UniCloud 数据管理平台

用户手册

紫光云技术有限公司
www.unicloud.com

资料版本：5W100-20230306
产品版本：UniCloud Data Management Platform (E6103)

前 言

本手册主要介绍了 UniCloud 数据管理平台（UniCloud Data Management Platform）的概述、访问方式、功能介绍、典型案例等内容。

前言部分包含如下内容：

读者对象

本手册主要适用于如下工程师：

- 网络规划人员
- 现场技术支持与维护人员
- 负责网络配置和维护的网络管理员

本书约定

1. 图形界面格式约定

格　式	意　义
< >	带尖括号“< >”表示按钮名，如“单击<确定>按钮”。
[ ]	带方括号“[ ]”表示窗口名、菜单名和数据表，如“弹出[新建用户]窗口”。
/	多级菜单用“/”隔开。如[文件/新建/文件夹]多级菜单表示[文件]菜单下的[新建]子菜单下的[文件夹]菜单项。

2. 各类标志

本书还采用各种醒目标志来表示在操作过程中应该特别注意的地方，这些标志的意义如下：

 警告	该标志后的注释需给予格外关注，不当的操作可能会对人身造成伤害。
 注意	提醒操作中应注意的事项，不当的操作可能会导致数据丢失或者设备损坏。
 提示	为确保设备配置成功或者正常工作而需要特别关注的操作或信息。
 说明	对操作内容的描述进行必要的补充和说明。
 窍门	配置、操作、或使用设备的技巧、小窍门。

3. 端口编号示例约定

本手册中出现的端口编号仅作示例，并不代表设备上实际具有此编号的端口，实际使用中请以设备上存在的端口编号为准。

资料意见反馈

如果您在使用过程中发现产品资料的任何问题，可以通过以下方式反馈：

E-mail：unicloud-ts@unicloud.com

感谢您的反馈，让我们做得更好！

目　录

1 概述

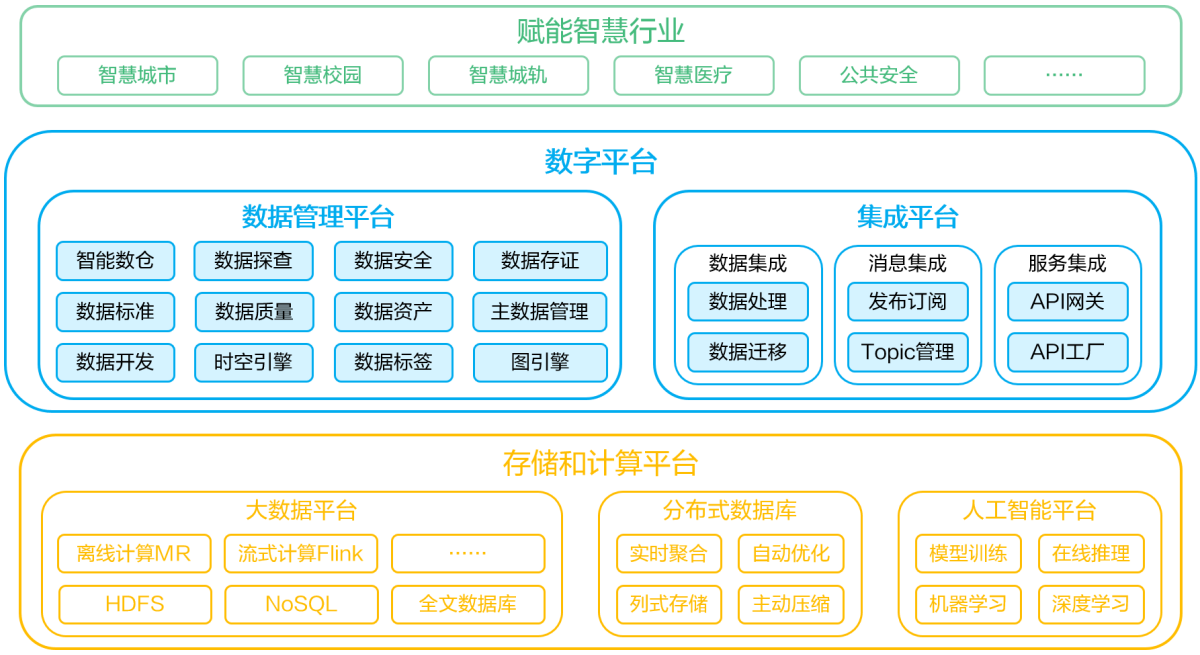
1.1 简介

DMP 数据管理平台基于新一代数据开发治理模型，提供了一站式数据开发与治理能力，极大简化治理流程并提高治理效率，为数据运营者提供一站式、自动化的数据治理及数据管控环境，由数据访问、实时计算、多维分析、业务开发、任务调度、全文检索、时空引擎、图引擎等核心子系统融合构成。将数据开发的各个环节融合在一套可视化的开发环境中，结合数据标准和数据资产相关能力，实现数据全域开发和治理，可以快速响应业务需求，通过创新带来业务价值。

1.2 产品架构

产品架构如图 1-1 所示：

图1-1 产品架构



DMP 数据管理平台依赖数据源管理功能，数据源管理模块的说明如下：

- 数据是核心资产，在 DMP 数据管理平台中可以纳管常见的数据源，用户仅需注册进来；可以对于数据源进行增删改查及测试链接等操作；屏蔽多种数据源的客户端差异。

DMP 数据管理平台产品核心功能模块及其说明如下：

- 智能数仓

  主要包括数仓规划、数据模型、表类别管理等功能，数仓规划支持对数仓数据依据业务提前进行主题和分层的规划，规范业务数据按层划分。数据模型提供可视化逻辑模型的创建和维护。表类别管理支持批量对表进行分层和主题划分，简单便捷。

- 数据治理

  通过标准管理、数据质量、数据资产、主数据管理，依照标准规范数据，全流程监控数据质量，清晰呈现数据来源走向，主数据重点管理。

  - 标准管理

    将以往文件形式的国家标准和行业标准进行系统化，帮助数据管理者构建自己的数据标准体系。通过定义数据规范，并实现标准的落地，来提升数据的可用性和关联价值。

  - 数据质量

    内置多种基础规则模型用于数据质量检测，也支持用户根据业务逻辑定义自己的可复用模型。通过规则模型与数据列进行绑定，建立数据质量指标库，即时或定时监控数据的问题发生率，及时帮助用户发现和分析数据问题。此外，还支持对数据源进行整体质量评估。

  - 数据资产

    以宏观视角对经过分析和治理的数据进行多种维度的统计，数据支持分层、主题、标签多种维度进行管理，统一管理多种数据源的元数据，拉通数据全生命周期流程形成数据全链路血缘关系，提供数据建模能力。

  - 主数据管理

    对共用的、核心业务相关的基础信息数据进行重点管理，支持灵活规划流向，同时支持数据订阅与分发。

- 数据开发

  提供实时计算、多维分析、业务流程、任务调度等数据加工处理端到端的工具集；支持复杂的数据处理模型构建；提供一站式可视化开发与管理界面，支持全托管的作业调度与灵活的调度策略；具有良好的扩展性，支持函数及作业的自定义开发，极大地降低了用户构建数据处理的复杂度，帮助企业专注于数据价值的挖掘和探索。

  此外，通过表管理可以对数据开发中的表（来源表、过程表、结果表等）进行全方位管理；通过文件管理可以对数据开发中函数、任务等相关的文件进行管理；通过数据标签将数据表实体化，自定义业务衍生标签，沉淀业务模型制作模型标签，加速应用开发；通过时空引擎可以对时空类数据进行存储、查询、分析计算等，可无缝融合 GIS。

- 数据安全

  数据通常不能直接且全量的暴露给业务使用，往往需要事先对数据中一些隐私、敏感类等信息进行掩盖或加密处理；有效降低或避免数据资产外泄的风险；数据脱敏提供脱敏规则、敏感等级、安全审计等功能，通过对敏感信息识别、数据变形等手段实现对隐私、敏感数据的可靠保护。数据水印功能，支持对泄露数据进行溯源。数据操作存证功能可以以资源为核心对用户操作资源进行行为存证。

- 数据探查

  针对多种数据源的数据查询需要，数据查询提供了智能 SQL 编辑器通过 SQL 方式的读取多种数据源的数据，简单容易上手。全文检索与数据搜索可以对 Elasticsearch 集群中所有 Elasticsearch 表的结构化数据和非结构化数据（文档、影像、视频等多种类型内容）的整体全局搜索与管理功能（支持内容上传与下载）。

