所述的非暂态计算机可读介质,其中,指示所述数据目标的所述信息标识所述数据目标的类型。

30.如权利要求29所述的非暂态计算机可读介质,其中,接收指示所述数据目标的所述信息包括:接收从一个或多个类型的集中对所述数据目标的类型的选择。

31.如权利要求24所述的非暂态计算机可读介质,其中,指示所述数据目标的所述信息包括指示所述数据目标的位置的信息。

32.如权利要求24所述的非暂态计算机可读介质,其中,所述过程的所述特性包括多个过程中的每个过程的特性。

33.如权利要求24所述的非暂态计算机可读介质,其中,所述指令使所述计算系统基于所述过程的类型确定与所述过程相关联的所述特性。

34.如权利要求24所述的非暂态计算机可读介质,其中,接收过程的特性包括:接收从一个或多个过程的集中对特定过程的选择。

35.如权利要求34所述的非暂态计算机可读介质,其中,所述集中的所述过程包括映射过程、过滤过程、去重过程、聚合过程、复制过程、规范化过程、划分和排序过程、改变数据捕获过程、以及合并过程中的一个或多个。

36.如权利要求26所述的非暂态计算机可读介质,包括在所述第一数据流图的运行时

自动地配置所述第一数据流图以引用所述第二数据流图。

37.如权利要求26所述的非暂态计算机可读介质,其中,自动地配置所述第一数据流图以引用所述第二数据流图包括配置所述第一数据流图以在所述第一数据流图的执行期间执行所述第二数据流图。

38.一种用于自动配置第一计算机可执行程序的计算系统,所述计算系统包括:

存储器和一个或多个处理器,所述存储器和一个或多个处理器被配置用于:

通过用户界面,接收指示数据源和数据目标的信息;

通过所述用户界面,接收过程的特性,所述特性包括所述过程的类型和与所述过程相关联的相应特性的值;

基于所接收到的信息,自动地为所述第一计算机可执行程序的相应参数分配值,以使所述第一计算机可执行程序在被执行时从所述数据源接收数据并将数据输出到所述数据目标;以及

基于所接收到的所述过程的特性,自动地配置所述第一计算机可执行程序以引用第二计算机可执行程序,所述配置包括:

基于所述过程的类型标识所述第二计算机可执行程序;以及

基于与所述过程相关联的所述相应特性的值,为所述第二计算机可执行程序的相应参数分配值。

39.如权利要求38所述的计算系统,其中,所述第一计算机可执行程序包括第一数据流图,所述第一数据流图包括表示数据处理组件的一个或多个节点和表示所述数据处理组件之间的数据流的一个或多个链路。

40.如权利要求39所述的计算系统,其中,所述第二计算机可执行程序包括第二数据流图,所述第二数据流图包括表示数据处理组件的一个或多个节点和表示所述数据处理组件之间的数据流的一个或多个链路。

41.如权利要求38所述的计算系统,其中,指示所述数据源的所述信息标识所述数据源的类型。

42.如权利要求41所述的计算系统,其中,接收指示所述数据源的所述信息包括:接收从一个或多个类型的集中对所述数据源的类型的选择。

43.如权利要求38所述的计算系统,其中,指示所述数据目标的所述信息标识所述数据目标的类型。

44.如权利要求43所述的计算系统,其中,接收指示所述数据目标的所述信息包括:接收从一个或多个类型的集中对所述数据目标的类型的选择。

45.如权利要求38所述的计算系统,其中,指示所述数据目标的所述信息包括指示所述数据目标的位置的信息。

46.如权利要求38所述的计算系统,其中,所述过程的所述特性包括多个过程中的每个过程的特性。

47.如权利要求38所述的计算系统,其中,所述一个或多个处理器和所述存储器被配置为基于所述过程的类型确定与所述过程相关联的所述特性。

48.如权利要求38所述的计算系统,其中,接收过程的特性包括:接收从一个或多个过程的集中对特定过程的选择。

49.如权利要求48所述的计算系统,其中,所述集中的所述过程包括映射过程、过滤过程、去重过程、聚合过程、复制过程、规范化过程、划分和排序过程、改变数据捕获过程、以及合并过程中的一个或多个。

50.如权利要求40所述的计算系统,包括在所述第一数据流图的运行时自动地配置所述第一数据流图以引用所述第二数据流图。

51.如权利要求40所述的计算系统,其中,自动地配置所述第一数据流图以引用所述第二数据流图包括配置所述第一数据流图以在所述第一数据流图的执行期间执行所述第二数据流图。

52.一种用于自动配置第一计算机可执行程序的计算系统,所述计算系统包括:

用于通过用户界面接收指示数据源和数据目标的信息的装置;

用于通过用户界面接收过程的特性的装置,所述特性包括所述过程的类型和与所述过程相关联的相应特性的值;

用于基于所接收到的指示所述数据源和所述数据目标的信息自动地为所述第一计算机可执行程序的相应参数分配值以使所述第一计算机可执行程序在被执行时从所述数据源接收数据并将数据输出到所述数据目标的装置,其中,分配至所述第一计算机可执行程序的参数的值定义所述第一计算机可执行程序在被执行时的行为;以及

用于基于所接收到的所述过程的特性自动地配置所述第一计算机可执行程序以引用第二计算机可执行程序的装置,所述配置包括:

基于所述过程的类型标识所述第二计算机可执行程序;以及

基于与所述过程相关联的所述相应特性的值,为所述第二计算机可执行程序的相应参数分配值,其中,分配至所述第二计算机可执行程序的参数的值定义所述第二计算机可执行程序在被执行时的行为。

数据流图配置

[0001] 优先权要求

[0002] 本申请要求2017年6月7日提交的美国专利申请序列号62/516,396和2018年6月4日提交的美国专利申请序列号15/997,278的优先权,这两个申请的内容通过援引以其全文并入本文。

技术领域

[0003] 本申请涉及数据流图配置。

背景技术

[0004] 可以生成用户界面并将其显示给用户,以允许用户与由计算系统处理的数据进行交互。某些用户界面是应用程序的静态组件,并且对于问题的多个用户,用户界面以相同的方式运行。某些用户界面可以由应用程序的用户定义,以便针对特定目的自定义用户界面。例如,应用程序可以允许指定多个用户界面,并且用户可以从所述多个用户界面中进行选择。

发明内容

[0005] 一方面,一种用于配置第一计算机可执行程序的方法,所述方法包括通过用户界面接收指示数据源和数据目标的信息;通过所述用户界面,接收过程的特性,所述特性包括所述过程的类型和与所述过程相关联的相应特性的值。所述方法包括基于所接收到的信息,自动地为所述第一计算机可执行程序的相应参数分配值,以使所述第一计算机可执行程序在被执行时从所述数据源接收数据并将所述数据输出到所述数据目标。所述方法包括基于所接收到的所述过程的特性,自动地配置所述第一计算机可执行程序以引用第二计算机可执行程序。所述配置包括基于所述过程的类型标识所述第二计算机可执行程序;以及基于与所述过程相关联的所述相应特性的值,为所述第二计算机可执行程序的相应参数分配值。

[0006] 实施例可包括以下特征中的一个或多个。

[0007] 所述第一计算机可执行程序包括第一数据流图,所述第一数据流图包括表示数据处理组件的一个或多个节点和表示所述数据处理组件之间的数据流的一个或多个链路。

[0008] 所述第二计算机可执行程序包括第二数据流图,所述第二数据流图包括表示数据处理组件的一个或多个节点和表示所述数据处理组件之间的数据流的一个或多个链路。

[0009] 指示所述数据源的所述信息标识所述数据源的类型。接收指示所述数据源的所述信息包括:从一个或多个类型的集中接收对所述数据源的类型的选择。所述一个或多个类型的集包括文件、表、HDFS数据集、HIVE表、和队列中的一个或多个。

[0010] 指示所述数据源的所述信息包括指示所述数据源的位置的信息。

[0011] 指示所述数据目标的所述信息标识所述数据目标的类型。接收指示所述数据目标的所述信息包括:从一个或多个类型的集中接收对所述数据目标的类型的选择。所述一个

或多个类型的集包括文件、表、HDFS数据集、HIVE表、和队列中的一个或多个。

[0012] 指示所述数据目标的所述信息包括指示多个数据目标的信息。

[0013] 指示所述数据目标的所述信息包括指示所述数据目标的位置的信息。

[0014] 所述过程的所述特性包括多个过程中的每个过程的特性。

[0015] 所述方法包括基于所述过程的类型确定与所述过程相关联的所述特性。

[0016] 接收过程的特性包括:从一个或多个过程的集中接收对特定过程的选择。所述集中的所述过程包括映射过程、过滤过程、去重过程、聚合过程、复制过程、规范化过程、划分和排序过程、改变数据捕获过程、以及合并过程中的一个或多个。

[0017] 所述方法包括在所述第一数据流图的运行时自动地配置所述第一数据流图以引用所述第二数据流图。

[0018] 自动地配置所述第一数据流图以引用所述第二数据流图包括配置所述第一数据流图以在所述第一数据流图的执行期间执行所述第二数据流图。

[0019] 一方面,一种非暂态计算机可读介质存储用于使计算系统配置第一计算机可执行程序的指令。所述指令包括用于使所述计算系统执行以下操作的指令:通过用户界面,接收指示数据源和数据目标的信息;通过所述用户界面,接收过程的特性,所述特性包括所述过程的类型和与所述过程相关联的相应特性的值;基于所接收到的信息,自动地为所述第一计算机可执行程序的相应参数分配值,以使所述第一计算机可执行程序在被执行时从所述数据源接收数据并将数据输出到所述数据目标;以及基于所接收到的所述过程的特性,自动地配置所述第一计算机可执行程序以引用第二计算机可执行程序。所述配置包括基于所述过程的类型标识所述第二计算机可执行程序;以及基于与所述过程相关联的所述相应特性的值,为所述第二计算机可执行程序的相应参数分配值。

[0020] 一方面,一种用于配置第一计算机可执行程序的计算系统包括存储器和一个或多个处理器,所述存储器和一个或多个处理器被配置用于通过用户界面接收指示数据源和数据目标的信息。通过所述用户界面,接收过程的特性,所述特性包括所述过程的类型和与所述过程相关联的相应特性的值;基于所接收到的信息,自动地为所述第一计算机可执行程序的相应参数分配值,以使所述第一计算机可执行程序在被执行时从所述数据源接收数据并将数据输出到所述数据目标;以及基于所接收到的所述过程的特性,自动地配置所述第一计算机可执行程序以引用第二计算机可执行程序。所述配置包括基于所述过程的类型标识所述第二计算机可执行程序;以及基于与所述过程相关联的所述相应特性的值,为所述第二计算机可执行程序的相应参数分配值。

[0021] 一方面,一种用于配置第一计算机可执行程序的计算系统包括:用于通过用户界面接收指示数据源和数据目标的信息的装置;用于通过用户界面接收过程的特性的装置,所述特性包括所述过程的类型和与所述过程相关联的相应特性的值;用于基于所接收到的信息自动地为所述第一计算机可执行程序的相应参数分配值以使所述第一计算机可执行程序在被执行时从所述数据源接收数据并将数据输出到所述数据目标的装置;以及用于基于所接收到的所述过程的特性自动地配置所述第一计算机可执行程序以引用第二计算机可执行程序的装置。所述配置包括基于所述过程的类型标识所述第二计算机可执行程序;以及基于与所述过程相关联的所述相应特性的值,为所述第二计算机可执行程序的相应参数分配值。

