(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112965793 B
(45)授权公告日 2023.11.21

(21)申请号 202110082479.8

(22)申请日 2021.01.21

(65)同一申请的已公布的文献号
申请公布号 CN 112965793 A

(43)申请公布日 2021.06.15

(73)专利权人 中国互联网络信息中心
地址 100190 北京市海淀区中关村南四街四号1号楼

(72)发明人 邓桂英 杨学 张立坤 孙从友

(74)专利代理机构 北京君尚知识产权代理有限公司 11200
专利代理师 邱晓锋

(51)Int.Cl.
G06F 9/48(2006.01)
G06F 16/25(2019.01)

(56)对比文件
CN 102981904 A,2013.03.20
CN 104050029 A,2014.09.17
CN 108427641 A,2018.08.21
CN 111090665 A,2020.05.01
US 2014229953 A1,2014.08.14
WO 2018219480 A1,2018.12.06
CN 104965754 A,2015.10.07
CN 110597611 A,2019.12.20
CN 111190892 A,2020.05.22
US 2004078105 A1,2004.04.22
梁毅等.面向大数据流式计算的任务管理技术综述.《计算机工程与科学》.2017,第39卷(第2期),215-226.
王建民.领域大数据应用开发与运行平台技术研究.《软件学报》.2017,第28卷(第06期),1516-1528.

审查员 严丽

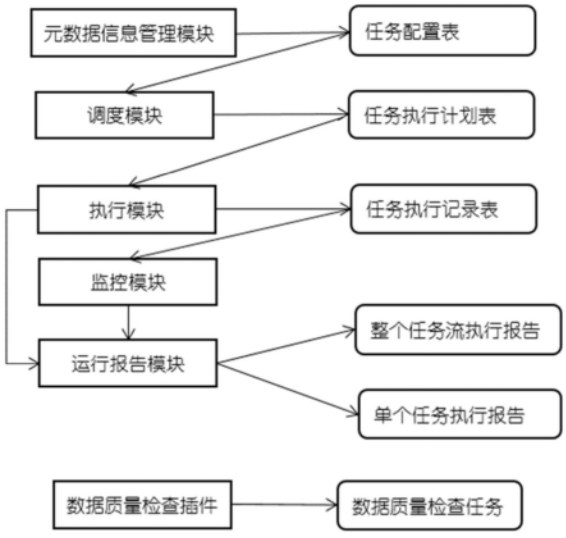
权利要求书2页 说明书6页 附图1页

(54)发明名称

一种面向标识解析数据的数据仓库任务调度方法和系统

(57)摘要

本发明涉及一种面向标识解析数据的数据仓库任务调度方法和系统。该方法的步骤包括：配置标识解析数据的数据任务的元数据信息；根据元数据信息解析生成执行任务流，执行任务流中的执行任务包含数据任务和数据时间；启动并运行执行任务流中满足执行条件的执行任务；监控执行任务的运行状态，并根据运行状态进行相应的处理；对执行任务流的运行结果进行报告。本发明提供了一种轻量级、易用性的尤其适用于标识解析数据的数据任务特点的调度方案，支持更加丰富的依赖关系，能够满足特定的数据分析场景，并引入数据质量检查环节，能够降低人工运维成本。



CN 112965793 B

1.一种面向标识解析数据的数据仓库任务调度方法,其特征在于,包括以下步骤:

配置标识解析数据的数据任务的元数据信息;

根据元数据信息解析生成执行任务流,执行任务流中的执行任务包含数据任务和数据时间;

启动并运行执行任务流中满足执行条件的执行任务;

监控执行任务的运行状态,并根据运行状态进行相应的处理;

对执行任务流的运行结果进行报告;

所述根据元数据信息解析生成执行任务流,包括:

采取轮询的方式,不断的生成最新的基于执行任务的任务流;

对于新插入的任务,及时合并到最新的任务流中;

如果需要启动历史数据修复,则将与错误的历史数据相关的首个执行任务置为待执行的状态,从而动态生成包含该首个执行任务的所有下游任务的整个执行任务流。

2.根据权利要求1所述的方法,其特征在于,所述元数据信息包括:数据周期,依赖的数据任务,依赖方式,超时时间,超时处理方案,执行命令,任务过期时间。

3.根据权利要求2所述的方法,其特征在于,所述依赖方式包括:

自依赖:数据任务的当前数据周期的统计结果依赖自己上一个数据周期的统计结果;

顺序依赖:两个不同的数据任务A和B,数据周期相同,B的统计结果依赖A的统计结果;

周期依赖:两个不同的数据任务A和B,B的统计结果依赖A的多个周期的统计结果;

混合依赖:包含自依赖、顺序依赖、周期依赖中的至少两种。

4.根据权利要求1所述的方法,其特征在于,所述监控执行任务的运行状态,并根据运行状态进行相应的处理,包括:

实时监控执行任务是否运行超时,是否失败,是否需要报警;如果超时,则根据对应数据任务的配置信息,或者将执行任务杀掉并重启,或者继续执行同时发出报警邮件。

5.根据权利要求1所述的方法,其特征在于,所述执行任务流中包含数据质量检查任务,所述数据质量检查任务完成以下操作:

统计最近若干个数据周期的数据结果,检查当前统计结果是否波动很大,波动超过设定的阈值则认定为疑似异常;

统计最近若干个数据周期的数据量,检查数据量是否波动很大,波动超过设定的阈值则认定为疑似异常;

检查数据结果中每个字段有无为空的情况,如果被检查数据不应该出现空值,则出现空说明数据异常;

检查数据结果中每个字段有无超过预期大小的情况,如果出现超过预期大小的情况说明数据异常。

6.根据权利要求1所述的方法,其特征在于,所述对执行任务流的运行结果进行报告,包括两个层面的报告:一是对整个执行任务流的运行结果的报告,二是对单个执行任务的运行结果的报告。

7.一种采用权利要求1~6中任一权利要求所述方法的面向标识解析数据的数据仓库任务调度系统,其特征在于,包括:

元数据信息管理模块,用于配置标识解析数据的数据任务的元数据信息;

调度模块，用于根据元数据信息解析生成执行任务流，执行任务流中的执行任务包含数据任务和数据时间；

执行模块，用于启动并运行执行任务流中满足执行条件的执行任务；

监控模块，监控执行任务的运行状态，并根据运行状态进行相应的处理；

运行报告模块，对执行任务流的运行结果进行报告；

所述根据元数据信息解析生成执行任务流，包括：

采取轮询的方式，不断的生成最新的基于执行任务的任务流；

对于新插入的任务，及时合并到最新的任务流中；

如果需要启动历史数据修复，则将与错误的历史数据相关的首个执行任务置为待执行的状态，从而动态生成包含该首个执行任务的所有下游任务的整个执行任务流。

8.一种电子装置，其特征在于，包括存储器和处理器，所述存储器存储计算机程序，所述计算机程序被配置为由所述处理器执行，所述计算机程序包括用于执行权利要求1~6中任一权利要求所述方法的指令。

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存储计算机程序，所述计算机程序被计算机执行时，实现权利要求1~6中任一权利要求所述的方法。

一种面向标识解析数据的数据仓库任务调度方法和系统

技术领域

[0001] 本发明属于信息技术领域，具体涉及一种面向标识解析数据的数据仓库任务调度方法和系统。

背景技术

[0002] 标识解析数据，既包括传统互联网DNS体系在运行过程中产生的解析数据，也包括工业互联网等新兴网络形态的标识体系产生的解析数据。在通过数据仓库对大规模的标识解析数据进行有效管理和深度分析挖掘的过程中，需要对数据仓库的数据任务进行科学有效地调度。

[0003] 数据仓库的数据任务具有下面的特点：1)数据任务种类众多。包括抽取、转化、清洗、备份、统计分析等。2)数据量巨大，数据任务量巨大。数据多源性高，数据并发连接多，数据种类多，数据持续性长，数据关联性高，统计指标众多，数据任务的量巨大。3)依赖关系复杂。有周期依赖又有顺序依赖，也有自己依赖自己，对掌握整体数据的拓扑关系的需求很强烈。4)对数据修复的要求比较高。