- 图引擎

图引擎是一个集成图数据库、图计算引擎和图可视化分析的一站式图服务平台。图数据库是一种用图模型来描述知识和建模世界万物之间关联关系的技术方法，旨在从数据中识别、发现和推断事物与事物之间的复杂关系，以及事物关系的可计算模型。

DMP 数据管理平台分为四个基础版本和五个特性功能版本，基础版本的差异如表 1-1 所示，特色功能版本如表 1-2 所示。

表1-1 DMP 数据管理平台基础版本

功能	基础版	标准版	教育增强版	增强版
标准管理	-	√	√	√
数据质量	-	√	√	√
数据标签	-	-	√	√
任务调度	√	√	√	√
实时计算	-	-		√
全文检索	-	-	√	√
数据访问	√	√	√	√
数据资产	√	√	√	√

表1-2 DMP 数据管理平台特色功能版本

服务组件	特色功能-数据脱敏	特色功能-数据存证	特色功能-时空引擎	特色功能-图引擎	特色功能-主数据
数据脱敏	√	-	-	-	-
数据存证	-	√	-	-	-
数据存证-区块链管理	-	√	-	-	-
时空引擎	-	-	√	-	-
图服务	-	-	-	√	-
主数据（集成在数据资产服务中）	-	-	-	-	√

DMP 数据管理平台不同版本的适用场景如表 1-3 所示。

表1-3 不同版本 DMP 数据管理平台建议使用场景

版本	建议使用场景
基础版	底层不依赖大数据平台，适合项目预算有限，需要数仓管理和小数据量数据开发的场

	景
标准版	底层不依赖大数据平台中的集群，适合项目预算有限，数据处理复杂度较底且数据量较小的场景
增强版	底层依赖大数据平台中的集群，适合有一定数据规模且对数据开发有较高要求的场景
智慧校园-教育增强版	底层依赖大数据平台中的集群，主要适用于智慧校园类项目对数据处理需求的分析场景，比增强版裁减了实时计算
图引擎独立版	底层依赖大数据平台中的集群，主要适用于在仅需使用图引擎功能进行知识图谱等方面分析与管理的场景中进行部署
特色功能-数据脱敏	必须提前部署基础版、标准版、增强版或教育增强版。适用于对数据中存在敏感字段，且有查询脱敏需求的场景
特色功能-数据存证	必须提前部署增强版或教育增强版。主要适用于对各资源操作进行记录存证，确保安全可信的场景
特色功能-图引擎	必须提前部署增强版或教育增强版。主要适用于对数据复杂关联关系有处理分析需求的智慧城市类项目
特色功能-时空引擎	必须提前部署增强版。主要适用于对时空数据处理分析需求的智慧城市类项目
特色功能-主数据	必须部署标准版、增强版或教育增强版。主要适用于需要对政企业务的主数据进行管理的场景

1.3 术语和定义

为方便用户理解 DMP 数据管理平台相关的重要概念，基本术语说明如表 1-4 所示。

表1-4 产品术语

术语	描述
全局索引	全局索引会把HBase二级索引数据放置在HBase中
全文索引	全文索引会把HBase二级索引数据放置在Elasticsearch中
数据表	数据源中的存储数据的表，是数据开发与治理的对象。数据表通常从数据源中采集识别或在本系统中创建，在表管理功能中统一管理
主数据	通常为企业或组织内跨部门业务协同需要的、核心业务相关的基础信息数据
质量指标	将匹配数据的规则模型与数据表中的字段结合，生成的用于衡量数据质量的指标
作业	指作业管理中下或业务流程中的作业。作业是按照系统调度规则生成的，包括可执行的代码程序包
业务流程	将不同作业进行组合生成一个针对业务的复杂任务执行流程，即可抽象成一个业务流程。一个业

术语	描述
	务流程包含多个作业，不同作业之间的顺序、依赖关系和触发条件都可以在业务流程中配置
实时作业	用于定义实时计算的作业，包括输入、输出、操作算子及其相关参数配置
数据同步作业	用于从Kafka中同步数据至指定的目标数据表中的作业
数据脱敏	指根据识别规则识别数据中的敏感信息，并通过脱敏策略进行处理，实现数据的脱敏化，保护数据安全

2 访问数字平台的 DMP 数据管理服务

DMP 数据管理平台作为数字平台的产品之一，在数字平台中以服务的形式提供 DMP 数据管理的各项功能。

2.1 登录

数字平台的系统服务安装成功以后，数字平台的 URL 地址会在安装脚本成功执行完成后显示，格式为：https://VIP:32015。

在浏览器中输入地址：https://VIP:32015，进入登录页面，如图 2-1 所示。输入正确的用户名和密码，单击<登录>按钮即可登录数字平台。数字平台缺省的超级管理员用户名为 admin，缺省密码为 Passw0rd@_。如果用户名或密码不正确，系统会弹出相应的错误提示。

图2-1 登录页面

登录成功后，跳转至图 2-2 首页。

2.2 首页

首页展示了数字平台的统计信息，首页如图 2-2 所示。

图2-2 首页

2.3 退出登录

登录成功后，在右上角当前登录用户的下拉菜单中选择[退出]菜单项，即可退出系统。

3 功能介绍

说明

- 数字平台包含联机帮助，其中包括对应 DMP 数据管理的联机帮助。在数字平台中 DMP 数据管理的各功能页面中，单击页面左上角的帮助按钮，弹出帮助窗口，窗口中提供了各功能的详细配置操作说明及注意事项等，可以帮助用户更好地使用 DMP 数据管理平台。
- 本章仅对 DMP 数据管理各功能进行概括说明，以便用户快速了解 DMP 数据管理各服务提供的主要功能，关于各功能的详细说明请参见 DMP 数据管理的联机帮助。

3.1 数据源管理

选择顶部导航栏中的[工程配置]，进入工程配置模块；然后再左侧导航树中选择[数据源管理]项，进入数据源管理页面。

数据源管理是对当前用户所在组织下各工作空间中的数据源记录进行管理，该功能管理的数据源，是 DMP 数据管理平台中[数据开发/表管理]中建表数据源的来源。数据源管理包含以下几个基本功能：查看数据源列表、新增数据源、编辑数据源、复制数据源、删除数据源、分配/回收、导入/导出数据源、查看数据源使用详情。

表3-1 数据源管理

特性	描述
查看数据源列表	• 数据源列表中显示当前用户已在系统中增加的所有数据源，展示信息包括数据源名称、数据源类型、IP 地址、业务部门、业务系统、所属工作空间、创建用户、创建时间、描述信息、连接状态，对每个数据源可进行编辑、复制、删除、查看使用详情操作，对于本工作空间的数据源，还可以执行分配/回收、导出操作 • 用户可通过上方的搜索条件展开不同维度的搜索，其中 IP 地址代表该数据源创建时所属的机器 IP
新增数据源	• 在数据源列表中单击<新增>按钮，可跳转至新建数据源页面。数据源管理支持多种类型的数据源，供数字平台中各产品服务使用。其中，DMP 数据管理平台支持的数据源类型有：Elasticsearch、Greenplum、HBase、HDFS、Hive2(Embedded Http)、Kafka、MySQL、Oracle、PostgreSQL、Redis、STDB、Vertica、达梦、DLH、DRDS、SeaSQL MPP、SQL Server、ClickHouse • 新增数据源时，部分数据源需要填写目标 IP 地址、端口号、用户名及密码等信息，页面提供了“测试连接”的功能，单击<测试连接>按钮来测试配置的参数是否可以连接数据源 • 在新增 Greenplum、MySQL、Oracle、PostgreSQL、Vertica、DRDS、SeaSQL MPP、SQL Server 这几种类型的数据源时，如勾选了属性“是否采集元数据”，那么在数据资产采集任务模块会创建该源的采集任务

