文章编号：2096-1472(2023)-02-20-04
DOI:10.19644/j.cnki.issn2096-1472.2023.002.004

基于可视化元数据配置的大数据治理方案

郑响萍

(浙江理工大学科技与艺术学院，浙江 绍兴 312369)
✉xiangping2002@163.com

摘　要：当下互联网运营竞争愈发激烈，中小企业急需大数据系统支持业务运营，但普遍缺乏大数据人才，并且实施大数据营销的成本较高。针对中小企业的业务场景，设计一套可视化的大数据治理方案，通过流程调度系统和大数据计算平台，实现大数据模型自动化计算，同时利用容器化技术实现快速部署，满足大数据场景。方案全部采用开源工具，大大降低了使用和维护成本，平台扩展性强，系统稳定性高，支持数据可视化，投入较低的成本就能满足中小企业的大数据运营需求。

关键词：可视化；配置；大数据；治理；运营
中图分类号：TP311　**文献标识码**：A

Big Data Governance Solution based on Visual Metadata Configuration

ZHENG Xiangping

(*Keyi College of Zhejiang Sci-tech University, Shaoxing* 312369, *China*)
✉xiangping2002@163.com

Abstract: With the increasingly fierce competition in Internet business at present, small and medium-sized enterprises are in urgent need of big data system to support business operation, but big data talents are generally in shortage and the cost of implementing big data marketing is high. Aiming at the business scenarios of small and medium-sized enterprises, this paper proposes to design a set of visualized big data governance solutions. The automatic calculation of big data models is realized through the process scheduling system and big data computing platform. At the same time, the container technology is used to realize rapid deployment to meet the big data scenarios. Open source tools are used in all solutions, which greatly reduces the use and maintenance costs. The platform has strong scalability, and the highly stable system supports data visualization. Big data operation needs of small and medium-sized enterprises are, therefore, met at a lower cost.

Keywords: visualization; configuration; big data; governance; operation

1 引言(Introduction)

全球数据量正飞速增长，据数据统计互联网公司Statistat统计预测，2020 年全球数据存储量已达到47 ZB，2035 年将达到2,142 ZB，目前企业运营中产生的数据以每年42.2%的速度快速增长，但是只有56%数据能被企业获取，而在获取的数据中也仅有57%的数据会被有效利用。2016 年《国家“十三五”时期文化发展改革规划纲要》正式提出，大数据发展进入深化阶段，2021 年国家把大数据列入《中华人民共和国国民经济和社会发展第十四个五年规划和2035 年远景目标纲要》中的重要一环，足见国家对大数据的重视[1]。近年来，大数据技术的发展日新月异，但是针对中小企业业务场景的大数据解决方案较少，并且实施成本高。

本文提出一种可视化配置的大数据治理方案，主要能解决中小企业使用大数据平台成本高的问题。企业大数据通常有“3V”属性，即高速度(Velocity)、多样性(Variety)和大体量(Volume)[2]，目前使用较多是Hadoop体系架构，Hadoop可以较好地解决“3V”属性带来的存储和计算难题，但Hadoop体系架构维护成本较高，并且日常云运营对专业大数据技术

基金项目：浙江理工大学科技与艺术学院院设科研项目资助(KY2022002).

人员的依赖程度高。本文的研究重点是通过可视化配置的方式，使得非大数据技术人员也可以方便地进行大数据分析。

2 方案简介(Solution introduction)

本文设计的大数据平台包括数据采集、数据处理、数据服务和应用服务四大模块。参考通用大数据架构设计和MPP架构[3]，将数据存储、数据处理和数据应用服务分开，实现元数据配置、数据模型可视化及数据处理流程自动化[4]。

数据处理流程包括数据的定义和采集、数据清洗和模型存储、数据分析和打标，以及标签与业务系统结合等操作步骤，通过流程、模型定义等的配置，满足企业的个性化业务场景需求，系统概览如图1所示。