[0004] 当前现有的一些通用调度方案，例如Oozie，Azkaban和大多数公有云上的workflow服务，都是DAG工作流类调度系统。Oozie和Azkaban采取的这两种方式，从系统设计的角度来说，对外部系统的关联和依赖比较小，是一个相对独立封闭的环境，演进起来比较自由。但这两个系统最大的问题是，周边的运维使用工具太过缺乏，易用性很差。作为工具使用可以，但是做为平台服务，缺失了太多内容，工作流的定义和维护成本太高。

发明内容

[0005] 本发明针对上述问题，提供一种轻量级、易用性的尤其适用于标识解析数据的数据任务特点的调度方法和系统，支持更加丰富的依赖关系，以满足特定的数据分析场景，并引入数据质量检查环节，降低人工运维成本。

[0006] 本发明采用的技术方案如下：

[0007] 一种面向标识解析数据的数据仓库任务调度方法，包括以下步骤：

[0008] 配置标识解析数据的数据任务的元数据信息；

[0009] 根据元数据信息解析生成执行任务流，执行任务流中的执行任务包含数据任务和数据时间；

[0010] 启动并运行执行任务流中满足执行条件的执行任务；

[0011] 监控执行任务的运行状态，并根据运行状态进行相应的处理；

[0012] 对执行任务流的运行结果进行报告。

[0013] 进一步地，所述元数据信息包括：数据周期，依赖的数据任务，依赖方式，超时时间，超时处理方案，执行命令，任务过期时间。

[0014] 进一步地，所述依赖方式包括：

[0015] 自依赖：数据任务的当前数据周期的统计结果依赖自己上一个数据周期的统计结

果；

[0016] 顺序依赖：两个不同的数据任务A和B，数据周期相同，B的统计结果依赖A的统计结果；

[0017] 周期依赖：两个不同的数据任务A和B，B的统计结果依赖A的多个周期的统计结果；

[0018] 混合依赖：包含自依赖、顺序依赖、周期依赖中的至少两种。

[0019] 进一步地，所述根据数据任务的元数据信息，解析生成执行任务流，包括：

[0020] 采取轮询的方式，不断的生成最新的基于执行任务的任务流；

[0021] 对于新插入的任务，及时合并到最新的任务流中；

[0022] 如果需要启动历史数据修复，则将与错误的历史数据相关的首个执行任务置为待执行的状态，从而动态生成包含该首个执行任务的所有下游任务的整个执行任务流。

[0023] 进一步地，所述监控执行任务的运行状态，并根据运行状态进行相应的处理，包括：

[0024] 实时监控执行任务是否运行超时，是否失败，是否需要报警；如果超时，则根据对应数据任务的配置信息，或者将执行任务杀掉并重启，或者继续执行同时发出报警邮件。

[0025] 进一步地，所述执行任务流中包含数据质量检查任务，所述数据质量检查任务完成以下操作：

[0026] 统计最近若干个数据周期的数据结果，检查当前统计结果是否波动很大，波动超过设定的阈值则认定为疑似异常；

[0027] 统计最近若干个数据周期的数据量，检查数据量是否波动很大，波动超过设定的阈值则认定为疑似异常；

[0028] 检查数据结果中每个字段有无为空的情况，如果被检查数据不应该出现空值，则出现空说明数据异常；

[0029] 检查数据结果中每个字段有无超过预期大小的情况，如果出现超过预期大小的情况说明数据异常。

[0030] 进一步地，所述对执行任务流的运行结果进行报告，包括两个层面的报告：一是对整个执行任务流的运行结果的报告，二是对单个执行任务的运行结果的报告。

[0031] 一种采用上述方法的面向标识解析数据的数据仓库任务调度系统，其包括：

[0032] 元数据信息管理模块，用于配置标识解析数据的数据任务的元数据信息；

[0033] 调度模块，用于根据元数据信息解析生成执行任务流，执行任务流中的执行任务包含数据任务和数据时间；

[0034] 执行模块，用于启动并运行执行任务流中满足执行条件的执行任务；

[0035] 监控模块，监控执行任务的运行状态，并根据运行状态进行相应的处理；

[0036] 运行报告模块，对执行任务流的运行结果进行报告。

[0037] 本发明的关键点是：

[0038] 1)执行任务的重新定义：基于标识解析数据的数据任务有个关键因素是数据时间dt，数据时间是分析和统计标识解析数据指标的重要维度。执行任务＝数据任务+数据时间dt，数据时间dt是可变参数，可以指定任意一天/小时/月等，执行任务是数据时间dt确定取值后的数据任务，参与调度的元素是执行任务而不是数据任务。