特性	描述
编辑数据源	• 在查看数据源列表页单击<编辑>按钮，跳转至编辑数据源页面，用户可以修改除“数据源类型”、“驱动”外的任何数据 • 为了保护用户信息的安全性，对于敏感信息会做脱敏处理，如用户密码只能通过重新输入的方式来覆盖已有的密码。所有信息输入完毕后，便可以进行测试连接并完成提交入库
复制数据源	在已经创建的数据源后面单击<复制>，输入新的数据源名称，就可以复制出一个相同数据源
删除数据源	对于系统中无意义的数据源记录，在数据源列表页中可自行删除。为了防止用户意外删除，单击<删除>按钮会提示删除确认，确认后将删除该条数据源相关的所有信息
分配/回收数据源	将数据源分配给各组织中的工作空间
导入/导出数据源配置	数据源管理支持通过单击<导入>按钮导入数据源配置文件，创建对应的数据源连接；还支持将系统中已有的数据源连接配置通过单击对应的<导出>按钮进行导出
查看数据源使用详情	单击<使用详情>按钮可在弹出中查看该数据源在数字平台各功能中的具体使用详情

3.2 智能数仓

智能数仓支持按照业务提前对数据进行分层和主题规划，支持批量对表进行分层、主题、部门的批量管理，数据模型功能用于构建逻辑表、定义数据表逻辑结构以及字段之间关系，数据模型中的逻辑表可以物理化生成对应的数据表。

表3-2 智能数仓

特性	描述
数仓规划	• 内置默认 6 类分层归属，可自定创建分层和主题。不可在“其他”归属下创建分层 • 分层支持配置检查器，对分层的表进行命名合格性检查
数据模型	• 支持对数据模型进行管理，包括新建、编辑、导入、导出、全部导出、导出记录查询、批量删除等功能 • 支持可视化创建逻辑表模型，逻辑表可以在多种数据源上一键创建物理表（依赖映射管理中的配置） • 支持直接新建、引用其他逻辑表、引用已有数据表（物理表）生成逻辑表 • 支持编辑、查看逻辑表，切换版本，生成建表语句 • 支持创建关联关系表，支持对现有业务表进行关联建模 • 支持对数据模型进行版本化管理，包括新建版本、切换版本、版本对比 • 数据模型支持逻辑表和物理表一致性比对 • 数据模型支持 MySQL、PostgreSQL、Oracle、Vertica、Greenplum、SQL Server、

特性	描述
	SEASQL MPP、DRDS、DLH、Hive、达梦数据源类型 • 逻辑表关联标准后可在物理化生成数据表时自动生成质量指标 • 数据模型支持逻辑表增加分区信息
表类别管理	• 表类别管理提供对表的快速分类功能。使用者可以根据已经注册的数据源，对该数据源下的表进行批量分类，可以按照自定义的主题、分层和业务部门批量添加以及删除 • 在表类别管理页面可以根据数据源，以及是否已经关联来等条件对表信息进行过滤，方便查看某个主题、分层和业务部门下已经挂载的表，或者某个数据源下挂载在指定主题、分层和业务部门下的表。挂载到不同主题、分层和业务部门下的表将在总览中统计并展示。目录树中对挂载表的个数进行统计并展示

3.3 数据治理

数据治理整合了系统中的基础资源和部分数据治理功能。基础资源包括数据标准、数据质量、数据资产、主数据管理等。

3.3.1 标准管理

标准管理提供了数据标准进行管理和维护的功能。其能够解决数据格式不统一、数据内容不规范的问题，并帮助用户对现有数据进行梳理，对新数据进行规范约束，从系统层面建立标准体系，用流程化的方式构建新的标准。

缺失标准管理这一环节，容易遇到因格式不一致，值域范围不统一导致的数据之间无法有效关联的问题，增加数据价值挖掘的难度，降低数据的可用性。因此，在全局范围内建立标准体系，可以有效保障新数据的质量和使用价值。

表3-3 标准管理

特性	描述
概述	概述页面包括功能介绍、功能模块、产品术语等区域，各区域的说明如下： • 功能介绍：介绍了数据标准管理的功能及各子功能模块的作用 • 功能模块：通过卡片形式展示了数据元管理、数据项管理和数据集管理的概念，并提供了相应功能模块的入口。单击区域中各模块卡片中的<立即进入>按钮，即可跳转至对应的功能模块页面 • 产品术语：通过卡片形式介绍了数据元相关、数据项相关、数据集相关的产品术语。点击区域中各卡片，会弹出相关的数据说明窗口。单击窗口底部的<确定>按钮，即可返回概览页面
数据集管理	数据集是数据项的集合，用于将数据项按照业务特性进行分类管理： • 数据集管理提供数据集的新增、导入、模板下载、搜索、查看详情、导出、批量导出、

特性	描述
	全部导出、导出记录查看、移动、批量移动、删除、批量删除、全部删除、查看变更记录、查看版本使用详情等功能 • 数据集可以通过引用数据项管理中的数据项、引用已审批通过的数据集中的数据项两种方式来添加数据项 • 新增或者编辑数据集时，可以通过删除、批量删除、全部删除的方式删除数据集中的数据项
数据项管理	数据项是构建数据集的最小单元： • 数据项管理提供数据项的新增、导入、模板下载、搜索、查看详情、导出、批量导出、全部导出、导出记录查看、移动、批量移动、删除、批量删除、全部删除、查看变更记录、查看版本使用详情
数据元管理	数据元是数据标准中的基础数据： • 数据元管理提供数据元的新增、导入、模板下载、搜索、查看详情、导出、批量导出、全部导出、导出记录查看、移动、批量移动、删除、批量删除、全部删除、查看变更记录、查看版本使用详情
代码管理	代码管理是对标准取值范围的管理，数据元、数据项可以引用代码来描述自身的取值范围： • 代码管理提供目录的新增、编辑、删除、搜索功能 • 码表列表提供新增、导入、模板下载、搜索、批量删除、批量导出、全部导出、全部删除、查看导出记录等功能 • 码表列表中提供查看详情、编辑、删除等功能 • 码表支持查看物理表和数据同步操作
标准发布	录入码表、数据元、数据项、数据集后，即可将这些标准内容汇总，作为数据标准的一个版本进行标准发布，发布操作支持流程审批 任意两个已发布版本的标准可以进行版本比对，对比结果中可以呈现两个版本的相同标准、差异标准、各自独有的标准等，能够帮助用户快速了解版本之间的变化

3.3.2 数据质量

数据质量模块通过为数据字段绑定规则模型，来定义字段指标。多个指标可以挂载到某一个质量监控任务中，形成一套质量监控方案，根据质量监控任务的执行结果形成质量报告，从字段的维度统计各指标的准确率。

表3-4 数据质量

特性	描述
规则模型	内置的校验规则，在配置指标时会用到这些内置规则，目前包含的内置规则有以下： • 空值校验

特性	描述
	• 值域校验 • 格式校验 • 长度校验 • 唯一约束校验 • SQL 条件校验 此外，还支持自定义创建规则模型，导入、导出、导出记录查看、全部导出、批量删除、全部删除、模板下载功能
指标管理	指标通过对数据字段绑定规则模型，来定义字段指标，指标管理提供指标的新建、全部删除、批量删除、导入、导出、导出记录查看、全部导出、模板下载功能 同时提供新增指标保存为草稿、更新草稿、删除草稿、草稿保存为指标功能
质量监控	多个指标可以挂载到某一个质量监控任务中，形成一套质量监控方案，质量监控提供对监控任务的创建任务、编辑、删除、共享、搜索、详情查看、执行结果查看、导入、导出、查看导出记录、全部导出、模板下载功能 手动调度提供立即执行、停止任务功能，自动调度提供启动任务、结束调度任务功能
质量报告	一个质量监控任务的执行结果可形成质量报告，包括数据表质量报告、指标趋势报告、部门报告（基于字段归属部分生成） • 数据表质量报告以表为维度，展示整个表的质量检测情况（展示检核数据量、错误数据量、错误率） • 指标趋势报告以指标为维度，展示不同任务中指标的历次质量检测结果（展示检核数据量、错误数据量、错误率） • 部门报告以部门为维度，展示其包含的存在质量指标的字段（创建表时为字段关联部门）相关的质量检测结果（展示检核数据量、错误数据量、错误率） 此外，还支持导出数据表质量报告
评估配置	评估配置：用于管理数据源的评估配置任务：用户通过创建不同的评估配置，评估配置即可以任务的形式根据指定的调度方式对数据源进行质量评估，并将结果呈现在评估报告中
评估报告	提供了数据源的质量评估报告：报告中提供了数据源的质量评分、接入率、表活跃度等信息，并提供了表详情列表