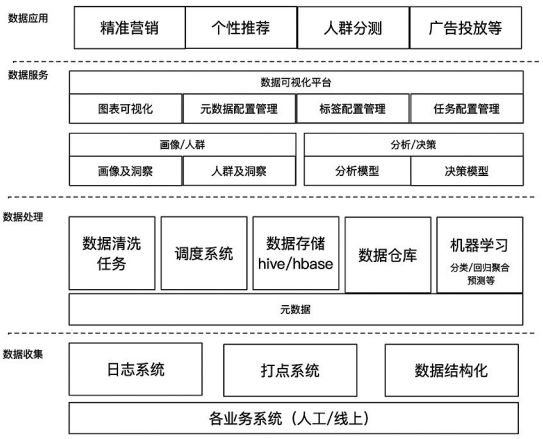

图1 系统概览

Fig.1 System overview

业界已有不少大数据平台解决方案，例如国内各大云服务商的数据治理平台、开源平台Hadoop体系。这些平台解决方案相较于本方案，云服务使用价格高，开源平台Hadoop体系的运行对专业技术人才的依赖程度更高，并且使用成本、维护成本也比较高，很难在中小企业中得到快速普及使用。本方案通过可视化配置组合开源工具，对可视化大数据架构和容器化进行了深入研究，弥补了大数据过于依赖专业技术人才和使用成本高等问题。各大数据平台解决方案比较如表1所示。

表1 大数据方案比较

Tab.1 Comparison of big data solutions

对比项	云服务数据治理平台	Hadoop体系	本文方案
使用成本	成本高	成本高，对专业技术人才依赖程度高	成本低，对专业技术人才依赖程度低
维护成本	维护成本低	系统搭建复杂，维护成本高	实现容器化和脚本化，维护成本低
数据开发和治理	支持可视化建模，支持快速数据查询，支持数据计算服务编排	不支持数据可视化，支持数据查询复杂，不支持服务编排	支持可视化建模，支持离线和实时，数据查询，支持数据计算服务编排
数据存储	扩展性强，成本高	扩展性强，成本低	扩展性强，成本低

从表1可以看出，中小企业最关心的几个指标为使用成本、维护成本、服务能力等，本文提出的方案都能较好地满足。

3 系统架构(System architecture)

系统采用开源方案，不额外增加企业成本，以自动化和可视化为前提，尽量降低企业对大数据专业技术人才的依赖程度。系统使用的开源工具包括Spark、Hive、MySQL、Snowplow等实现，系统架构图如图2所示。

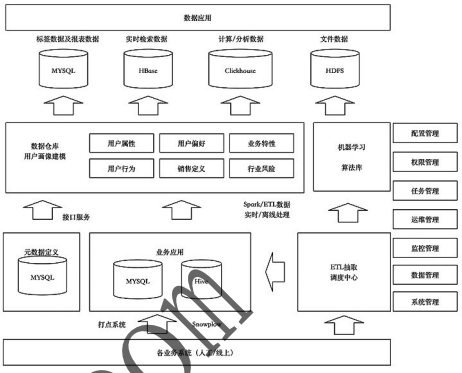

图2 系统架构图

Fig.2 System architecture diagram

数据采集层Snowplow作为业务系统的数据采集工具，其为自动化数据流而设计，通过API管理数据结构定义，可提升采集数据的质量，减少无效数据带来的成本。Snowplow通过Kafka将数据传输到后端存储。后端存储选型为Hive，考虑到采集数据的多样性，KV存储特性能有效支持Snowplow自动采集的数据。

需采集的原始数据模型通过元数据定义描述，元数据被存储到关系型数据库MySQL中，通过Echarts和Vue等前端技术实现元数据的可视化配置。业务系统通过API获取事件的元数据定义，构建采集的数据结构，将事件数据填充好并通过采集器传入Snowplow采集器中。采集到的数据将被Snowplow采集器传入Kafka中，通过消息清洗平台ETL调度中心，将Kafka中数据消费并进行结构化处理后再次保存到Hive中，即可完成原始数据的存储。