附图说明

[0022] 图1是系统图。

[0023] 图2是数据流图的图。

[0024] 图3是用户界面的图。

[0025] 图4是流程图。

[0026] 图5A至图5G是示例屏幕截图。

[0027] 图6至图8是示例屏幕截图。

[0028] 图9A和图9B是示例屏幕截图。

[0029] 图10是系统图。

具体实施方式

[0030] 在此描述了一种用于基于用户提供的特性来配置数据流图的方法。所述特性可以定义要由数据流图处理的数据源、从数据流图输出的数据目标,并且可以定义要由数据流图对数据执行的处理。可以从可用选项库中选择要执行的处理类型。根据这些特性,可以对通用的、可再使用的数据流图进行参数化。

[0031] 参照图1,用于配置数据流图的示例系统10包括数据源12,所述数据源可以包括一个或多个数据源(诸如存储设备或与在线数据流的连接),每个数据源可以以各种存储格式(例如,数据库表、电子表格文件、纯文本文件、或主机使用的本机格式)中的任意一种存储数据。执行环境14包括图配置模块16和用户界面模块22。执行环境14可以在诸如UNIX操作系统之类的合适的操作系统的控制下被托管在一个或多个通用计算机上。例如,执行环境14可以包括多节点并行计算环境,所述多节点并行计算环境包括使用多个中央处理单元(CPU)的计算机系统的配置,所述中央处理单元可以是本地的(例如,诸如SMP计算机之类的多处理器系统)、或本地分布式的(例如,耦合为集群或MPP的多个处理器)、或远程的、或远程分布式的(例如,经由LAN或WAN网络耦合的多个处理器)、或其任意组合。

[0032] 如以下更详细地描述的,图配置模块16改变数据流图的配置。用户界面模块22向用户30显示配置信息,并从用户30接收配置动作。用户界面模块22还与图配置模块16通信,所述图配置模块基于用户的动作来配置数据流图。例如,数据流图可以存储在数据源12中。提供数据源12的存储设备可以在执行环境14本地,例如,存储在连接到运行执行环境14(例如,硬盘驱动器18)的计算机的存储介质上,或者可以远离执行环境14,例如,被托管在通过局域网或广域网与运行执行环境14的计算机通信的远程系统(例如,主机20)上。

[0033] 执行环境与数据存储系统26通信,所述数据存储系统包含由用户界面模块22用来显示用户界面的信息。数据存储系统26还可以被开发环境28访问,在所述开发环境中,开发者30能够开发存储在数据存储系统26中的用户界面,所述用户界面被用户界面模块22用来显示用户界面。

[0034] 在一些实施方式中,数据源12是用于将应用程序开发为数据流图的系统,所述数据流图包括通过顶点之间的有向链路(表示工作元素流)连接的顶点(组件或数据集)。例如,在标题为“基于图的应用程序的参数管理(Managing Parameters for Graph-Based Applications)”的第2007/0011668号美国公开中更详细地描述了这样的环境,所述美国公开通过引用并入本文。

[0035] 数据流图是在数据流图执行环境中执行的计算机程序,所述计算机程序处理来自一个或多个数据源的数据。来自数据源的数据根据数据流图进行才操作和处理,并导出到一个或多个数据接收器。例如,数据源和接收器可以包括文件、数据库、数据流、或队列。数据流图被表示为有向图,所述有向图包括:表示数据处理组件的节点,每个节点包括用于处理来自至少一个数据输入的数据并将数据提供给至少一个数据输出的代码;以及表示用于访问数据源和/或接收器的数据集对象的节点。节点通过有向链路进行连接,所述有向链路表示组件之间的数据流,始于数据源,并终止于数据接收器。上游组件的数据输出端口连接到下游组件的数据输入端口。可以将数据流图重新用于由数据集对象表示的不同数据源和不同数据接收器。例如,用于实现数据流图的数据结构和程序代码可以通过参数化来支持多种不同的配置,以使不同的源和接收器易于替换。此外,在一些布置中,可以通过使用参数来更改数据流图的流动,使得可以绕过一个组件或一系列组件。通常,参数表示可以被配置或改变的数据流图的属性。例如,可以在数据流图的使用之间改变属性,并且由于所述改变,数据流图可以不同的方式执行操作。可以在编译时通过更改数据流图的预编译以执行特定操作、或者在运行时通过设置参数或更改由数据流图使用的配置文件来配置数据流图。

[0036] 例如,参照图2,通用数据流图100包括数据源102、组件104和数据接收器106a至106n。数据源102连接到组件104的输入端口。组件104的输出端口连接到数据接收器106a至106n。数据源、组件、和数据接收器之间的连接定义了通过数据流图的数据流。

[0037] 数据源102、组件104、和数据接收器106a至106n中的每一个分别与参数集108、110、112a至112n相关联,每个参数集包括一个或多个参数。参数定义了数据流图的行为。例如,参数可以定义数据源或数据接收器在物理磁盘上的位置。参数还可以定义组件的行为,诸如排序组件如何将输入到组件中的数据排序。在一些示例中,在数据流图的运行时填充参数集中的参数的值。

[0038] 在一些示例中,一个参数的值可以取决于另一参数的值。例如,数据源可以存储在特定目录中的文件中。数据源的参数集可以包括名为“DIRECTORY(目录)”的第一参数和名为“FILENAME(文件名)”的第二参数。在此示例中,FILENAME参数取决于DIRECTORY参数(例如,DIRECTORY可以是“/usr/local/”,而FILENAME可以是“input.dat”)。参数还可以取决于其他组件的参数。例如,用于数据流图的数据接收器的物理位置可以取决于用于数据流图的数据源的物理位置。例如,数据接收器可以包括参数集,所述参数集包括取决于数据源的DIRECTORY参数的FILENAME参数(例如,数据接收器的FILENAME参数可以是“/usr/local/output.dat”,其中值“/usr/local/”是从数据源的DIRECTORY参数中获得的)。

[0039] 组件104可以是引用一个或多个其他数据流图120a至120n(有时称为子图)的图形界面组件。在运行时,数据流图100动态地加载并执行组件104引用的(多个)子图,例如,使数据流图100能够灵活地访问子图120a至120n提供的各种功能。组件104的一个或多个参数定义了组件104所引用的(多个)特定子图120a至120n。每个子图120a至120n还与包括一个或多个参数的参数集122a至122n相关联,每个参数定义相应子图的行为。

[0040] 在一些情况下,数据流图的构造本质上可以是高度技术性的。为达到特定的业务目的而编写时,图形的基础结构和构造是基于技术考虑来确定的。例如,可以选择图形组件以最大化可再使用性,或支持并行处理。另一方面,如何使用图形以及在何处使用图形可能

在很大程度上是业务决策。与参数化的数据流图相关联的一些参数可用于使业务用户能够自定义数据流图,而不要求用户理解其实施背后的技术复杂性。参数化的数据流图简化了自定义化并有助于再使用。

[0041] 可以在客户机上呈现用于标识用于构造数据流图的参数值的界面。在一些实施方式中,客户机可能正在使用提供参数界面的客户机上的Web浏览器并使用为客户机侧处理提供某些功能的脚本语言访问在服务器上运行的开发环境。脚本语言可以与服务器通信以更新参数并执行其他必要的操作。此通信可以经由网桥机发生,所述网桥机转换客户机与运行开发环境的服务器之间的通信,所述服务器存储用于被构造的图形的对象和相关联参数值。即使用户缺乏有关数据流图和数据流图配置的技术知识,界面也允许用户配置参数化数据流图的参数。

[0042] 呈现在客户机设备上的配置界面使用户能够访问图1的图形配置模块16。通过配置界面,用户可以指定数据源、数据接收器的特性,以及数据流图要执行的转换,而无需关于数据流图配置的技术知识。基于用户指定的特性,可以为参数集108、110、112a至112n、122a至122n分配参数值,从而根据用户指定的特性定义数据流图的行为。

[0043] 在配置界面内,参数集108、110、112a至112n的参数可以被组合并重组为用于与用户交互的组,例如,反映业务考虑而非技术考虑。用于基于用户输入来接收参数的值的配置界面可以根据参数之间的关系以灵活的方式显示不同的参数,所述方式不必受到服务器上开发环境方面的限制。在第2011/0145748号美国公开中描述了配置接口的示例,其内容通过援引整体并入本文。

[0044] 参照图3,配置界面300使用户能够为数据流图100的参数分配值。在图3的示例中,参数被分为第一组参数,所述第一组参数由表示数据源102的参数的源图标302、表示组件104的参数的转换图标304、和每一个表示数据接收器106a至106n中的相应一个的参数的接收器图标306a至306n表示。用户可以输入与每个图标相关联的特性。基于用户输入的特性,可以确定数据源102、数据接收器106a至106n、和组件104的参数的值。

[0045] 在图3的示例中,与源图标302相关联的第一特性被显示为文本框308,用户可以在所述文本框中输入文本。文本的类型可以受到约束(例如,整数、浮点值、字符串、或其他类型),也可以接受任何类型的文本。第二特性显示为具有预填充值的下拉列表310。第三特性显示为单选按钮312。用于第四特性的字段314允许用户输入更大量的文本。也可以使用其他类型的显示。例如,用户可能能够打开文件选择器以选择要用作数据源的特定数据集,或者选择要通过网络访问或下载数据集的位置。也可以输入与接收器图标306a至306n和转换图标304相关联的特性。

[0046] 与源图标302或接收器图标306a至306n相关联的特性可以包括指示用于数据源102或数据接收器106a至106n之一的数据集的类型的特性。例如,可以提示用户从被支持的类型列表中选择数据集的类型,诸如文件、表、HDFS(Hadoop分布式文件系统)数据集、HIVE表、队列、或其他类型。所谓文件,是指具有可描述为文件或用与数据流图100的实施兼容的记录格式定义的形式的数据。例如,MVS(多虚拟存储)文件就是文件的示例。可以提示用户指定数据集的路径或位置。提示用户的其他特性可以取决于用户选择的数据集的类型。例如,用户可以输入与数据集的记录格式、数据集的密钥、表数据集的配置文件、压缩或加密的类型、或其他特性相关的特性。当使用多个数据接收器时,每个数据接收器可以具有其自

己的数据集类型和其他相关联的特性。

[0047] 与源图标302相关联的输入特性用于自动地参数化数据源102,以便在被执行时,数据流图从适当的源接收数据。类似地,与每个接收器图标306a至306n相关联的输入特性用于自动地参数化每个数据接收器106a至106n,以便在被执行时,从数据流图100输出的数据被发送到适当的目的地。

[0048] 与转换图标304相关联的特性可以包括指示将对来自数据源的数据执行的处理的类型的特性。例如,可以提示用户从被支持的过程库中选择过程类型。每个过程对应于可以被组件104引用的先前构造的、存储的子图120a至120n。例如,系统可以支持以下过程,诸如映射过程、过滤过程、去重过程(有时称为Dedupe过程)、聚合过程、复制过程、规范化过程、划分和排序过程、改变数据捕获过程、合并过程、或其他过程。提示用户的其他特性可以取决于过程的类型。例如,对于映射过程,用户可以输入映射规则并标识映射的目标数据集。对于过滤过程,用户可以输入过滤规则。对于去重过程,用户可以输入去重密钥并指定记录选择规则。对于合并过程,用户可以标识合并数据集并定义合并规则。在一些示例中,用户可以选择多个过程并指示所选择过程的期望顺序。

[0049] 与转换图标304相关联的输入特性被用于自动地参数化组件104以及任何相关的子图120a至120n。参数化组件104以引用与用户选择的过程类型相对应的(多个)子图。每个引用的子图也会基于用户为相应过程输入的特性被参数化。在被执行时,数据流图100将执行组件引用的每个子图120a至120n,其中,每个引用的子图的行为由基于用户输入的特性的参数化来定义。例如,数据流图100可以以串行、流水线过程来执行每个引用的子图120a至120n,其中,在子图的执行期间不存储任何数据。