3.3.3 数据资产

数据资产以宏观维度展示了客户数据资产全貌，实现多种维度的资产图表统计展示，并提供数据血缘和智能搜索，以盘清数据资产、理清数据链路、管妥数据分层和搞懂数据价值为使命。

表3-5 数据资产

特性	描述
总览	支持统计MySQL、Oracle、PostgreSQL、Greenplum、HBase、Elasticsearch、Hive2(Embedded Http)、DLH、DRDS、SQL Server、SeaSQL_MPP、达梦、Vertica、ClickHouse类型数据源的数据，以图表的形式展示多维度的数据统计情况。支持按不同数据源类型配置数据的统计情况和统计周期 • 支持展示数据源总数、表总数、数据字段总数、数据总存储大小、数据总条数、同前一天的比较。按照主题或分层统计数据表总数分布、数据字段总数分布、数据条数总数分布、数据存储量总数分布、数据量趋势(按主题或分层)。对于每一种数据源类型，统计上述指标项，表来源于两部分，采集的表和数据开发中创建的表。再将不同数据源下的表挂载到自定义的主题以及分层下，按照主题和分层的维度以上述指标项对挂载表的数据进行统计以及展示 • 支持自定义数据资产展示看板
元数据采集	支持对PostgreSQL、Greenplum、MySQL、Oracle、Hive、Vertica、DLH、SQL Server、SeaSQL MPP、DRDS类型数据源中的表信息、字段信息的采集，不采集视图、函数、存储过程 • 采集任务：支持周期性调度，任务运行后支持查看采集任务的日志信息。在创建数据源的时候，选择采集元数据可以自动生成采集任务并执行。目录树中将任务个数进行统计并展示 • 运行监控：展示了采集任务的运行记录和日志，并可以对采集任务进行重跑操作
血缘管理	• 支持血缘概览，界面展示系统所有表、作业、接口、应该、业务系统的血缘统计情况，支持按照以上各维度进行下钻，按血缘节点类型和血缘趋势统计 • 数据开发中涉及到数据处理、流转的服务包括数据采集、数据同步、实时计算、多维分析等，在数据处理过程中记录数据的关系，从中解析出完整的数据血缘关系 • 在界面中，支持按照表-表名、作业-作业名、数据 API-api、应用名称-app 进行搜索，结果以可视化的方式在界面展示表和字段的血缘关系。对于自动解析作业中的数据血缘关系失败的情况，可以通过手动添加的方式生成作业中的血缘关系

3.3.4 主数据管理

主数据管理支持协调和管理与政企的核心业务主体数据，使得主体数据的变动和更新具有一致性、共享性、权威性，保证上层应用对主体数据的一致，从而实现更好的跨部门协同和数据治理。

表3-6 主数据管理

特性	描述
主数据管理	• 支持将已注册的 Oracle、MySQL 数据源绑定为主数据的数据源 • 支持将主数据源和非主数据源中的数据表添加为主数据表，并进行流程审批

特性	描述
	• 主数据表中数据的添加、修改、删除需要进行流程审批，并支持查看操作记录 • 支持将主数据表中的数据备份以及根据备份文件进行数据恢复 • 支持其他使用方以接口的形式订阅主数据 • 主数据绑定推送接口后，展示每次推送给订阅者的状态 • 支持数据流向规划，数据 U/C 矩阵（一数一源），可以查看谁创建的数据源被哪个用户所使用 • 支持数据分发和数据版本管理能力 • 主数据管理支持灵活可配置的编码引擎 • 支持对主数据进行版本比对 • 主数据管理支持将主数据的变更信息推送到配置的消息队列

3.4 数据开发

数据开发涵盖表管理、多维分析、实时计算、业务流程、数据标签、时空引擎等能力集，打造出全域数据开发和全链路的数据监控，让用户轻松能够看到整个开发链条上每个节点的开发状态和统计监控，结合数据资产、数据质量、数据安全等功能，方便用户掌控数据资产，为企业政府的发展决策提供依据。

表3-7 数据开发

特性	描述
表管理	• 表管理提供了可视化的建表功能，支持主题视图和分层视图对表资源进行列表展示，表结构支持根据标准管理中定义好的行业数据模板进行选择，统一元数据定义，便于数据质量分析和治理。分为可管理表、可使用表、表上架管理、已订阅表 • 表管理提供了草稿箱功能，可统一管理未创建完成的草稿 • 表管理提供了 HBase、Kafka、Elasticsearch、MySQL、PostgreSQL、达梦、Greenplum、Hive、Oracle、Vertica、STDB、DLH、DRDS、SeaSQL MPP、SQL Server、ClickHouse 数据源的建表功能，其中，ClickHouse、DLH、DRDS、Greenplum、Hive、MySQL、Oracle、PostgreSQL、SeaSQL MPP、SQL Server、Vertica、达梦类型的表，支持对字段名、类型、描述的增删改，其他数据源的表只支持追加字段的功能 • 表管理支持查看详情、列属性、预览、血缘、索引、全文索引、版本等信息查看 • 数据上架支持将表上架到资产市场 • 表中单个字段可以通过选择数据集中的数据项或者直接选择数据项作为对应标准进行关联，已关联标准的字段支持重新关联 • 支持对 HBase、MySQL、Greenplum 类型数据源下的表添加索引 • 支持对 Greenplum、MySQL、PostgreSQL、Vertica、达梦类型数据源下的表添加

特性	描述
	全文索引 • 支持在已指定为主数据源的 MySQL、Oracle 类型数据源下建表时配置表为主数据表 • 根据业务需求对表配置不同的主题和分层，方便管理 • 支持全程可视化操作建表，无需在线编写建表 SQL 语句 • 支持建表关联国家、行业、企业标准，自顶向下，自成一体 • 支持对多种维度对表进行分类管理和搜索 • 提供表的查看、编辑、删除、清空、共享、授权、订阅等操作 • 支持通过表的授权功能可以实现表的跨组织跨工作空间共享 • 导出功能可以批量导出平台中的表，然后导入其他 DMP 数据管理平台环境 • 支持通过上传 SQL 文件进行批量建表 • 支持使用标准管理中的码表进行建表 • 发生元数据变更时，支持针对指定邮箱进行邮件预警 • 表管理支持复制功能 • 建表支持符合规范的前提下快速命名的功能特性
作业开发	• 作业管理支持创建实时、数据同步两种类型作业： o 实时作业：支持创建 FLINK_GRAPH、FLINK_JAR 和 FLINK_SQL 三种作业类型，支持对作业进行编辑、删除、查询等功能 o 数据同步作业：支持创建同步任务，将 Kafka 中的数据实时同步到不同的数据源中，支持的目标数据源包括 Elasticsearch、MySQL、PostgreSQL、Greenplum、达梦、Vertica、STDB、DLH、DRDS、HBase、ClickHouse 以及文件（File）中。支持作业的删除、查询等功能 • 函数管理支持对 Spark 和 Flink 引擎的用户函数进行管理和展示： o 函数管理对函数进行创建、编辑、查看、删除、导入、导出等 • 任务管理是对离线任务模板的管理，支持内置作业管理和自定义作业管理，任务模板在[调度中心]的业务流程中被离线分析组件调用时，通过调整参数作为具体的作业运行： o 任务管理支持的离线任务类型有 5 种，分别是：SparkJar 任务、Java 任务、MR（MapReduce）任务、Shell 和 PySpark 任务 o 所有任务的操作类型均包括新建、禁用、删除、编辑任务、编辑附件、共享、查看和导出，不同的任务类型所需要的参数及任务配置项有所区别 o 内置作业包括 HBase 表数据统计、HBase 表数据批量加载、时空表数据批量导入导出、数据写入 Hudi 并同步 Hive 表（SparkAPI 方式）、数据实时写入 Hudi（Hudi DeltaStreamer 工具）、聚合统计分析任务、增量脱敏最大值计算（Column_Max_Value_Calc）、数据存证相关任务