[0039] 2)支持数据任务自依赖方式。任务自依赖指的是有些数据任务，要统计当前数据

周期的结果,需要依赖自己上一个数据周期的结果。

[0040] 3)数据质量检查任务纳入调度系统,数据质量检查任务具备自己的特有特性,有三种任务结果:完成(done),错误(error),待定(uncertain)。对于待定的情况,可以选择继续执行下游任务,也可以终止下游任务。

[0041] 由于采用了以上的方案,本发明具有以下优点:

[0042] 1)区分数据任务和执行任务,执行任务=数据任务+数据时间,调度系统是针对执行任务的调度,便于更加灵活的调度业务逻辑复杂的数据任务。

[0043] 2)提供了丰富的任务依赖关系,能够满足数据仓库建设中各种任务依赖情况。

[0044] 3)将数据质量检查任务纳入调度DAG工作流。

附图说明

[0045] 图1是实施例中执行任务流的有向无环图示意图。

[0046] 图2是实施例中任务调度系统的任务调度流程图。

具体实施方式

[0047] 为使本发明的上述目的、特征和优点能够更加明显易懂,下面通过具体实施例和附图,对本发明做进一步详细说明。

[0048] 本发明的一实施例提出的一种面向数据仓库的数据任务调度系统,包括以下模块或子系统。

[0049] 1)元数据信息管理模块(meta):该模块主要通过服务接口的方式,提供给用户配置标识解析数据的数据任务的元数据信息的功能,形成任务配置表。任务元数据信息包括下面几个方面:

[0050] a)配置数据任务的元数据信息的接口:

[0051] 元数据信息包括,数据周期(5min,10min,小时,天,周,月,半年,年),依赖的数据任务(job),依赖方式,超时时间,超时处理方案(杀掉任务,或者继续等待,是否发报警),执行命令,任务过期时间。

[0052] 数据周期:数据周期是指数据指标的周期,数据周期可以是:分钟,小时,天,周,季度,半年,年。

[0053] 依赖的数据任务:“依赖”是指,如果数据任务A的统计结果需要基于数据任务B的结果,则A依赖B。

[0054] 依赖方式:依赖方式是指数据任务对上游数据任务的依赖需求。包括以下方式:

[0055] 自依赖:自依赖是指数据任务当前数据周期的统计结果依赖自己上一个数据周期的统计结果,表示为A_dt-1--->A_dt,其中A_dt-1表示上一个数据周期的数据任务A,A_dt表示当前数据周期的数据任务A,--->表示箭头右边的数据任务依赖箭头左边的数据任务。

[0056] 顺序依赖:顺序依赖是指两个不同的数据任务A和B,数据周期相同,数据任务B的统计结果依赖数据任务A的统计结果,表示为A_dt--->B_dt,其中A_dt表示当前数据周期的数据任务A,B_dt表示当前数据周期的数据任务B。

[0057] 周期依赖:周期依赖是指两个不同的数据任务A和B,B的统计结果依赖A的多个周期的统计结果,比如B的每天的统计结果依赖于A的前七天的统计结果,表示为A_dt-n～A_

dt-->B_dt,其中A_dt-n～A_dt表示当前数据周期以及往前n个数据周期,是n个任务的组合。

[0058] 混合依赖:混合依赖是指包含前面所说的自依赖,顺序依赖,周期依赖中的至少2种,表示为A_dt-n～A_dt,B_dt--->C_dt,其中其中A_dt-n～A_dt表示当前数据周期以及往前n个数据周期,是n个任务的组合,B_dt表示当前数据周期的数据任务B,C_dt是当前数据周期的数据任务C。

[0059] 任务过期时间:对于定义了最晚运行时间的任务,认为是过期未执行的任务,对于定义了最晚成功结束的时间的任务,认为是过期未成功的任务。比如有些任务生成的数据需要每天早晨8点呈现在审核者面前,则需要设置过期未成功的时间,及时干预以确保数据按时生成。

[0060] b)检查任务上下游的接口:

[0061] 支持2种,只显示任务直接上游和任务直接下游,或者显示所在的整个任务流。

[0062] c)修改任务运行状态的接口:

[0063] 如果遇到极端的情况,整个调度系统崩溃,通过强行修改任务运行状态,重新纳入调度执行。

[0064] 2)调度模块(scheduler):

[0065] 根据配置文件提供的数据任务的元数据信息,解析生成当下的执行任务流,形成任务执行计划表。执行任务＝数据任务+数据时间。调度模块采取轮询的方式,不断的生成最新的基于执行任务的任务流;对于新插入的任务,也可以及时合并到最新的任务流中;如果需要启动历史数据修复(即发现历史数据有错误需进行修复),可以将与错误的历史数据相关的首个执行任务置为待执行(todo)状态,即可动态生成包含该首个执行任务的所有下游任务的整个执行任务流。

[0066] 其中,配置文件的格式如下:

[0067] 数据任务名称＝”A”

[0068] 数据周期＝”天”

[0069] 执行任务流实际上是个有向无环图,如图1所示。

[0070] 3)执行模块(executor):

[0071] 如果某个执行任务满足执行条件,即状态为待执行(todo)状态,则由执行模块进行启动执行任务,并记录执行任务的开始时间start_time,状态设置为运行(running)。其中满足执行条件是指,如果任务配置表里设置该任务是某个时间点启动,则当系统时间为启动时间点时就为满足执行条件,如果任务配置表里设置的该任务还有上游依赖任务,则所有上游任务的状态都是完成(done)时就为满足执行条件。对于监控模块扫描发现已经超时,需要杀掉(kill)的执行任务,也通过执行模块对任务进行杀掉,或者是杀掉并且重启任务。对于成功结束的执行任务,执行模块负责将任务状态修改为完成(done),并且记录结束时间end_time。执行模块也负责把所有运行的任务执行日志记录入任务执行记录表,用于问题查找和其他分析。

[0072] 4)监控模块(monitor):

[0073] 扫描每个运行的执行任务是不是超时,如果超时则根据对应数据任务的配置信息,或者通知执行模块将执行任务杀掉并重启,或者通知运行报告模块发出报警邮件。对于

那些设置了最晚执行时间或者最晚完成时间的执行任务，扫描其完成情况，超时则调用运行报告模块发出报警邮件。

[0074] 5)运行报告模块(reporter)：

[0075] 该模块属于被其他模块调用执行的模块，包括2个层面的邮件报告和短信报警。一个层面是整个工作流的整体运行状态的汇总报告，这个层面的报告一般是发给调度系统的运维人员。另外一个层面是单个任务级的运行状态，发给对应的负责人。分级报告有利于既能做到及时报告和发现问题，又能避免邮件过多，漏掉发现问题。运行报告模块中，有默认的报告模板，用户定义统计方式，如果不定义，则报告邮件中没有这部分内容，同时也可以复用数据检查任务的结果。任务运行状态报告的示例如表1所示。

[0076] 表1.任务运行状态报告

[0077]

运行状态	数据任务	数据量(行数)			存储量(M)		
		20200830（本次）	20200829（上期）	变化率	20200830（本次）	20200829（上期）	变化率
done	dwd_cn_info_daily	21824213	21832652	-0.04%	2115.29	2116.698	-0.07%
done	dwa_higncnt_daily	29181	27596	5.74%	1.462	1.445	1.19%

[0078] 6)数据质量检查任务(数据质量检查插件)：

[0079] 本发明的执行任务流中可以包含数据质量检查任务，也可以不包含数据质量检查任务。数据质量检查任务具备自己的特有特性，有三种任务结果：完成(done)，错误(error)，待定(uncertain)。对于待定的情况，可以选择继续执行下游任务，也可以终止下游任务。数据质量检查任务可以由开发人员自己开发，同时本调度系统也支持数据质量检查插件，该插件是通过抽取整理一些共性的数据质量检查方式，提供一些通用函数或者方法，供数据开发人员直接配置生成数据质量检查任务。而数据质量检查任务可以直接配置到任务依赖中，作为数据真正就绪的最后一道关卡。