[0050] 在本文描述的示例中,系统支持单个数据源和多个数据接收器,诸如多至两个数据接收器、五个数据接收器、十个数据接收器、二十个数据接收器、或其他数量的数据接收器。可以分别定义每个数据接收器的参数,例如,使得每个数据接收器可以是不同类型的数据集和/或可以具有其自己的参数集。在一些示例中,系统可以支持多个数据源,例如,可以使用合并过程、查找过程、或其他类型的过程来处理来自多个源的数据。例如,可以通过合并过程、查找过程、或合并和查找过程的组合将多个源合并为有效的单个源。在一些示例中,系统可以支持单个数据接收器。

[0051] 在一些示例中,系统可以基于提供数据源和一组处理步骤的细节的用户输入来自动地生成配置。例如,用户可以通过诸如电子表格之类的用户界面提供数据源和一组处理步骤的细节,并且系统可以基于用户输入的细节来自动地配置计算机程序。

[0052] 参照图4,在用于配置第一计算机可执行程序的示例过程中,接收指示数据源和数据目标的信息(400)。例如,所述信息可以指示数据集的类型、数据集的位置或路径、以及关于数据集的其他特性。接收过程的特性(402)。特性包括过程的类型和过程的相应特性的值。

[0053] 基于接收到的信息,自动地为第一计算机可执行程序的相应参数分配值(404),以使第一计算机可执行程序在程序被执行时从数据源接收数据并将数据输出到数据目标。例如,第一计算机可执行程序可以是数据流图。

[0054] 基于接收到的过程的特性,第一计算机可执行程序被自动地配置为引用第二计算机可执行程序(406)。所述配置包括基于过程的类型来标识第二计算机可执行程序(408)。

所述配置还包括基于与过程相关联的相应特性的值,为第二计算机可执行程序的相应参数分配值(410)。例如,第二计算机可执行程序可以是子图。

[0055] 图5A至图5G是配置界面的示例,用户可以通过所述配置界面输入关于数据源或数据目标的信息。参照图5A,提示用户选择源或目标数据集的类型。

[0056] 参照图5B和图5C,如果数据集的类型是文件,则提示用户输入验证细节和特性,诸如文件的URL、文件的记录格式、文件是否被压缩或被加密或既被压缩又被加密、是否对文件进行读或写、文件的自然密钥、文件的数据顺序密钥、或其他参数中的一个或多个。

[0057] 参照图5D,可以输入关于数据集的元数据,诸如物理文件的rpath或存储库名称。

[0058] 参照图5E,如果数据集的类型是表,则提示用户输入特性,诸如是否使用表或查询(例如,SQL查询)、是否对表进行读或写、表的数据库配置文件、表的自然密钥、表的数据顺序密钥、或其他参数。如果选择了查询,则用户可以指定查询。

[0059] 参照图5F,如果数据集的类型是HDFS数据集,则提示用户输入描述HDFS数据集的特性,诸如是否对HDFS文件进行读或写、HDFS文件系统的路径、HDFS文件的名称、HDFS文件的后缀(如果有的话)、记录格式文件的路径、或任何各种其他特性中的一个或多个。用户可以提供指示文件类型或格式的特性,诸如要读取或写入的HDFS文件的类型,诸如未解译的文件(例如,未压缩、未格式化的文本(例如,ASCII或UTF8)或二进制数据)、行列格式(诸如Apache优化行列格式)、列存储格式(诸如Apache Parquet)、包含密钥值对的二进制文件(诸如Hadoop序列文件)、二进制文件格式(诸如Apache Avro)、文本数据或其他类型。用户可以提供指示压缩类型的特性,诸如gzip、Snappy、Deflate、bzip2、zlib、LZ4、LZO、或其他压缩类型。用户可以为HDFS数据集提供可直接从现有的HDFS数据集导出的记录格式,也可以提供记录的格式。在一些示例中,可以根据对文件类型、压缩类型、或两者的选择来提示用户附加特性。

[0060] 参照图5G,如果数据集的类型为HIVE表,则提示用户输入描述HIVE表的特性,诸如是否对HIVE表进行读或写、HIVE表的文件格式、记录格式、记录格式文件的路径、数据库、表是否存在、表、或任何各种其他特性中的一个或多个。在一些示例中,可以根据对表类型和/或压缩类型的选择来提示用户附加特性。

[0061] 图6是配置界面600的示例,在所述配置界面中用户可以定义处理步骤。在图6的示例中,已经定义了源数据集602和两个目标数据集604a、604b。处理步骤窗口606允许用户选择要在源数据集上执行的处理步骤。

[0062] 图7是示出了用户已经指定了一系列处理步骤的提要的配置界面700的示例。已经定义了源数据集702。还定义了四个处理步骤:映射步骤704、去重步骤706、第二映射步骤708、和合并步骤710。已经为每个过程指定了各种特性。

[0063] 当基于配置界面700中示出的提要对数据流图进行参数化时,数据源将被配置为从源表Teradata_cc_trans接收数据,如源数据集窗口702中指定的那样。数据流图的组件将被配置为引用映射子图、去重子图、再次映射子图、和合并子图。每个引用的子图将根据窗口704至710中示出的特性对相应过程进行参数化。

[0064] 从配置界面700中的提要中可以看出,用户可以输入数据流图的配置特性,而无需对图配置进行详细的技术理解。例如,在图7的示例中,用户知道其想要处理的数据的源以及所处理的数据的目的地(未示出)。用户知道其想将源数据映射到自身,例如,以改变一个

或多个字段的映射。他知道他想使用去重密钥{amount}对映射的数据进行去重,并保留每个组中的第一记录。在去重步骤之后,他想再次将数据映射到自身。最后,他想将经去重的和映射的数据合并到表mixed_cc_accounts中。也就是说,为了设置此流程,用户可以了解通用数据处理框架,而无需知道如何配置数据流图。

[0065] 在一些示例中,通过配置界面700,用户可以选择测试数据集来用作源数据集702。例如,用户可以选择可以被更快地处理的较小的数据集,而不是使用极大的数据集(例如,具有实际客户记录的数据集)来测试数据流图。在一些示例中,例如,出于测试目的,用户可以为处理后的数据选择一个目的地,所述目的地不同于在实际操作中数据流图将发送其处理后的数据的目的地。例如,用户可以选择测试数据目的地,以免用实际的客户记录覆盖数据集。无论源数据集702和已处理数据的目的地如何,其他过程706至708的配置均保持不变,并且因此用户可以如所期望的直接更新源数据集和/或目的地。

[0066] 在一些示例中,配置界面700可以提供关于用户输入的参数是否有意义的指示。例如,用户输入的过程密钥可以相互验证。在示例中,如果用户创建了使用第一密钥对数据集进行去重并使用第二密钥对所得的去重数据进行排序的数据流图,则可能在配置界面中显示警报。

[0067] 在一些示例中,在用户已经配置了数据流图之后,系统可以确定如何执行数据流图,例如以减少处理时间或计算成本。例如,对于给定的过程,系统可以基于一个或多个上游过程的参数(诸如上游过程中使用的密钥)来判定是否重新划分数据。在一个示例中,如果在上游过程中使用了相似的密钥,则系统可以确定在下游过程中使用相同的密钥对数据进行重新划分是不必要的。重新划分或排序是否适当的判定可以自动地执行,而无需用户输入。

[0068] 在一些示例中,指示数据流图的性能的信息可以在数据流图的执行期间显示在配置界面700中。所述信息可以包括指示进出每个处理步骤的记录的数量的信息。例如,去重步骤706可以记录并显示输入的记录数量、输出的记录数量、以及作为重复项丢弃的记录数量。合并步骤710可以记录并显示来自第一数据集的输入的记录数量、来自第二数据集的输入的记录数量、输出的记录数量、以及从每个数据集丢弃的记录数量。此信息可以向用户提供快速反馈,例如,指示用户是否按照其意图配置了数据流图,从而为非技术用户提供了对数据流图进行故障排除和调试的方法。例如,参照图8,看到115,911条记录(其中,108,191条记录作为重复项被丢弃)被读取到去重步骤中的非技术用户可以使用此信息来确定诸如去重密钥之类的参数可能被错误地配置。

[0069] 在一些示例中,配置界面700可以被呈现为电子表格界面,用户可以通过所述电子表格界面来构建和配置数据流图。

[0070] 在一些示例中,可以例如基于标识要使用的数据(诸如,电子表格、数据库、或其他位置中的数据)的特性来创建源数据集、目标数据集、或两者。

[0071] 参照图9A和图9B,在一些示例中,可以在将处理后的数据从数据流图发送到数据目标之前分别应用目标特定的过程,诸如映射过程或过滤过程。

[0072] 参照图10,客户机系统802可以正在向用户显示配置界面804。基于通过配置界面804与用户的交互而生成的参数集814可以存储在服务器808上。因此,由配置接口804进行的改变经由网桥806从客户机802发送到服务器808。由箭头820表示,客户机802将消息以一

种格式发送到网桥806,例如,使用简单对象访问协议(SOAP)发送的消息。网桥806将消息转换为新格式,并且如果必要,则开始与服务器808的客户会话。由箭头822表示,网桥806将消息以服务器808可理解的格式(例如COM+消息)发送到服务器808。服务器808接收消息并更新参数集。由箭头824表示,服务器808将包含由于客户机802接收的输入而对参数集发生的任何改变的应答发送到桥806。网桥806解码消息并为客户机802创建应答消息。由箭头826表示,网桥806将应答消息发送到客户机802。客户机802更新配置界面804以反映改变,包括显示如上所描述的由于先决条件的失败而先前隐藏的任何组件。

[0073] 用户还可向客户机802指示其希望基于当前参数集执行使用样本数据构造的数据流图,所述参数集可以是完整的也可以是不完整的。如上,客户机802经由网桥806将消息发送到服务器808。服务器808将任何改变应用于参数集,并且在服务器上运行的过程816编译数据流图。经编译的数据流图从样本数据集810、812接受数据,并执行经编译的数据流图。数据流图将请求的输出产生到输出数据集818。数据流图的输出是客户机802请求的中间数据,而不一定是将通过完全执行数据流图所产生的数据。

[0074] 如上所描述的,所得数据经由网桥806从服务器808发送到客户机802。

[0075] 可以使用用于在计算机上执行的软件来实施上述图形配置方法。例如,软件在一个或多个计算机程序中形成程序,所述程序在一个或多个编程或可编程计算机系统(其可以是诸如分布式、客户机/服务器、或网格之类的各种架构)上执行,所述编程或可编程计算机系统中的每一个包括至少一个处理器、至少一个数据存储系统(包括易失性和非易失性存储器和/或存储元件)、至少一个输入设备或端口、以及至少一个输出设备或端口。软件可以形成较大程序(例如,提供与计算图的设计和配置有关的其他服务的程序)的一个或多个模块。图的节点和元素可以实施为存储在计算机可读介质中的数据结构或与存储在数据存储库中的数据模型一致的其他组织数据。

[0076] 软件可以被提供在诸如CD-ROM之类的存储介质上,所述软件可以由通用或专用可编程计算机读取,或者可以通过网络的通信介质被传递(被编码在传播的信号中)到执行所述软件的计算机。所有功能都可以在专用计算机上执行,或者使用专用硬件(诸如协处理器)执行。可以以分布式方式来实施软件,其中,由软件指定的计算的不同部分由不同的计算机执行。每个这样的计算机程序优选地被存储在或下载到通用或专用可编程计算机可读的存储介质或设备(例如,固态存储器或介质、或磁性或光学介质)上用于当计算机系统读取存储介质或设备以执行本文所描述的过程时配置并操作计算机。本发明的系统也可以被认为是实施为配置有计算机程序的计算机可读存储介质,其中,如此配置的存储介质使计算机系统以特定且预定的方式操作以执行本文所描述的功能。