特性	描述
调度中心	调度中心的核心是业务流程数据管理，一个业务流程数据对象可以统筹操作多个类型的作业，如同步作业、实时作业等，多个不同类型作业各自完成不同的目标，但最终是为了实现一个实际业务目的 • 新建业务流程 • 业务流程画布中支持数据集成、离线分析、实时计算、控制节点 • 业务流程支持简单调度和高级调度，高级调度使用 Cron 表达式控件 • 业务流程支持批量删除、查看批量操作记录（批量操作包括批量删除和复制的记录）功能 • 业务流程支持根据业务流程名称和更新时间、创建者等列进行排序 • 单个 Spark 节点支持补数据，对 SQL 语句中的时间变量，界面中可以传递具体的值 • 支持 Java、Shell 作业输出执行结果；Java、Shell、SparkJar、PySpark 作业接收上游作业节点的结果作为参数参与作业执行 • 业务流程通过子流程节点支持业务流程嵌套 • 业务流程支持发布、回退、同步 • 业务流程支持对节点配置通知，当节点运行异常时可以发送通知 • 业务流程支持将 MR、PySpark、SparkJar、Shell、Java 类型的新建作业保存为作业模板使用 • 业务流程支持查看画布的操作记录日志 • 支持对业务流程划分业务分组和业务标签
运维概览	运维概览页支持查看当前通知数量、业务流程过期调度数量、任务队列的资源使用趋势，运行状态分布情况，运行耗时Top排名，任务类型分布情况
调度运维	• 展示了所有业务流程 • 支持业务流程的批量提交和停止 • 支持分组和标签对业务流程进行管理 • 支持对业务流程实例进行监控和调度信息查看，其中监控页面可以对节点的运行情况、日志、数据进行查看和预览 支持查看批量操作记录日志
运行实例	展示了业务流程在运行过程中，各任务运行实例信息，支持对实例进行重跑和查看实例详情
运维通知	支持查看节点运行异常的消息通知信息，支持异常消息通知配置
文件管理	通过常见的列表方式展示了文件系统中的目录和文件 提供了文件或文件夹的常用管理操作，包括新增、上传、下载、移动、重命名等功能 资源管理支持对DMP数据管理平台中常见资源的文件进行管理，如内置任务、自定义函

特性	描述
	数、自定义作业等

表3-8 数据标签

特性	描述
标签模型	基于OLT（对象、关系、标签）模型来抽象物理数据形成了的模型，这个模型就叫标签模型。构建语义层，为业务人员“看懂数据”从而“使用数据”提供基础 每个标签模型包含标签管理、对象管理、关系管理三部分： • 对象管理：对象描述一个客观存在实体或事件；从数据角度看，是一个多种数据描述同一实体或事件的集合，因此数据标签中的对象可以对应数据治理层的一张表，也可以对应数据治理层的多张表。对象管理用于管理该模型下的所有对象，支持对象的创建、删除、导入、导出、编辑等功能 • 标签管理：标签是挂在对象之上的，分为原生标签和衍生标签，从物理层来讲，原生标签就是原生物理表中的字段，衍生标签是根据已有的标签（字段）计算出来的新标签（字段），标签管理用于展示当前模型下所有对象创建的标签，支持针对标签的上下线、编目、值类型编辑、编辑、删除等功能。只有上线的标签，才能用于数据探查、数据分析 关系管理：关系图表达对象与对象之间的关联关系，让用户可以在视图上直接进行各类的关联计算，不需要预先对物理数据进行大规模的加工处理，即用即算
标签方案	提供业务衍生标签加工功能，降低标签加工的门槛，使得业务人员可以根据业务场景进行标签数据加工方案，即标签方案。衍生标签是在已有标签之上做一些标签加工生成的标签。生成的衍生标签字段可绑定到实体对象上
标签任务	配置的标签方案生成标签生产任务并运行。标签任务可以一次性运行和周期调度运行
标签探查	标签模型创建完成后，可以在标签探查模块中对标签数据进行可视化的查询，支持查询所有对象（实体）的标签及标签内容数据，支持多标签的内容筛选能力

表3-9 时空引擎功能特性

特性	描述
概览	展示时空引擎整体运行情况，包含了时空数据源状态预警、类型分布、数据分布情况，热点时空数据表访问量，重点概览数据展示。以上功能均按照时空数据源进行分类，可按照时空数据源切换数据
数据目录	根据时空数据表结构，展示时空数据源和时空catalog的树结构，根据时空数据源和catalog展示了时空数据表的基本信息，可对时空表进行查看、离线表注册、扩展表、清空表数据等操作 时空表详情展示时空表的基本表信息和所有字段信息，可对表进行数据导入操作，查看时

特性	描述
	空表在物理集群上的状态信息
时空查询	对时空表进行数据查询，查询方式支持查询转换、实时状态、聚合统计、数据导出、轨迹查询： • 查询转换对数据使用 CQL 查询语法进行数据过滤查询，支持对查询结果函数转换，针对时空表的 HBase 和 Redis 数据源进行查询 • 实时状态查询是对时空对象的最新位置信息进行查询，针对时空表 Redis 数据源进行查询 • 聚合统计查询是根据 CQL 过滤条件对数据进行多种聚合分析 • 数据导出功能是使用分布式导出方式根据 CQL 过滤条件对大数据量进行导出 • 轨迹查询是根据 CQL 条件过滤的位置信息重组成轨迹信息的查询方式
运行维护	对时空引擎进行总体配置，设置时空数据统计周期，配置时空引擎对接的地图引擎信息，以及对时空概览重点数据的配置

3.5 数据探查

数据探查提供了对数据和系统资源进行全方位查询探索的功能。其中，数据查询提供了简单易用的数据查询工具，操作简单容易上手，不需要掌握大数据组件的原始用法，同时具备多种便捷操作，例如历史 SQL 查看，多窗口查询、查询结果导出等。全文检索与数据搜索功能用于对 Elasticsearch 集群中所有 Elasticsearch 表的结构化和非结构化数据的整体全局搜索。

表3-10 数据查询

特性	描述
数据查询	• SQL 编辑器支持对 HBase、MySQL、Oracle、PostgreSQL、Elasticsearch、Greenplum、达梦、Kafka、Vertica、Hive、DLH、SQL Server、SEASQL MPP、DRDS、ClickHouse 类型表进行查询 • 支持对数据进行 insert、update、delete 等更新操作 • 支持对 SQL 语句进行格式化、保存 SQL、查看历史 SQL 列表等便捷功能 • 查询结果以结构化表格展示，支持翻页和查询结果导出 • 左侧导航树以数据源列表展示，选中可查看已有表中数据量、字段等详情 • SQL 编辑器支持多窗口，方便同时执行多种 SQL • SQL 编辑器中执行 SQL 语句添加 limit 限制，查询结果最多返回 10000 条 • SQL 调试支持调试需要在 SparkSQL 或 Hive 数据源执行的 SQL 语句 • SQL 调试支持查看表字段信息 • SQL 调试支持对 SQL 进行执行、SQL 上传、选中执行、格式化、语法校验等便捷

特性	描述
	操作 • SQL 调试支持多个 SQL 串行执行 • SQL 调试支持执行记录、动态日志、执行结果、表信息、字段信息的查看
全文检索	主要用来对[表管理]中创建的Elasticsearch类型表中的数据以及全站范围内的资源进行检索 • 支持全站范围内进行数据和资源搜索 • 支持对指定主题下的所有 Elasticsearch 表的全局搜索 • 支持对指定数据源下的一个或多个表进行表的全局搜索。如果搜索到的数据本身含有附件类型的数据结构，还提供文件的下载功能
数据搜索	• 支持自定义查询 • 支持按时间序列进行查询，可以根据用户的时间相关数据，渲染出某些指标在某段时间的变化情况，从而便于用户根据现有趋势做出决策 • 支持按表格进行查询，支持明细查询、聚合查询、模板查询
数据上传	• 支持对[表管理]中创建的 Elasticsearch 类型表插入数据 • 支持对 int、keyword、text、double、date、long、boolean 类型数据的上传 • 支持对 attachment 类型字段的文件、图片、视频的上传 • 数据导入：数据导入用于将 MySQL、PostgreSQL、GreenPlum、达梦和 Vertica 类型的数据源中的某张表的数据导入到 Elasticsearch 类型数据源中的目标 Elasticsearch 表，并且支持持续监控来源表中的数据新增情况，同步将新增数据更新到对应的目标 Elasticsearch 表中。这部分的数据导入需要在数据源中定义 Elasticsearch 类型的索引表（不包含附件类型的表）；另外，对于来源的外部表需要提供时间类型和主键字段的支持，事件类型用来判断数据的增加情况，主键主要用来进行数据的去重
集群监控	• 集群监控以图表形式展示 Elasticsearch 集群总览信息、集群节点信息、Elasticsearch 索引信息、全文检索服务实例信息和服务接口调用数据