[0080] 本实施例的采用上述各模块实现的任务调度方法的流程如图2所示。其步骤描述如下：

[0081] 1)元数据信息管理模块meta，收集记录数据任务的元数据信息，形成任务配置表。同时会对是否存在数据任务依赖环进行判断。

[0082] 2)调度模块scheduler以轮询的方式，负责根据数据任务的元数据信息，解析生成当下的执行任务流，形成任务执行计划表，同时对已经生成的执行任务流做动态调整。对于需要启动历史数据修复的执行任务流，可以将首个任务置为待执行状态，即可动态生成包含所有下游任务的整个执行任务流。

[0083] 3)对于满足启动执行条件的执行任务，调用执行模块executor执行，把所有运行的任务执行日志记录入任务执行记录表。并通过监控模块monitor实时监控执行任务运行结果。

[0084] 4)监控模块monitor实时监控执行任务是否运行超时，是否失败，是否需要报警。

[0085] 5)运行报告模块reporter负责对整个执行任务流运行状态的运行结果进行汇报。包括两个层面的报告，一是整个执行任务流的运行结果统一报告，二是对单个执行任务的运行结果的报告。

[0086] 6)执行任务流中可以包含数据质量检查任务,可以通过数据质量检查插件实现,主要包含如下功能:

[0087] a)统计最近几个数据周期(默认是7)的数据结果,检查当前统计结果是否波动很大,波动超过20%则认定为疑似异常。

[0088] b)统计最近几个数据周期(默认是7)的数据量(行数),检查数据量是否波动很大,波动超过20%则认定为疑似异常。

[0089] c)检查数据结果中每个字段有无为空的情况,如果被检查数据不应该出现空值,则出现空说明数据异常。

[0090] d)检查数据结果中每个字段有无超过预期大小的情况,比如有些字段预期是128字节长度,如果出现超过128字节的情况说明数据异常。

[0091] 基于同一发明构思,本发明的另一实施例提供一种电子装置(计算机、服务器、智能手机等),其包括存储器和处理器,所述存储器存储计算机程序,所述计算机程序被配置为由所述处理器执行,所述计算机程序包括用于执行本发明方法中各步骤的指令。

[0092] 基于同一发明构思,本发明的另一实施例提供一种计算机可读存储介质(如ROM/RAM、磁盘、光盘),所述计算机可读存储介质存储计算机程序,所述计算机程序被计算机执行时,实现本发明方法的各个步骤。

[0093] 以上公开的本发明的具体实施例,其目的在于帮助理解本发明的内容并据以实施,本领域的普通技术人员可以理解,在不脱离本发明的精神和范围内,各种替换、变化和修改都是可能的。本发明不应局限于本说明书的实施例所公开的内容,本发明的保护范围以权利要求书界定的范围为准。