[0077] 已经描述了本发明的多个实施例。然而,将理解,可以在不脱离本发明的精神和范围的情况下进行各种修改。例如,上述一些步骤可以与顺序无关,并且因此可以按照与描述的顺序不同的顺序执行。

[0078] 其他实施方式也在所附权利要求的范围内。

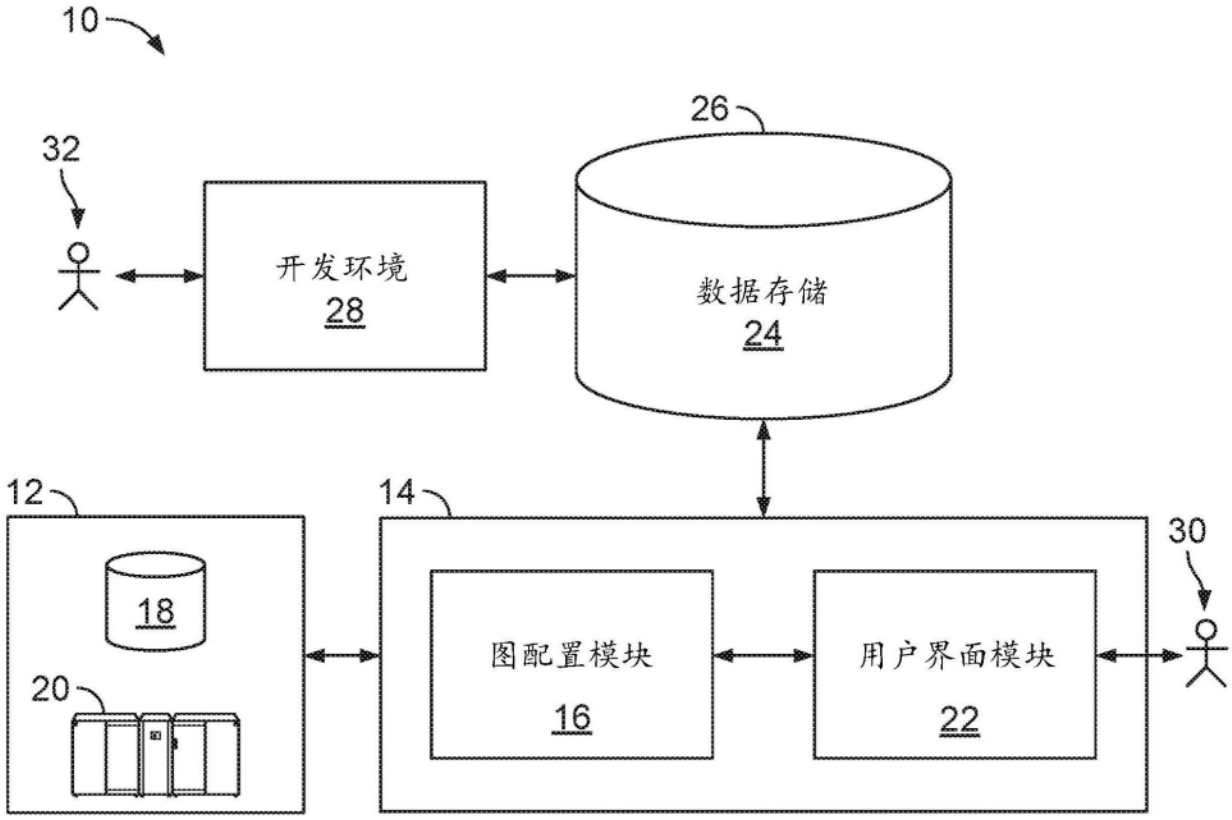


图1

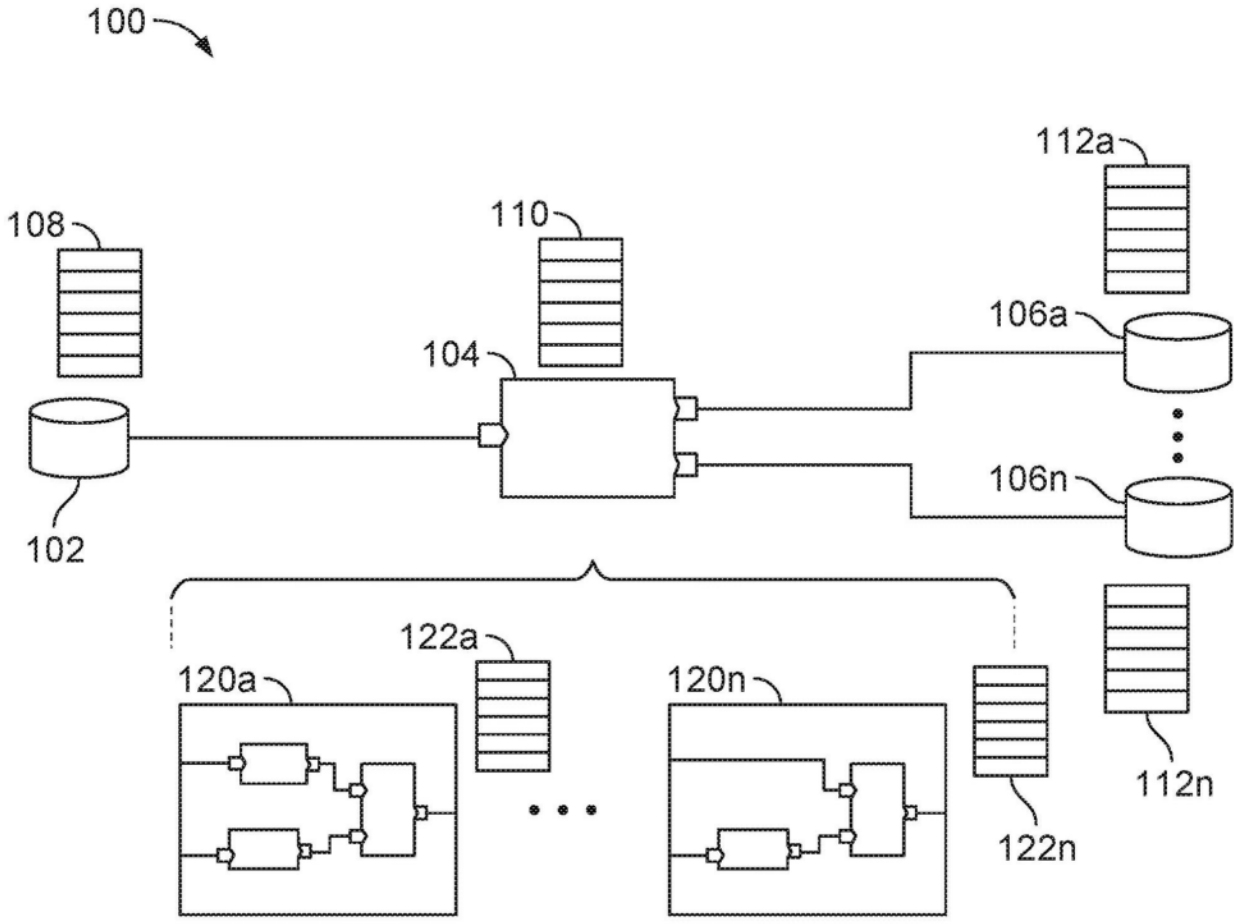


图2

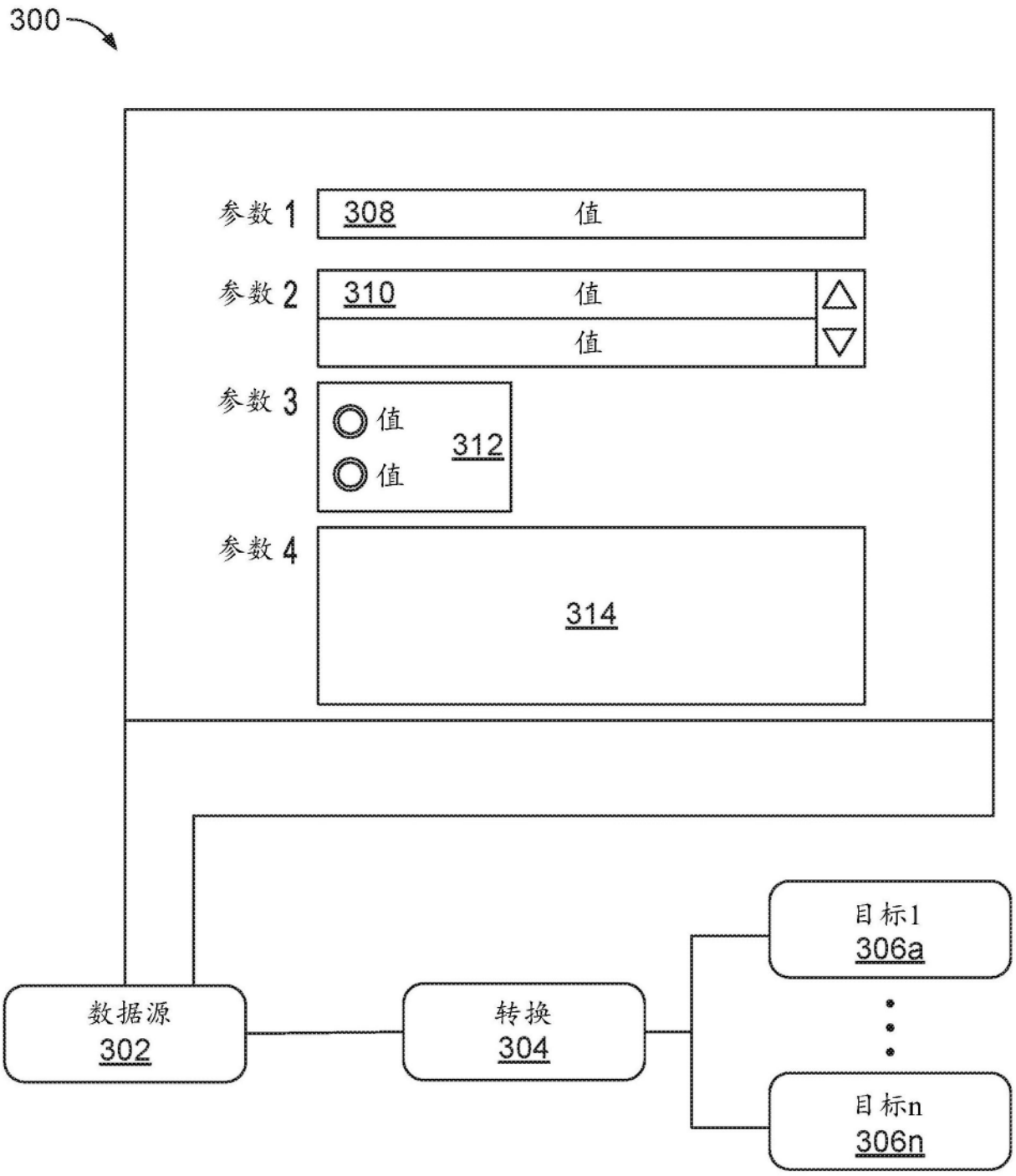

300
参数 1
308
值
参数 2
310
值
值
参数 3
值
值
312
参数 4
314
数据源
302
转换
304
目标1
306a
目标n
306n