ETL任务流交由调度中心配置，数据模型由元数据定义描述，ETL任务将原始数据作为输入源，与元数据定义的输出数据进行映射，实现数据清洗的自动化，ETL清洗处理完的数据将被保存到Hive中。工作人员可通过可视化报表系统快速获得模型数据，实现用户画像、业务模型与报表的快速实时获取。

系统按数据处理流程共分为基础服务、数据采集、数据处理、数据服务四大子系统。

3.1 基础服务

基础服务包括元数据管理和任务调度两大系统。

元数据管理系统贯穿整个流程，包括数据采集时元数据配置、业务模型元数据定义等，在数据采集、数据清洗、数据建模期间都需定义数据模型绑定关系。

元数据管理系统采用微服务架构，通过Vue前端技术和SpringBoot后端技术实现元数据的配置功能，进行可视化的元数据定义管理[4-6]。元数据定义存储在MySQL中，并在Redis中缓存备份，以提升响应速度。元数据定义根据场景分为数据采集元数据定义、数据清洗元数据定义和业务模型元数据定义三大模块。

以业务模型元数据定义为例(表2)，定义了元数据字段Order.Price，该字段含义为订单金额，数据来源是trade表的price字段。

表2 业务模型元数据定义表

Tab.2 Business model metadata definition table

元数据定义	元数据编号	元数据字段类型	元数据所属模型	字段名	类型	存储定义
订单金额	1001	double	Order:100	price	业务模型	Hive:Order
交易时间	1002	date	Order:100	trade_time	业务模型	Hive:Order

元数据定义完成后，通过调度任务完成数据清洗和构建数据模型工作。将元数据定义与ETL任务进行绑定，绑定信息包括任务输入、输出及流程规则(表3、表4)，定义了订单交易数据任务清单。

表3 元数据任务绑定表

Tab.3 Metadata task binding table

模型编号	元数据编号	元数据模型	元数据字段名	ETL任务
100	1001	order	price	Trade/Task
100	1002	order	trade_time	Trade/Task
100	1003	order	month_price	Trade/Task

表4 TradeTask任务定义表

Tab.4 TradeTask definition table

任务编号	上级任务	定义	规则	输入	输出	存储
T0	null	提取订单原始数据	{‘command’:‘Fetch’,‘Data’:‘order’,‘col’:‘all’}	null	Obj:Order	Mq:tmp_1312123
T1	T0	按月汇总交易额	{‘command’:‘Sum’,‘col’:‘order.price’,‘rul’:‘mouth_of’}	Obj:Order	KV	Mq:tmp_1312124
T2	T0	按月汇总订单数	{‘command’:‘Sum’,‘col’:‘order.id’,‘rul’:‘mouth_of’}	Obj:Order	KV	Mq:tmp_1312125
T3	T1，T2	汇总数据存储	{‘command’:‘merge’}	KV	KV	Mq:tmp_1312125
T4	T3	存储	{‘command’:‘store’,‘type’:‘Hbase’,‘table’:‘tmp_trade_mdl_21323’}	KV	file	Hbase:tmp_trade_mdl_21323

调度系统通过定义表的映射关系创建ETL任务队列，任务自动获取数据，按流程处理数据。ETL任务调度系统参考业界流式数据清洗架构，并在此基础上进行优化，将元数据管理和调度系统结合，整合Spark、ClickHouse及MQ等技术[7]。规则字段定义的mouth_of等模块，由Clickhouse、Hive平台的能力支持，包装成ETL任务通用计算模块。

任务系统是一个集群，由Zookeeper选举获取Master节点，其余为Worker节点。Master节点负责编排、调度和分发，确保任务的执行均衡，Worker节点负责执行任务。Master的任务编排模块会梳理任务执行链路、任务类型等，整理出任务流程，分发到任务执行平台执行任务。Worker节点获取原始数据定义，并自动从MySQL、Hive等平台中获取原始数据，依次在Master节点的指挥下并行或串行完成任务链[8]。任务系统架构图如图3所示。