搜索管理	• 自定义词库
	  ◦ 自定义词库用于全文检索自定义分词的管理，在此可以添加行业或业务等需要的特殊词汇，优化索引和搜索效果
	  ◦ 用户可以通过文件方式去批量导入，也可选择单个词汇去单条上传
	• 快照管理：快照管理用来备份某一时刻 HBase 及 Elasticsearch 的数据状态，以便异常情况下恢复 HBase 及 Elasticsearch 的数据，防止数据丢失

3.6 数据安全

数据安全提供了数据脱敏，数字水印和数据存证功能。

- 数据脱敏：对数据中的敏感信息通过规则进行识别和脱敏处理，使敏感数据得到有效保护。通过对敏感信息自动发现、分级分类、数据变形、安全审计等功能实现对敏感隐私数据的可靠保护。
- 数据水印：数据脱敏支持在脱敏过程中对数据进行水印嵌入，通过修改最低有效位和零宽度空格算将水印信息编码后嵌入到原始数据列中，通过泄露数据文件提取水印编码信息进行溯源。
- 数据存证：基于区块链以资源维度对用户的操作行为记录存证，保障数据的安全可信，提供行为画像、存证概览、存证分析可视化界面，助力客户快速进行行为分析。

表3-11 数据脱敏

特性	描述
首页	首页展示了数据脱敏系统状态的总览信息，包含近一周新增敏感字段总数、近一周捕获的风险总数、敏感数据访问量趋势、风险操作量趋势、用户授权的等级分布和数据源授权占比，旨在让用户掌握系统敏感数据的分布及访问状况，调整措施，提高数据保护能力。该模块提供展示和切换组织的功能
授权管理	支持对数据源授权和用户授权两种。数据源授权支持对数据源进行扫描权限、脱敏权限、审计权限的配置，细分对数据源的操作权限，支持新建、展示、编辑、删除数据源授权。用户授权目的是给用户指定访问等级，以及授权角色，授权用户仅能访问到小于等于自身等级的未脱敏信息，支持开发、审计、普通三种角色的选择，支持展示、新建、编辑、删除用户授权
分级分类	• 分级管理：定义和管理数据的敏感级别，按照数据的价值、内容敏感程度、影响和分发范围不同对数据进行敏感级别划分，方便进行权限控制，并实现根据分级进行资产打标，安全管控等。不同敏感级别的数据有不同的管控原则和数据开发要求。系统支持为数据识别规则和授权用户设置等级。该功能页面提供新建、导入、导出、展示、编辑、管理分级规则的能力 • 分类管理：定义和管理数据识别规则的分类模板，便于对数据识别规则进行分类管理
数据识别	数据发现基于不同维度展示识别到的敏感字段的统计信息，包含识别字段数、识别表总数、敏感信息在各分级分布、数据源分布以及明细统计 • 识别模型：通过指定扫描方式、扫描类型和数据识别逻辑，构建出对数据识别的基础模型。系统内置了常用的敏感字段识别模型，并支持用户根据其行业特点自定义模型，可被识别规则引用 • 识别规则：基于识别模型，结合数据敏感级别，构成对数据进行扫描识别的规则。支持根据配置的识别规则对指定范围的数据进行扫描、分级 • 识别任务：基于数据识别规则，对已授权数据源进行扫描，识别其中匹配规则的数据 • 识别日志：展示了敏感数据识别日志和调度任务的日志。敏感数据识别日志以表粒度展示扫描的表信息、扫描状态、结果信息和扫描时间等内容；针对扫描失败的表，可以按结果信息进行处理并重试，该模块提供了展示、搜索和切换组织的功能。调度日

特性	描述
	志展示调度任务的启动时间、与单次调度扫描的表数量，该模块提供了展示和切换组织的功能 • 敏感信息维护：用于展示系统识别到的敏感字段信息，包括敏感字段所属的表以及数据源信息，敏感字段对应的识别规则等。支持对敏感信息的搜索、编辑与删除，方便对识别结果进行手动修正。此外还支持将敏感表发布动态脱敏接口
数据脱敏	• 访问概览记录并展示对敏感信息访问的统计和详细信息。包含访问量趋势图、访问量、访问人数和访问记录的展示 • 脱敏策略：支持对应识别规则创建数据脱敏规则作为数据脱敏的默认策略，或选择默认策略中的规则组成自定义策略。通过脱敏策略中的规则，可以实现对数据识别规则识别出的敏感数据进行脱敏 • 静态脱敏：支持对 MySQL、Oracle、达梦、Vertica、PostgreSQL、Greenplum、Hive 等数据源到同源数据库、HDFS 文件的静态脱敏，支持全量和增量的抽取脱敏方式。生成的脱敏任务默认在调度中心的数据脱敏分组中
风险审计	系统支持自定义风险审计规则，对潜在危险操作进行识别并记录 • 审计规则：对用户的行为进行分析，将触发到规则的事件进行记录，推送给安全管理员进行风险审计，支持针对数据源、表、分级、敏感数据类型、操作数量、操作时间等场景进行布控。该功能页面提供新建、展示、搜索、编辑、添加配置、删除、启用或者禁用规则的功能 • 审计日志模块展示识别的风险操作统计信息和详情展示，支持用户对结果进行审计操作，确认或排除风险项。通过此项，管理员能够发现数据风险项，进而采取针对性操作提高系统的数据安全。该功能页面提供了展示、搜索、审计、删除和切换组织的功能

表3-12 数据水印

特性	描述
水印嵌入	提供水印任务管理页面，支持对数据库表提供水印嵌入的任务，数据处理流程复用静态脱敏的过程，提供对原始表进行数据抽取后经过水印嵌入后加载到目标表的能力，原始表数据不进行修改，通过开发SparkSQL自定义函数实现水印的嵌入过程
水印提取	提供水印提取任务管理页面，支持针对各种方式嵌入水印后泄漏的数据文件进行水印的提取溯源，溯源过程是水印嵌入的逆过程，支持针对只有部分数据的水印提取，支持部分水印编码丢失后尝试进行水印还原的能力

表3-13 数据存证

特性	描述
存证主题管理	• 支持存证主体的创建和维护，支持是否开启区块链存证，如果不开启，则不将该主体数据存入区块链中。如果开启，则会创建相应的区块链资源 • 在任务监控页面，会显示当前主体的任务总数，主要包括两部分：数据接入任务和指标分析任务。数据接入任务，用于从消息队列中消费数据到对应的存储介质中，指标统计任务，是周期性任务，会统计当前存证主体中的数据，梳理统计指标，用于界面展示 • 组织管理：用于存证主体的权限控制，当组织被注册到存证主体后，被注册组织的组织管理员才有权限访问存证主体 • 认证管理：被注册的组织，有权限进行认证管理，通过认证管理生成一个认证密钥。通过该密钥进行数据写入的权限校验。
存证查询	• 支持按照时间维度和用户维度对存证数据进行查询，查询结果支持表格视图和时间轴视图 • 提供用户行为画像，以组织维度进行统计，最常使用服务、最常访问资源、最多使用资源、最多执行操作、最常执行的操作类型以及各种 top5 数据统计 • 提供存证概览，全局维度进行统计：当前存证数据的组织数量、用户数、操作总数、服务总数、资源总数 • 存证分析：按时间周期（每月）对存证数据进行统计分析，该周期内操作用户数，操作总次数，设计组织数。按操作类型统计，按 用户、资源、服务、操作次数 统计。

3.7 图引擎

图引擎是一个集图数据库、图计算、图可视化为一体的一站式图服务平台。针对高度互联数据的存储和查询场景进行设计，提供一种更好的组织、管理和理解海量信息的能力。适用于数据之间存在复杂或深度关联关系的场景，利用高度连接的数据中复杂、动态的关系来产生洞察力和竞争优势。

表3-14 图引擎功能特性

特性	描述
图谱	图库中所有图集中管理，提供图的创建、修改、删除以及配置管理。支持同工作空间内的共享和不同工作空间之间的授权
图概览	展示了图的业务流程以及图的各项信息，并提供了各种操作的入口： • 提供清空、备份/还原、提交统计、定时统计配置等功能
图模型	图模型页面中展示了图中所有标签模型，包括模型管理、模型管理（表格视图）和索引管理两部分： • 模型管理：包括可视化建模、模型管理（增删改查），从数据中提取实体、关系、属