图3

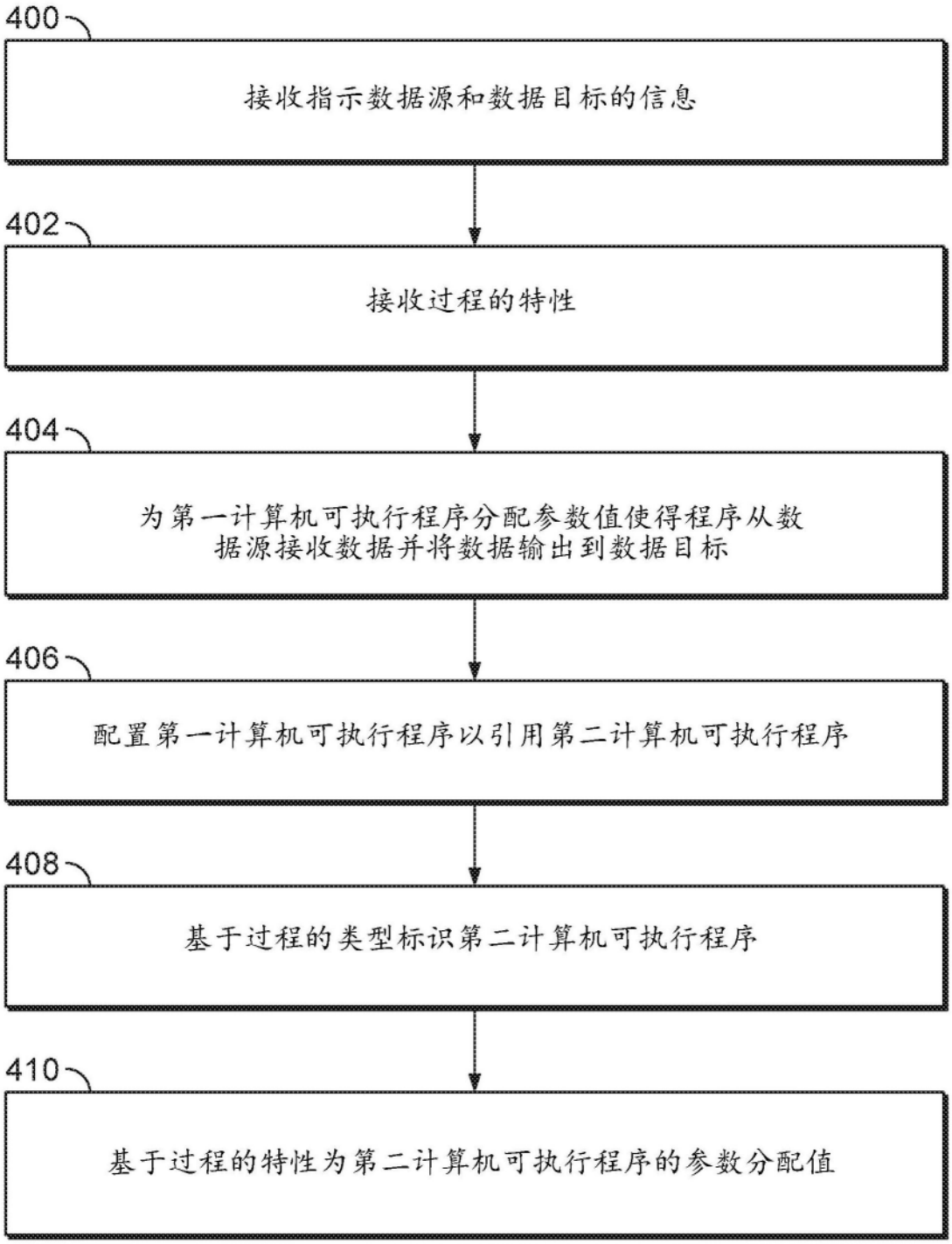

400
接收指示数据源和数据目标的信息
402
接收过程的特性
404
为第一计算机可执行程序分配参数值使得程序从数据源接收数据并将数据输出到数据目标
406
配置第一计算机可执行程序以引用第二计算机可执行程序
408
基于过程的类型标识第二计算机可执行程序
410
基于过程的特性为第二计算机可执行程序的参数分配值

图4

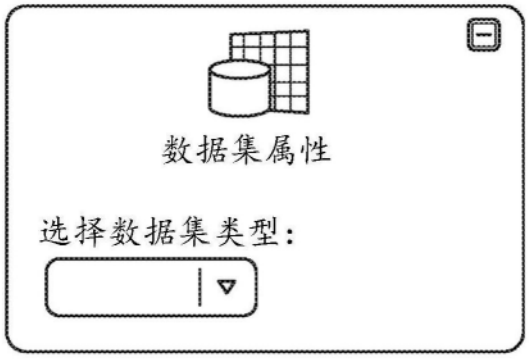

数据集属性
选择数据集类型:

图5A

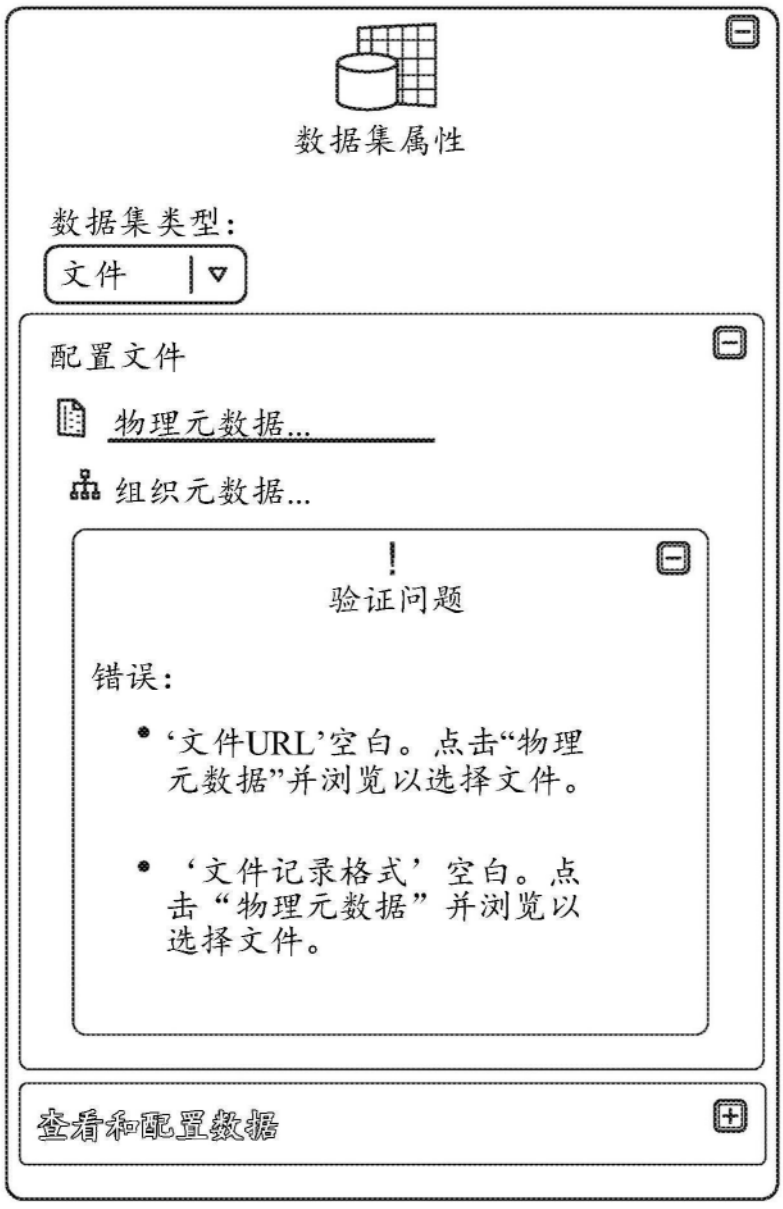

数据集属性
数据集类型:
文件
配置文件
物理元数据...
组织元数据...
验证问题
错误:
• '文件URL'空白。点击"物理元数据"并浏览以选择文件。
• '文件记录格式'空白。点击"物理元数据"并浏览以选择文件。
查看和配置数据

图5B


物理元数据
访问选项
读文件
写文件
查找文件
写模式
替换
文件详情
文件URL:
浏览...
参数化文件名
数据结构
记录格式选项:
浏览现有的记录格式
文件记录格式:
浏览...
存储特性:
压缩
确认
取消