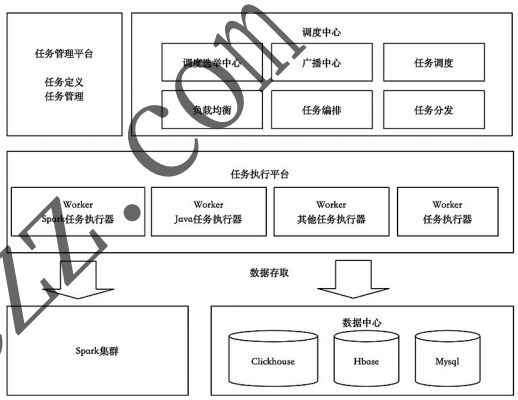

图3 任务系统架构图

Fig.3 Task system architecture diagram

将调度系统与大数据存储、计算平台结合，构建可视化的配置系统和数据报表系统，完成从原始数据、数据清洗到大数据运算结果的流程自动化。考虑到系统运维服务搭建的复杂度，对系统创建docker镜像，通过docker容器化管理工具快速完成部署[9-10]。

3.2 数据采集系统

数据采集包括采集工具和管理系统两个部分组成。

采集工具选型Snowplow数据采集器，在元数据管理中配置好数据采集定义后，Snowplow可以通过API获取最新定义的采集元数据定义，控制客户端采集数据模型，采集的数据暂存到Kafka缓冲区中，等待数据处理层处理，采集过程如图4所示。

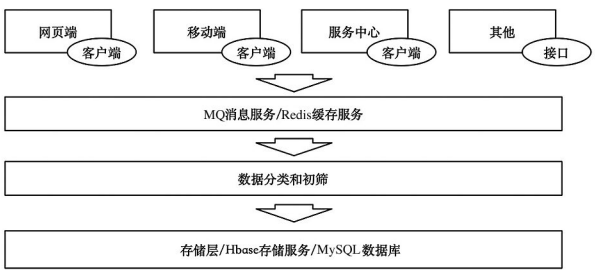

图4 数据采集流程

Fig.4 Data collection process

采集管理系统提供管理界面，系统通过SpringBoot微服务和Vue实现模块管理，实现上传Excel、导入数据及管理采集元数据定义等功能，采集数据绑定如图5所示。

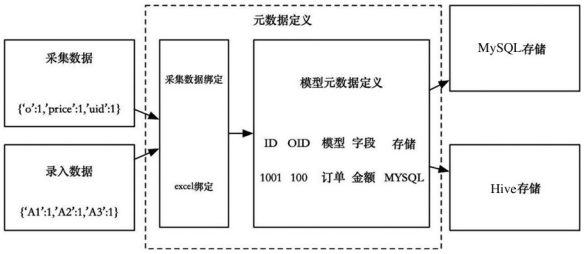

图5 采集数据绑定

Fig.5 Binding of collected data

3.3 数据处理系统

采集数据暂存在Kafka中，由数据处理系统接收并处理后，输出模型数据并持久化存储。

数据处理系统依赖基础服务的任务调度系统，通过运算模块完成数据处理。任务模块主要包括二元和多元运算、算法平台实现等，任务配置通过SpringBoot微服务实现，Vue实现前端可视化的绑定配置，例如订单交易额为原始数据，而用户的历史累计交易、单月交易额等需要多元运算，例如表4中规则字段定义的mouth_of等方法，包装Clickhouse、Hive平台查询语法，沉淀为通用运算。元数据存储在关系型数据库中，最终经过数据处理系统实现原始数据模型及二次数据模型的运算和存储。数据存储在Hive和Clickhouse等持久化平台中，以便数据服务系统进一步实现用户画像、标签化等[11-12]。数据处理链路图如图6所示。