特性	描述
	性要素，进行模型构建，满足业务场景建模需求 • 模型管理（表格视图）：列表方式展示属性、顶点标签和边标签，并提供了更新等管理功能 • 索引管理：支持通过索引加速点、边的查询效率，提供索引状态监控以及索引的分布式重建、删除功能 • 此外，模型管理和索引管理均提供模型批量加载与导出 Schema 文件功能
图入库	图入库功能包括图模型展示以及数据入库管理。针对不同场景对数据接入的需求，提供单机、分布式、实时三种数据导入工具 可支撑亿级数据的高效入库，支持增量/全量两种入库模式。可通过图任务实时监控入库任务状态，支持入库过程中点数据校验、错误数据记录功能
图检索	图检索用于图数据的查询，支持多种查询方式，包含点查询、边查询、路径查询、扩展查询、模板查询以及Gremlin查询 针对点、边查询，封装了基于属性条件过滤的查询功能。查询结果可视化展示，提供树形、圆形、力导向图三种布局方式，支持查询结果的过滤及导出。Gremlin查询支持输入Gremlin语句检索图数据，提供历史查询语句保存及查看功能
图算法	图算法功能用于关系数据的推理运算，挖掘隐藏关系 内置丰富的图算法库，包含PageRank、PersonalPageRank、连通体、增强连通体、三角计数、单源L-Hop、多源L-Hop、标签传播、最短路径、单源最短距离、最短距离，基于Spark GraphX提供分布式的图数据挖掘，满足各种场景的算法分析需求。提供可视化的算法分析界面，通过界面选择相应算法、数据进行分析，提供算法分析任务的可视化监控，并对分析结果可视化展示
图任务	图任务提供图的任务监控功能，可监控的任务包括图算法、图入库、图管理、图索引、顶点中心索引、数据统计，可以实时查看任务状态，任务失败可以检查异常信息。针对入库任务，提供任务取消及任务详情监控功能，实时查看入库数据量

3.8 系统

系统模块提供了对用户、权限、组织、系统等进行配置的功能，并支持查看操作日志。

表3-15 系统功能特性

特性	描述
组织管理	组织管理功能用于对系统内的所有组织进行管理，如新建、编辑或删除组织等： • 用户可以单击<新建组织>按钮，新建组织 • 用户可以单击组织对应操作列中的<详情>按钮，查看组织详情 • 用户可以单击组织对应操作列中的<编辑>按钮，编辑组织名称、上级组织等组织信

特性	描述
	息 ● 用户可以单击组织对应操作列中的<删除>按钮，删除组织 ● 组织管理支持组织名称、组织全称等条件进行查询
用户管理	用户管理用于管理本系统中的用户： ● 用户可以单击<新建用户>按钮，新建用户 ● 用户可以单击用户对应操作列中的<禁用>按钮，禁用该用户 ● 用户可以单击用户对应操作列中的<编辑>按钮，编辑该用户的用户名、隶属组织、电子邮箱等用户信息 ● 用户可以单击用户对应操作列中的<修改授权>按钮，修改该用户的角色信息 ● 用户可以单击用户对应操作列中的<删除>按钮，删除该用户 ● 用户管理支持用户名等条件进行查询
角色管理	角色管理用于管理本系统中用户的角色： ● 用户可以单击<新建角色>按钮，新建角色 ● 用户可以单击<编辑>按钮，编辑角色 ● 用户可以单击<删除>按钮，删除角色 ● 角色管理支持角色名称、角色类型等条件进行查询
操作日志	操作日志指具体的应用产生的访问记录日志： ● 用户可以查看自己的操作记录 ● 操作日志支持通过 IP 地址、操作、操作时间段、级别和结果等条件进行查询
系统配置	系统提供了系统相关的配置管理功能，包括系统菜单管理、流程配置、安全设置、基础设置和大数据集群相关配置： ● 菜单管理：用于配置数字平台中各功能的布局及是否显示等，方便系统管理员对界面功能展示进行配置。用户可以单击<编辑>按钮，编辑菜单的图标、排序、是否打开新页面等菜单信息；用户可以单击<隐藏>按钮，隐藏菜单 ● 流程配置：为用户提供了一个多组织的、自助的流程服务。用户可以单击<新建>按钮可以新建流程，单击流程配置名称可以查看流程详情，未启用的流程点击<启用>按钮启用该流程，已启用的流程单击<禁用>按钮禁用该流程，未启用的流程单击<编辑>按钮可以编辑该流程，未启用的流程单击<删除>按钮可以删除该流程 ● 安全设置：密码策略用于管理用户登录系统的认证鉴权策略；登录认证策略支持开启/关闭图片认证方式；SSO 认证用于配置对接 SSO 单点登录服务器信息；会话超时策略用于配置会话超时时间，超时未操作的用户会自动退出系统 ● 基础设置：消息设置支持对邮件服务器和企业微信进行配置和启动；大数据平台访问地址支持开启或关闭配置大数据平台访问地址，开启并配置地址后，会在顶部导航栏增加大数据平台的跳转入口

特性	描述
	• 全局监控：实例监控模块主要监控本实例及注册到该实例的子级实例内部运行的各项数据指标，呈现各实例的运行情况，多实例指标涉及数据资产，服务集成，业务系统，应用系统，知识图谱的多实例指标统计 • 多级部署配置：数字平台支持多级部署，子级实例支持向上级实例进行注册，各数字平台实例可以独立运行
帮助文档	用来维护资产市场中的资产信息，管理员可在该页面中对资产的来源、用途、使用方法等信息进行介绍，帮助资产用户更好地了解和使用资产
软件授权	软件授权支持对License Server进行配置，并对授权信息进行展示
配置管理	• 参数管理，支持对系统的部分参数进行自定义设置 • 行业套件：行业套件的导入和导出入口
大数据环境管理	• 大数据集群管理：用于管理数字平台所使用的的大数据集群相关信息，支持配置多个集群 • 运行环境管理：用于管理系统中各组织与工作空间可用的集群资源

3.9 运维

运维提供了数字平台的运行维护功能，包括服务的管理、资源监控及系统巡检等功能。

表3-16 系统功能特性

特性	描述
概览	展示了系统运维相关的监控统计信息，主要包括系统的节点信息、系统基础组件信息、以及系统中各项服务的信息
安装部署	• 提供了部署、扩容、卸载、升级、回退功能 • 提供了统一的操作流程 • 支持查看整体日志和服务节点日志
服务管理	• 提供了服务的启动、停止、重启、卸载服务、日志查看等服务运维相关的功能，并提供了批量操作 • 提供了服务状态等信息展示
告警管理	• 以列表的形式展示数字平台中各功能服务的告警情况，包括当前告警信息查询、历史告警信息查询、当前告警信息确认、告警查询 • 支持配置告警联系人和告警级别 • 支持配置告警通知方式，可选择向企业微信、邮箱地址发送通知
资源监控	资源管理提供主机管理、监控详情的功能：

特性	描述
	• 主机管理：对当前已部署服务或加入系统的主机进行统一管理，并支持查看主机的状态，以便根据使用情况灵活扩容、缩容；此外还支持统一修改所有主机密码 • 监控详情：对部署服务的主机节点进行监控，包括主机节点信息、MySQL 信息和服务信息、重要的系统基础组件信息，帮助用户了解资源使用状况，以便更合理地使用主机资源
防火墙管理	展示了系统的防火墙状态，可执行开启或关闭防火墙的操作，也可以添加或移除白名单，开放或者关闭相关服务端口，同时支持检查防火墙状态及一键同步等功能
系统巡检	系统维护支持一键巡检、定时巡检、查看报告、下载报告、删除巡检记录的功能
系统备份	提供了对系统数据进行备份，并管理备份的功能，支持一键备份和定时备份

3.10 个人中心

个人中心用于用户统一维护个人信息，处理相关申请流程，查看用户的订阅、收藏、工作空间等信息。

表3-17 个人中心功能特性

特性	描述
个人信息	个人信息提供了当前登录用户修改个人信息、修改密码等功能： • 在个人信息页签中，用户可以修改用户名、电子邮箱、手机号码、企业微信账号等信息 • 在修改密码页签中，用户可以通过输入原密码、输入新密码来修改密码
资产订阅	展示了当前用户所在组织下的应用信息及应用下的资产订阅信息： • 应用：展示了当前登录用户所在组织下的所有应用 • 资产：展示了当前登录用户所在组织下申请订阅的资产信息
我的申请	我的申请展示了当前登录用户提交的申请流程列表： • 我的申请支持按照时间段、状态、名称和审批结果等条件进行查询 • 用户可以单击申请对应操作列中的<撤销>按钮，对申请进行撤回处理 • 用户可以单击申请对应操作列中的<删除>按钮，对于申请进行删除处理
待办审批	待办审批展示了当前登录用户待办的审批流程： • 待办审批支持按照时间段、名称、申请人等条件进行查询 • 用户可以单击<同意>意或者<驳回>按钮，对申请流程进行审批 • 用户可以单击<更改责任人>按钮，在责任人下拉框中选择变更后的责任人，进行责任人修改