图5C

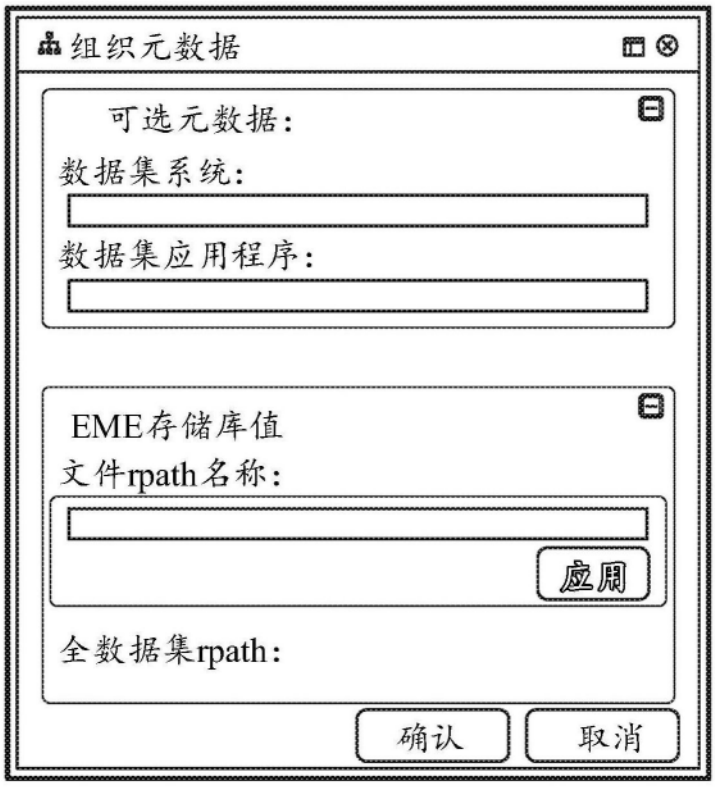

组织元数据
可选元数据:
数据集系统:
数据集应用程序:
EME存储库值
文件rpath名称:
应用
全数据集rpath:
确认
取消

图5D

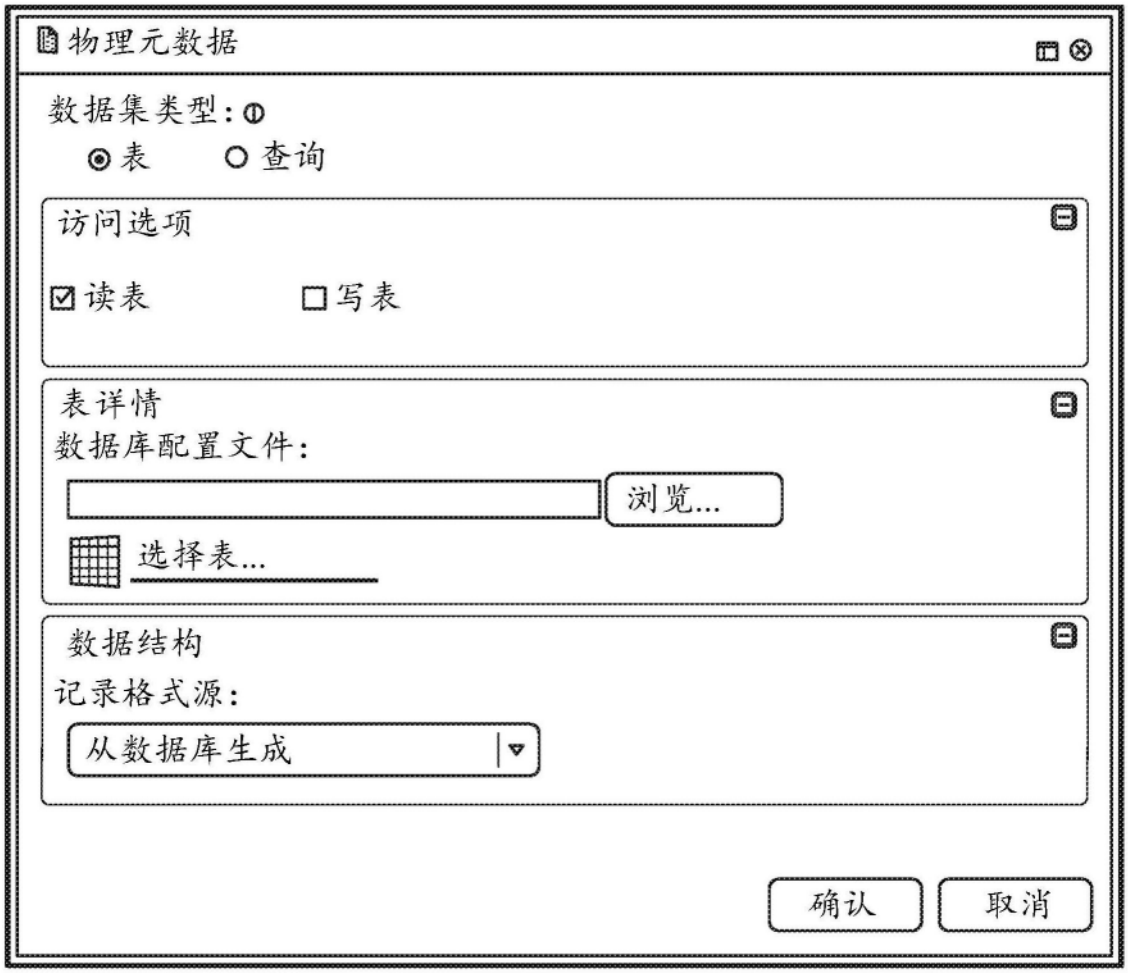

物理元数据
数据集类型:
表
查询
访问选项
读表
写表
表详情
数据库配置文件:
浏览...
选择表...
数据结构
记录格式源:
从数据库生成
确认
取消

图5E

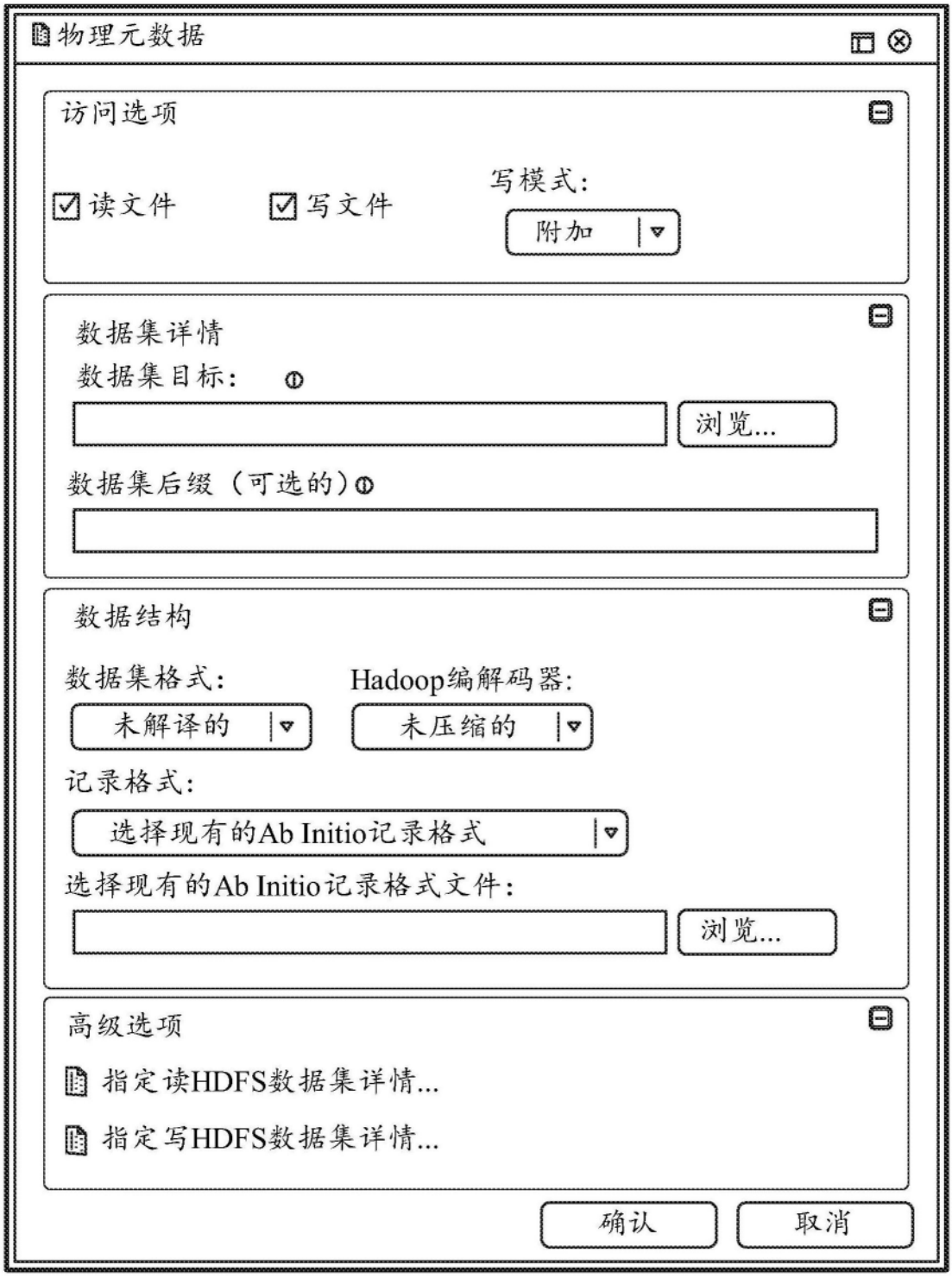

物理元数据
访问选项
读文件
写文件
写模式:
附加
数据集详情
数据集目标:
浏览...
数据集后缀（可选的）
数据结构
数据集格式:
未解译的
Hadoop编解码器:
未压缩的
记录格式:
选择现有的Ab Initio记录格式
选择现有的Ab Initio记录格式文件:
浏览...
高级选项
指定读HDFS数据集详情...
指定写HDFS数据集详情...
确认
取消

图5F

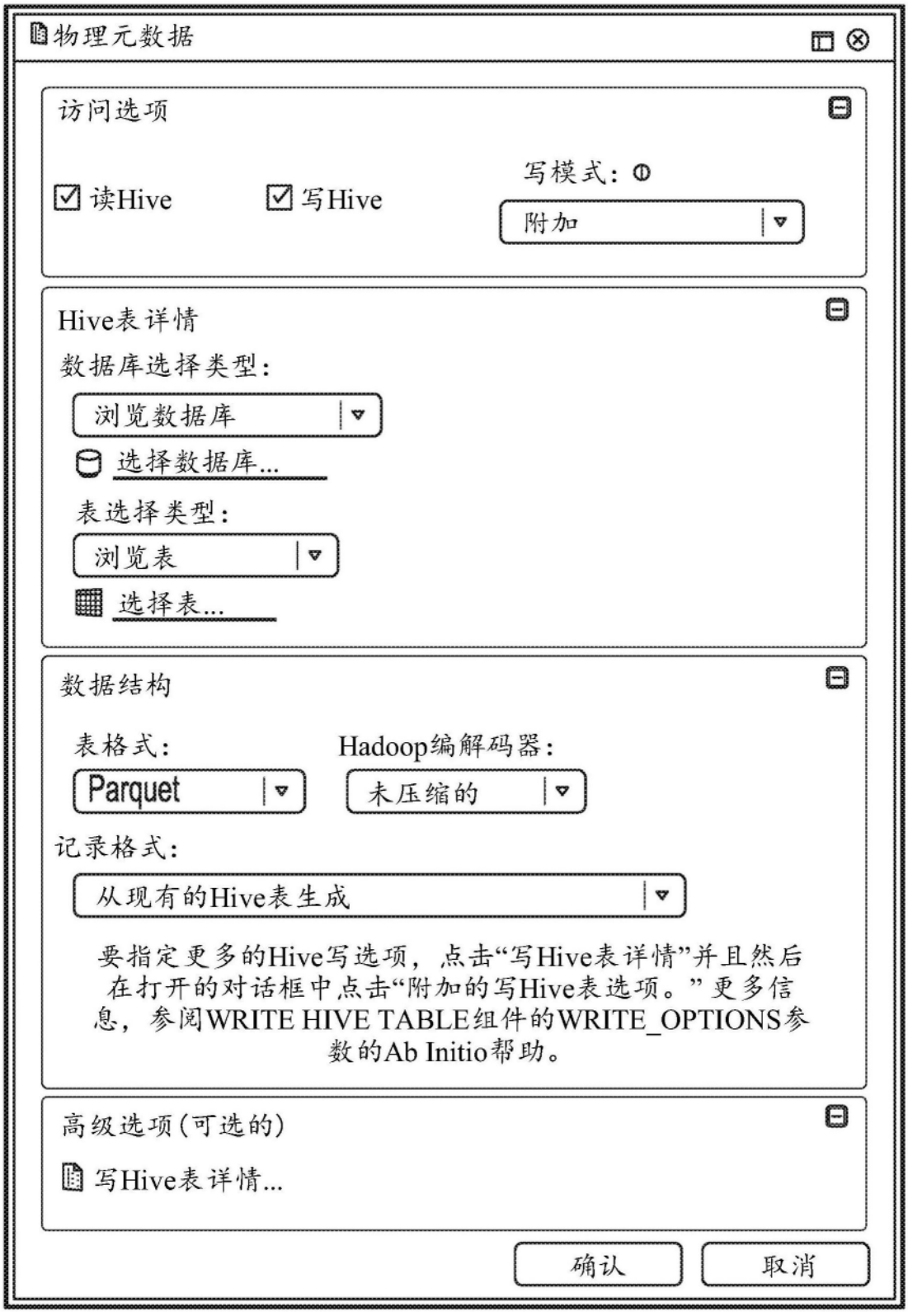

物理元数据
访问选项
读Hive
写Hive
写模式：
附加
Hive表详情
数据库选择类型：
浏览数据库
选择数据库...
表选择类型：
浏览表
选择表...
数据结构
表格式：
Parquet
Hadoop编解码器：
未压缩的
记录格式：
从现有的Hive表生成
要指定更多的Hive写选项，点击“写Hive表详情”并且然后在打开的对话框中点击“附加的写Hive表选项。”更多信息，参阅WRITE HIVE TABLE组件的WRITE_OPTIONS参数的Ab Initio帮助。
高级选项(可选的)
写Hive表详情...
确认
取消