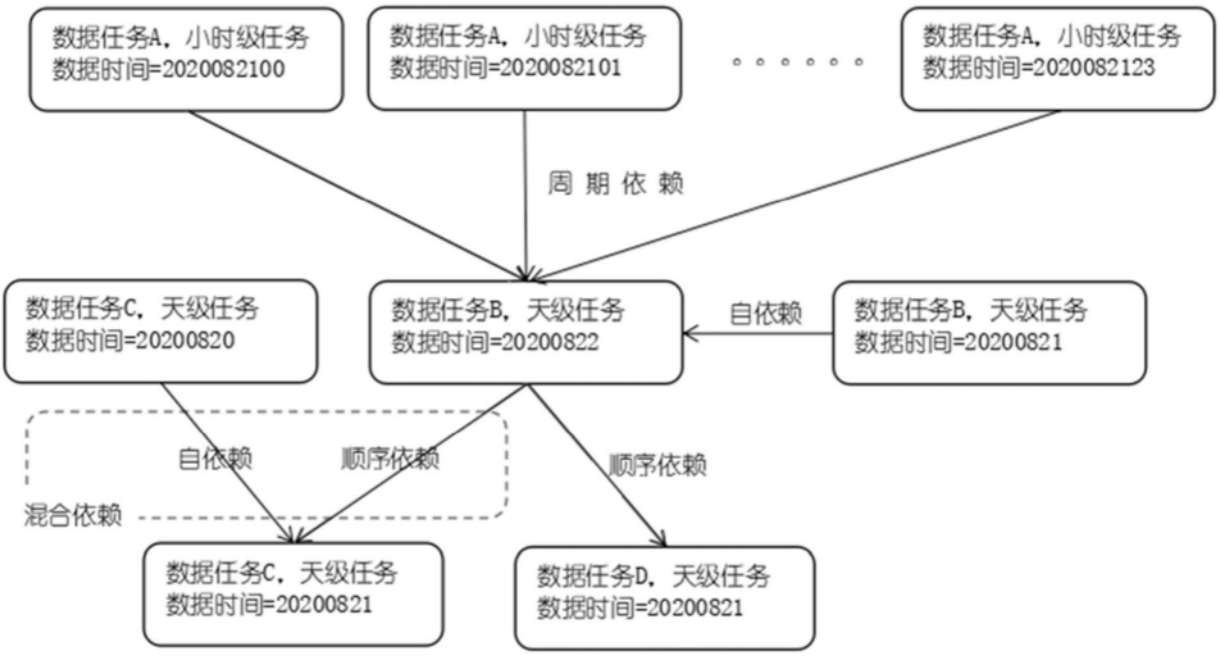

数据任务A，小时级任务
数据时间=2020082100
数据任务A，小时级任务
数据时间=2020082101
数据任务A，小时级任务
数据时间=2020082123
周 期 依 赖
数据任务C，天级任务
数据时间=20200820
数据任务B，天级任务
数据时间=20200822
自依赖
数据任务B，天级任务
数据时间=20200821
自依赖
顺序依赖
混合依赖
顺序依赖
数据任务C，天级任务
数据时间=20200821
数据任务D，天级任务
数据时间=20200821

图1

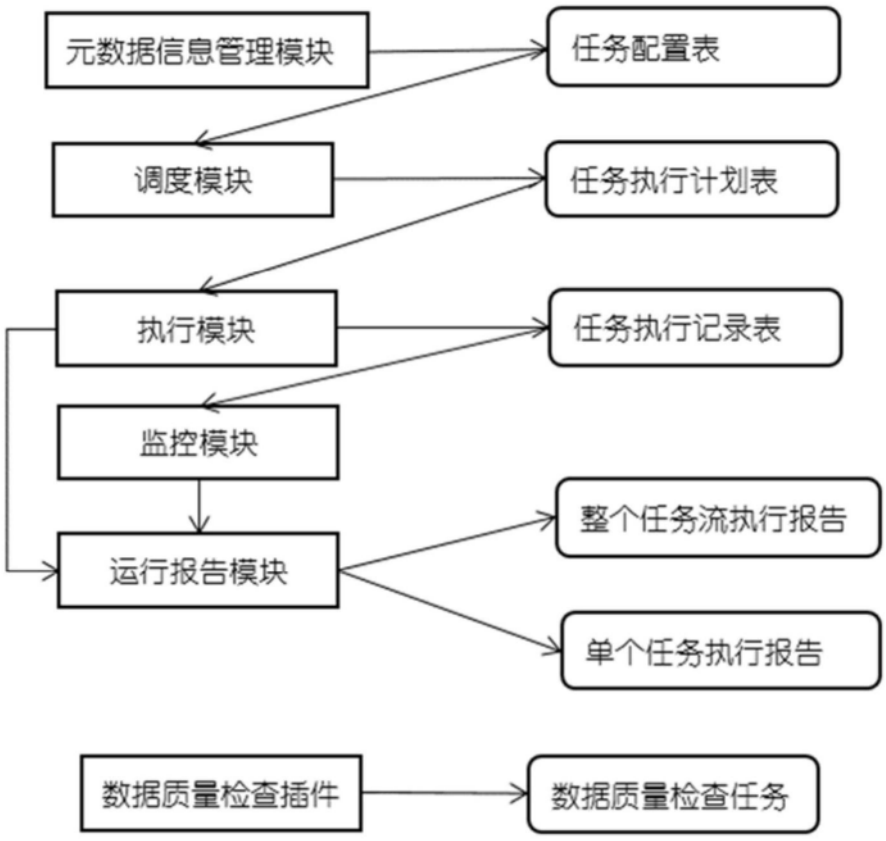

元数据信息管理模块
任务配置表
调度模块
任务执行计划表
执行模块
任务执行记录表
监控模块
整个任务流执行报告
运行报告模块
单个任务执行报告
数据质量检查插件
数据质量检查任务

图2