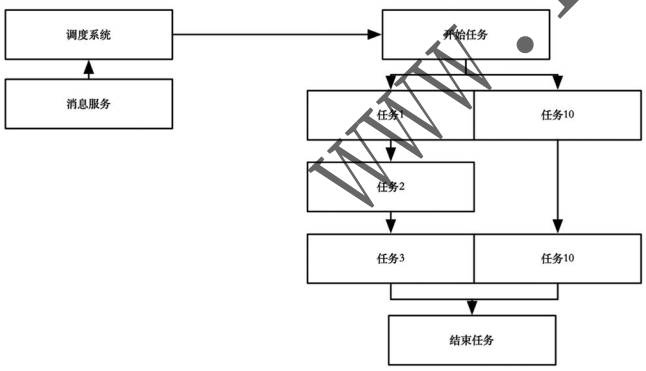

图6 数据处理链路图

Fig.6 Data processing link diagram

3.4 数据服务系统

数据服务系统为面向业务人员使用的系统，将数据处理系统完成的模型通过数据可视化报表和表格等形式展示给业务人员。

常用的可视化工具为Highcharts、Echarts、D3等，本文数据服务系统选择Echarts数据可视化图表库，原因是从兼容性角度考虑，Echarts兼容IE9及所有主流浏览器且开源免费，支持较多图表类型，可封装成通用组件，并且Apache官网自带有编辑工具，可快速完成编程。Highcharts的使用是需要收费的，D3虽然编程灵活，但是操作复杂。

数据服务系统构建可视化的数据报表供业务人员选择，组件包括数据表格、折线图、柱状图等，将数据处理系统完成的模型数据定义绑定到以上可供选择的组件中，数据可视化组件封装了从Hive、Clickhouse等平台自动获取模型数据的功能，通过简单配置就可完成数据可视化。

数据服务系统整合了数据处理系统和基础服务平台能力，可以配置数据采集、数据处理的元数据定义和任务定义链，完成数据清洗到模型的配置过程。

业务人员可以利用数据处理系统可视化配置模型的能力，构建一套数据指标体系及创建用户画像、交易模型等业务模型。业务人员可利用数据标签指导商业活动，例如构建用户画像标签后，根据场景圈定不同的标签人群做具体业务场景的商业化服务，可圈定标签为某地域组合、某时间段、交易额在一定范围的多个人群组进行下一阶段的精准营销，可以创建多组分组测试数据，用于判断哪个商业化行为更有优势。

3.5 测试结果

系统模拟电商平台中的1万用户和100万单订单数据，在4台4核8 G服务器上部署完成整套系统，通过可视化任务平台建立用户画像标签50 项，包括商品类目喜好、大促敏感、交易能力等，分钟级别地完成标签的输出和更新，并通过可视化表格的方式呈现给业务人员。

4 结论(Conclusion)

本文提出的基于可视化配置的中小企业大数据解决方案，利用开源工具，结合容器化技术，能快速完成系统搭建，并且成本低。企业非专业技术人员通过可视化平台进行数据收集和数据定义，即可完成模型和标签的大数据计算和存储；工作人员通过系统输出的模型数据报表指导业务运营，整个操作简单直观且不需要专业技术人员介入，能有效降低中小企业大数据运营成本。

参考文献(References)

[1] 中国信息通信研究院.大数据白皮书(2021)[R].北京:中国信息通信研究院,2021.

[2] 马克·冯·里吉门纳姆[荷].企业的大数据战略[M].杭州:浙江人民出版社,2017:21-22.

[3] 李智慧.大数据技术架构:核心原理与应用实践[M].北京:电子工业出版社,2021:47-68.

[4] 梁晨,王耀俊.微服务架构下企业元数据管理平台的设计和应用[J].电脑知识与技术,2020,16(17):49-50.

[5] 翟永超.Spring Cloud微服务实战[M].北京:电子工业出版社,2017:61-84.

(下转第19页)