图5G

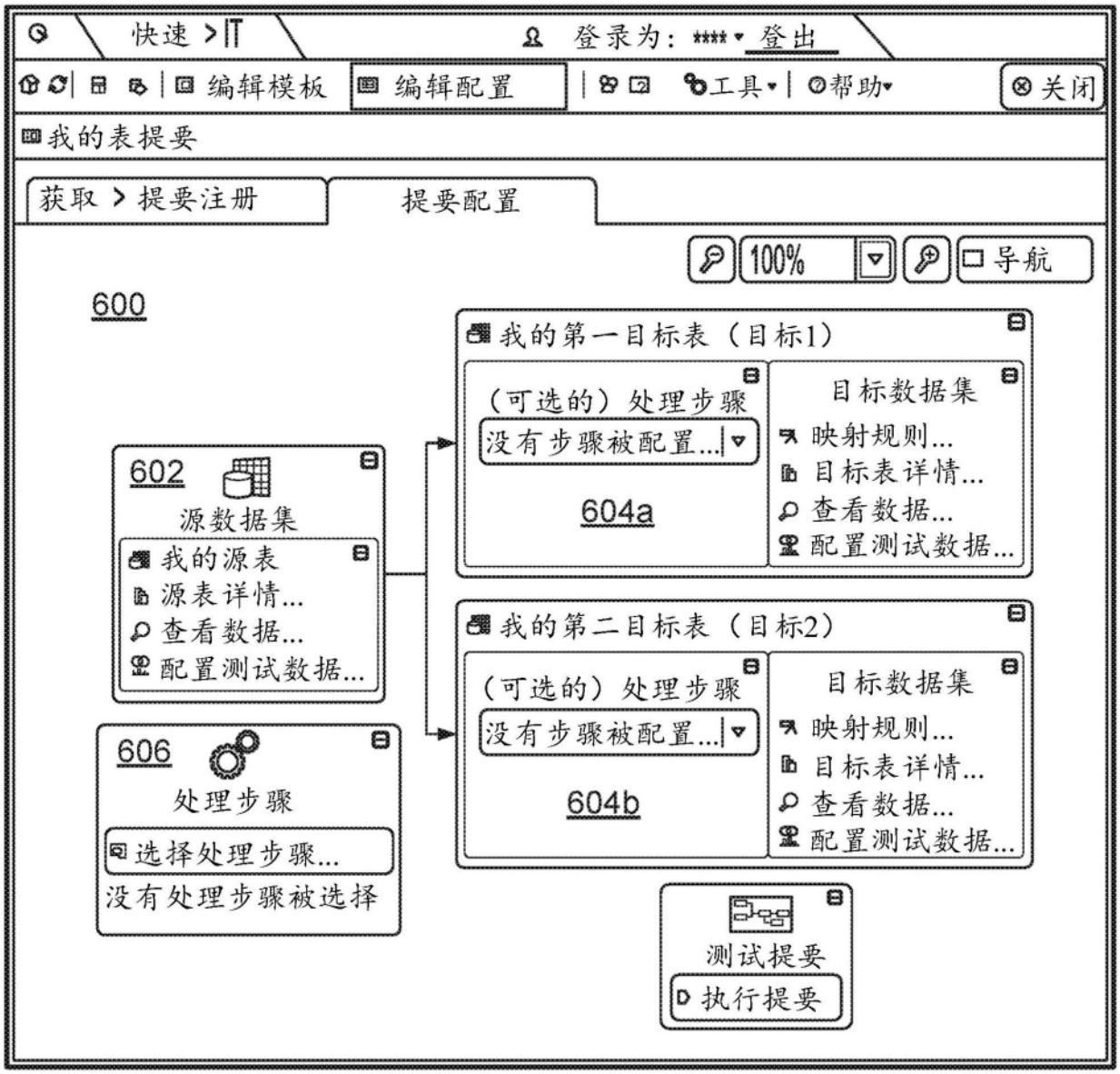

快速 > IT
登录为：**** 登出
编辑模板
编辑配置
工具
帮助
关闭
我的表提要
获取 > 提要注册
提要配置
100%
导航
600
我的第一目标表（目标1）
（可选的）处理步骤
没有步骤被配置...
604a
目标数据集
映射规则...
目标表详情...
查看数据...
配置测试数据...
602
源数据集
我的源表
源表详情...
查看数据...
配置测试数据...
我的第二目标表（目标2）
（可选的）处理步骤
没有步骤被配置...
604b
目标数据集
映射规则...
目标表详情...
查看数据...
配置测试数据...
606
处理步骤
选择处理步骤...
没有处理步骤被选择
测试提要
执行提要

图6

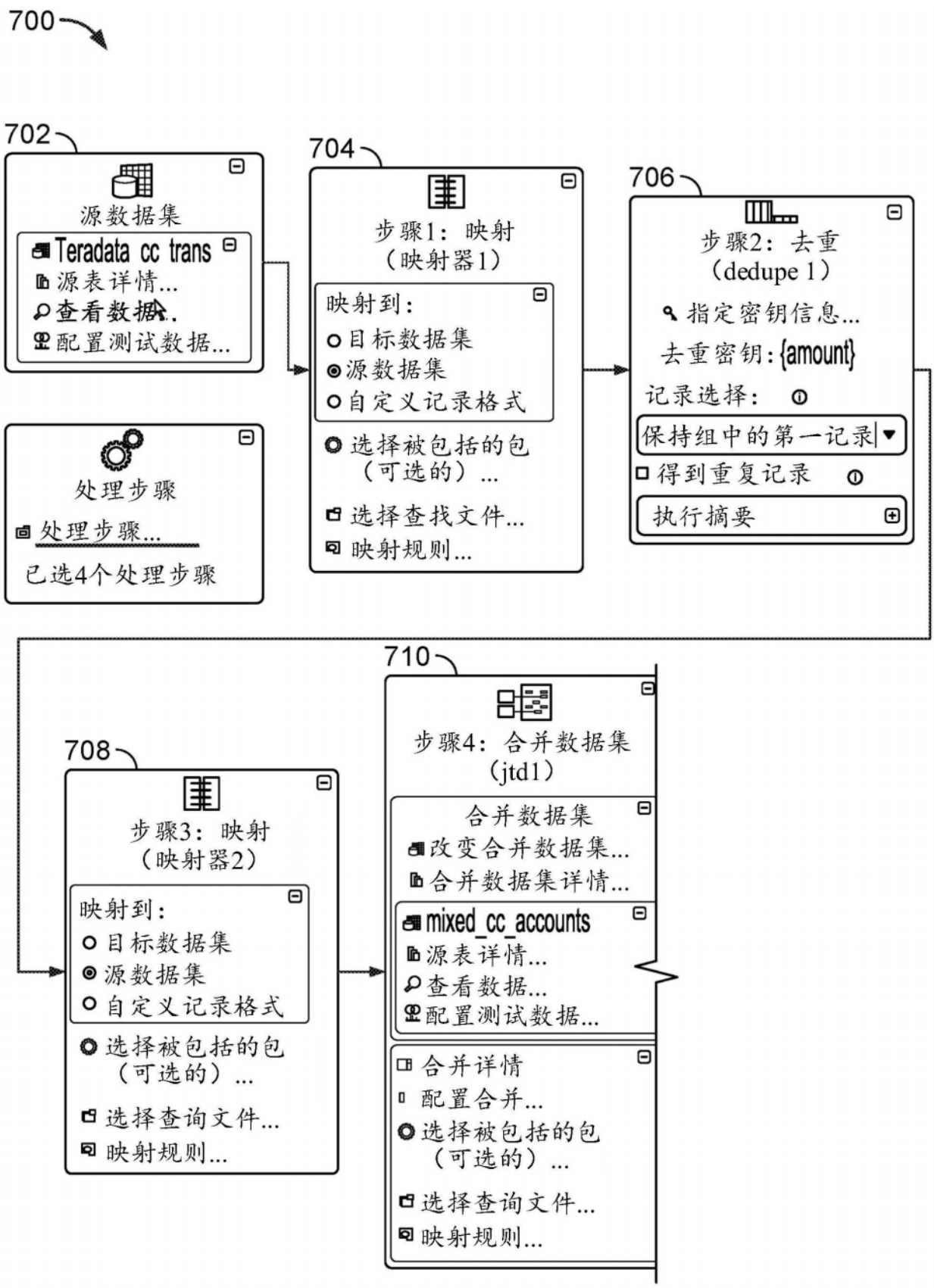

700
702
源数据集
Teradata cc trans
源表详情...
查看数据...
配置测试数据...
处理步骤
处理步骤...
已选4个处理步骤
704
步骤1：映射
（映射器1）
映射到：
目标数据集
源数据集
自定义记录格式
选择被包括的包
（可选的）...
选择查找文件...
映射规则...
706
步骤2：去重
（dedupe 1）
指定密钥信息...
去重密钥：{amount}
记录选择：
保持组中的第一记录
得到重复记录
执行摘要
708
步骤3：映射
（映射器2）
映射到：
目标数据集
源数据集
自定义记录格式
选择被包括的包
（可选的）...
选择查询文件...
映射规则...
710
步骤4：合并数据集
（jtd1）
合并数据集
改变合并数据集...
合并数据集详情...
mixed_cc_accounts
源表详情...
查看数据...
配置测试数据...
合并详情
配置合并...
选择被包括的包
（可选的）...
选择查询文件...
映射规则...

图7

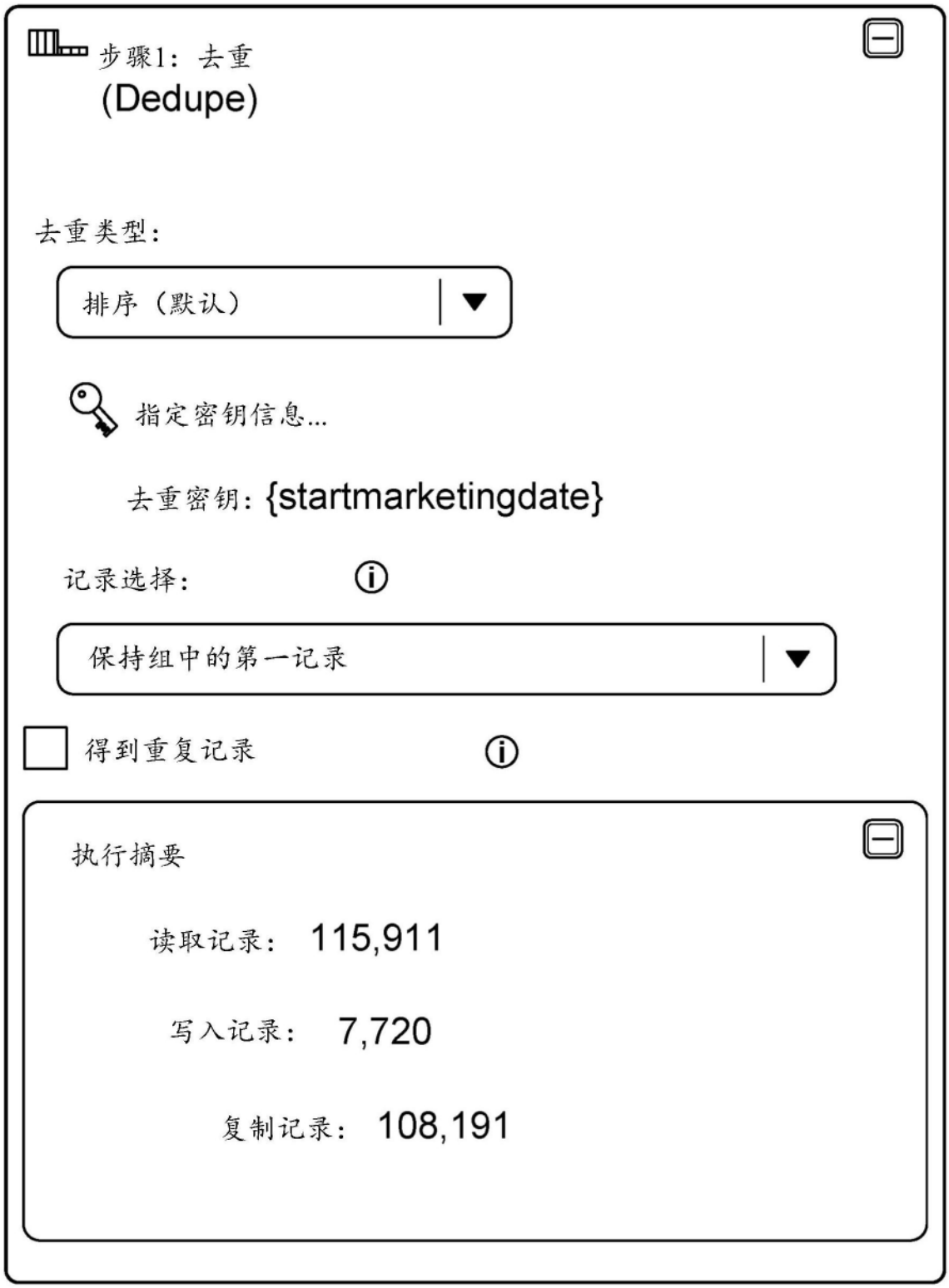

步骤1：去重
(Dedupe)
去重类型：
排序（默认）
指定密钥信息...
去重密钥：{startmarketingdate}
记录选择：
保持组中的第一记录
得到重复记录
执行摘要
读取记录： 115,911
写入记录： 7,720
复制记录： 108,191

图8

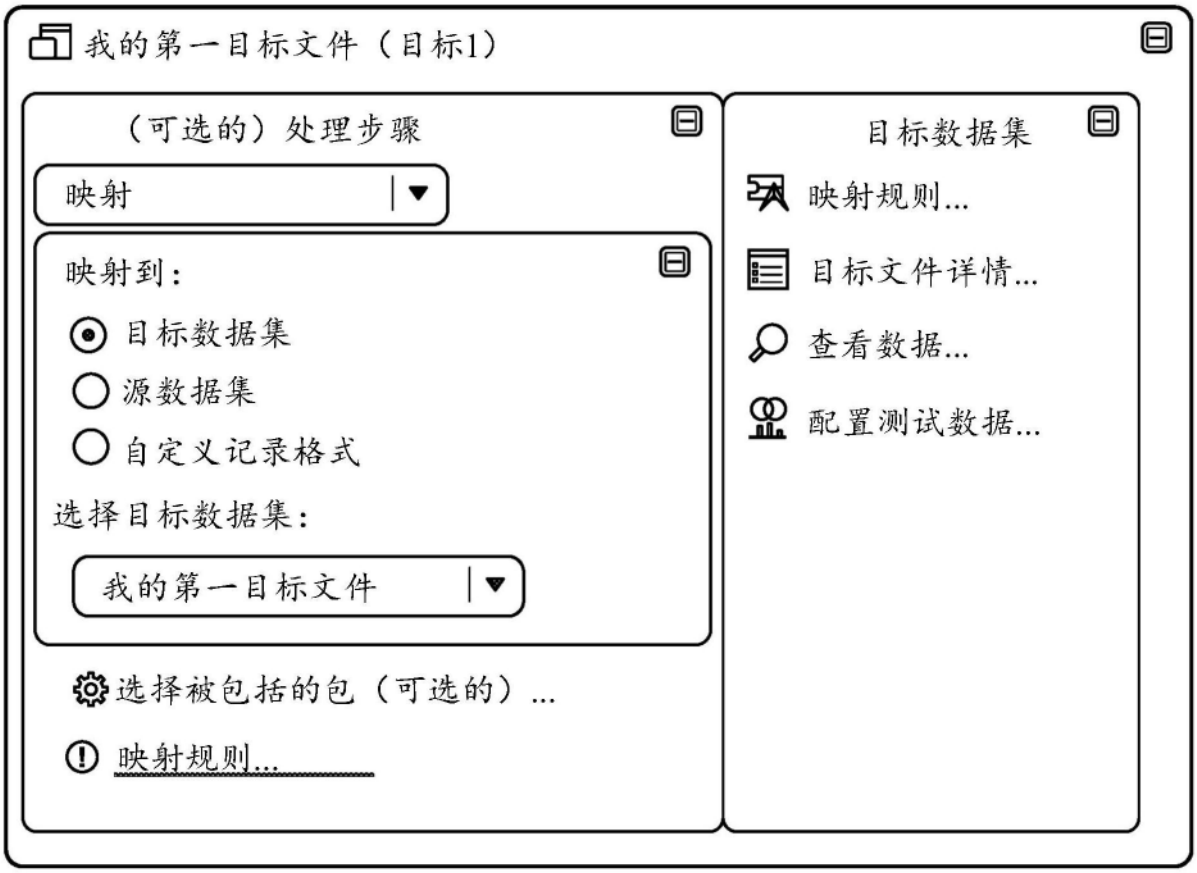

我的第一目标文件（目标1）
（可选的）处理步骤
映射
映射到：
目标数据集
源数据集
自定义记录格式
选择目标数据集：
我的第一目标文件
选择被包括的包（可选的）...
映射规则...
目标数据集
映射规则...
目标文件详情...
查看数据...
配置测试数据...

图9A

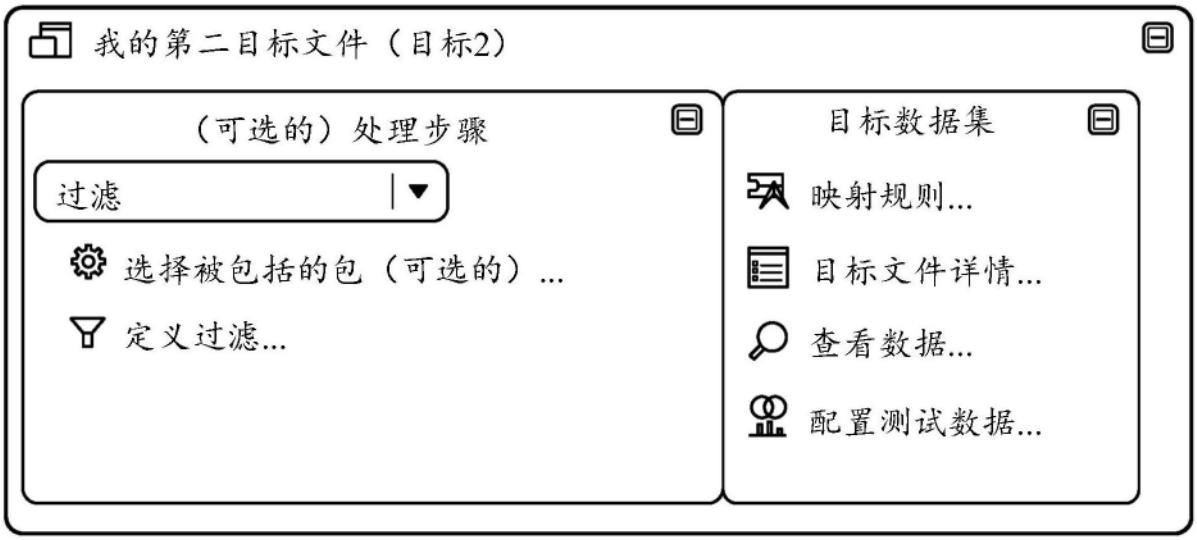

我的第二目标文件（目标2）
（可选的）处理步骤
过滤
选择被包括的包（可选的）...
定义过滤...
目标数据集
映射规则...
目标文件详情...
查看数据...
配置测试数据...

图9B

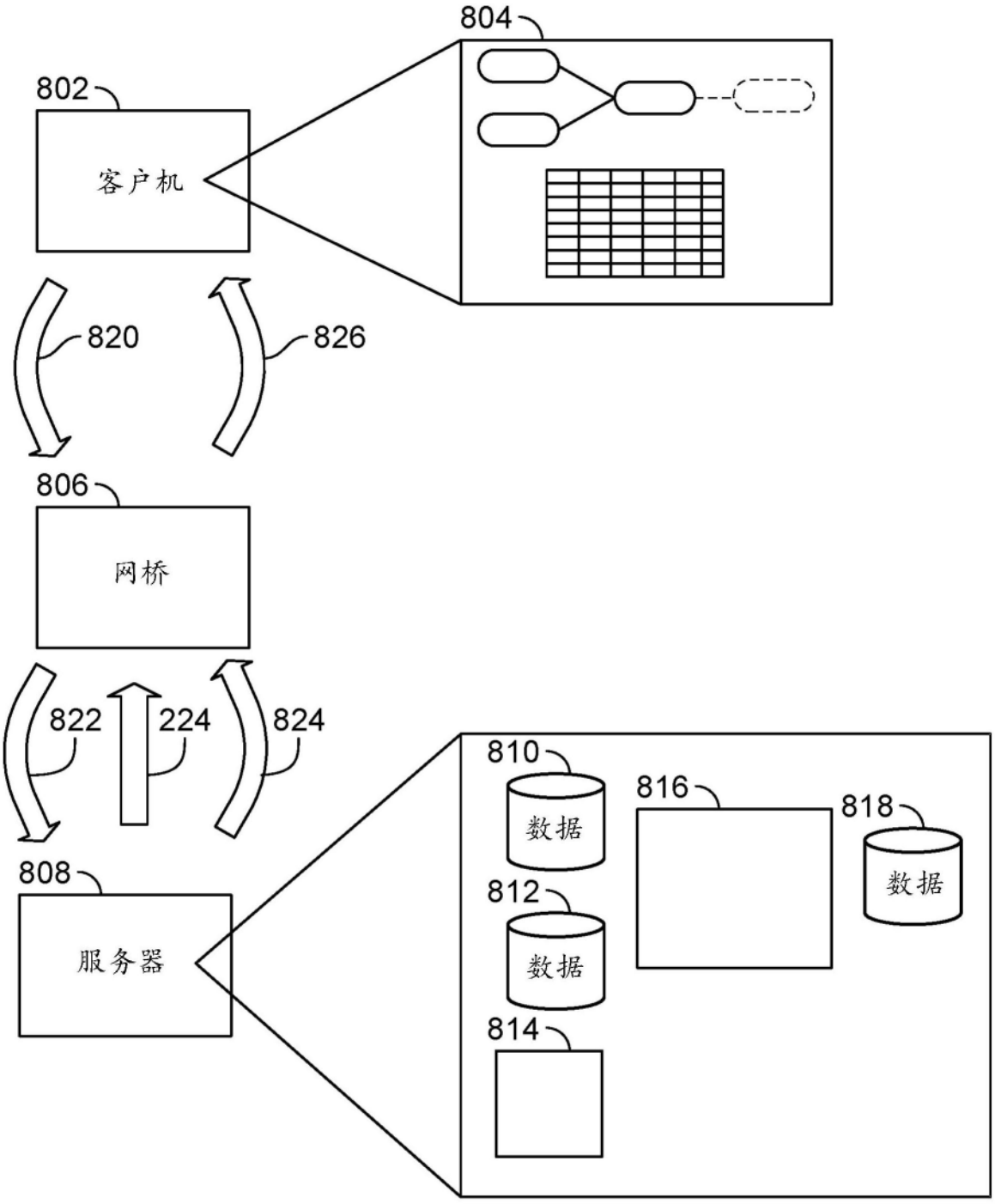

804
802
客户机
820
826
806
网桥
822
224
824
810
数据
816
818
数据
812
数据
808
服务器
814

图10