特性	描述
所有申请	所有申请展示了当前登录用户相关的申请流程： • 所有申请支持按照时间段、状态、名称、审批结果和申请人等条件进行查询 • 用户可以点击申请的名称，查看该申请的详细信息 • 用户可以对自己负责审批的申请流程，执行更改责任人、同意申请、驳回申请操作
我的纠错	展示用户提交的资产纠错信息
我的收藏	展示了当前登录用户收藏的资产信息
我的资产	展示了当前登录用户负责的资产列表，并可对相应资产的纠错和客户评价进行处理
我的评价	以列表的形式展示了当前登录用户已评价的资产

4 共享自行车案例

DMP 数据管理平台是通过数据技术，对海量数据进行采集、清洗、加工，成为标准化、统一化的数据存储，形成大数据资产，进而为客户提供高效的、创新的服务。其中又以数据开发模块最为核心，将实际业务需求抽象为一个个实体，提供基于任务类型的代码组织方式。

本章旨在以业务流程中的“SparkSQL”节点为核心创建离线任务，展示开发及运维等过程的完整使用步骤。

4.1 案例说明

自行车共享系统是一种租赁自行车的方法，注册会员、租车、还车都将通过城市中的站点网络自动完成。通过该系统，人们可以根据需要从一个地方租赁一辆自行车然后骑到自己的目的地归还。

但在共享单车的运维过程中，不时出现一些共享单车长时间未有运行轨迹上报或无法准确监控到位置的情况。这就导致，在 APP 中可以看到很多可用单车，但走到对应位置时，才发现该位置没有单车或无法使用。这些单车中，有些单车不知道被藏到了哪里；有些车或许是在高楼的后面，因 GPS 的误差而找不到；有些车被放到了小区内，一墙之隔使骑车人无法取到单车，甚至使单车直接从 GPS 定位中消失等等。

为了获得这些单车的数据，确认这些车是否已经变成了“僵尸车”，本案例将通过业务流程中的 SparkSQL 节点将单车基本数据信息与实时更新的数据信息进行关联分析，筛出“僵尸车”。

4.2 准备操作

4.2.1 配置运行环境

本例使用根组织下的默认工作空间 defaultWorkspace，在进行数据处理的配置前，需要给改工作空间配置集群资源。

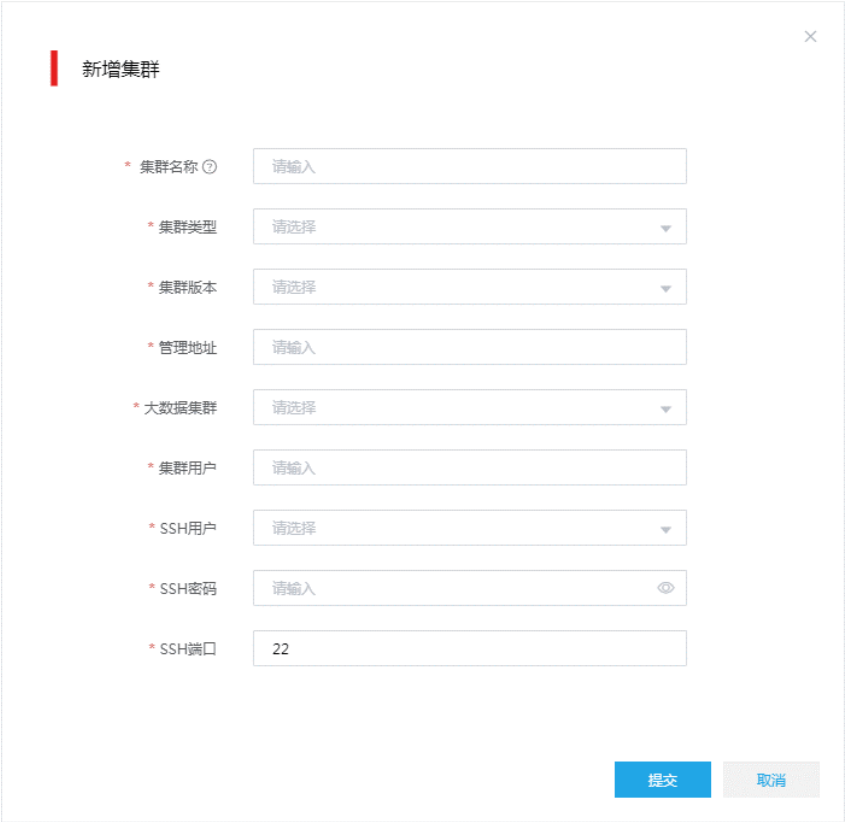
(1) 在顶部导航栏中选择[系统]，进入系统模块。

(2) 在左侧导航树中，选择[大数据环境管理/大数据集群管理]菜单项，进入大数据集群管理页面。

(3) 在页面中，确认是否已存在集群。如已存在，请直接执行下一步骤。如不存在，请增加集群，步骤如下：

a. 单击列表上方的<配置>按钮，弹出新增集群窗口。

图4-1 新增集群



b. 配置集群的参数信息。

- 集群名称：配置大数据集群在本系统中的名称，不可使用 default 关键字作为集群名称。
- 集群类型：选择集群的类型，即对接的大数据集群类型。
- 集群版本：选择集群的版本，对于 BDP 大数据平台类型集群，支持 E6105 版本。
- 管理地址：输入大数据集群所在大数据平台的管理地址。输入完成后，系统会自动获取该地址对应大数据平台中的信息。
- 大数据集群：从下拉列表中选取需要的大数据集群。列表中的集群均为系统从大数据平台中自动获取，如无可选值，说明大数据平台中为创建符合要求的集群，或本系统与大数据平台连接故障。
- 集群用户：系统自动根据所选的大数据集群填充。
- SSH 用户：配置大数据平台的 SSH 用户名，数字平台部分功能需要该 SSH 用户名和密码连接大数据平台节点，修改配置文件，进行功能适配。
- SSH 密码：配置 SSH 用户的登录密码。
- SSH 端口：指定 SSH 连接使用的端口，默认为 22。

c. 单击<提交>按钮，集群配置完成。

(4) 在左侧导航树中，选择[大数据环境管理/运行环境管理]菜单项，进入运行环境页面。

(5) 在“组织”页签中，确认是否已为根组织配置了集群资源。如已配置，请直接执行下一步骤。如未配置，请为根组织配置集群资源：

a. 单击列表上方的<配置>按钮，弹出分配运行环境窗口。

图4-2 配置组织运行环境

b. 配置组织的运行环境参数。

- 集群名称：从下拉列表中选择大数据集群。列表中的集群即之前配置的大数据集群。
- 集群用户：从下拉列表中选择大数据集群的用户。
- 集群度列：从下拉列表中选择使用的队列，本例中选择 root.default。
- 组织：选择集群资源分配给的目标组织。本例中选择根组织。

c. 单击<提交>按钮，配置完成。

(6) 切换至“工作空间”页签，确认是否已为根组织的默认工作空间 defaultWorkspace 配置了集群资源。如已配置，请直接执行下一步骤。如未配置，请为工作空间配置资源：

a. 单击列表上方的<配置>按钮，弹出分配运行环境窗口。

图4-3 配置工作空间运行环境

b. 配置工作空间的运行环境参数。

- 集群名称：从下拉列表中选择大数据集群，可选择的集群为之前分配给工作空间所属组织的集群。
- 集群用户：选择集群后会自动填充。
- 集群度列：选择集群后会自动填充。
- 任务并行度：配置工作空间中，任务执行的并行度，本例中使用默认值。
- 工作空间：选择集群资源分配给的目标工作空间。本例中选择根组织下的默认工作空间defaultWorkspace。

c. 单击<提交>按钮，配置完成。

(7) 至此，运行环境配置完成。

4.2.2 新增数据源

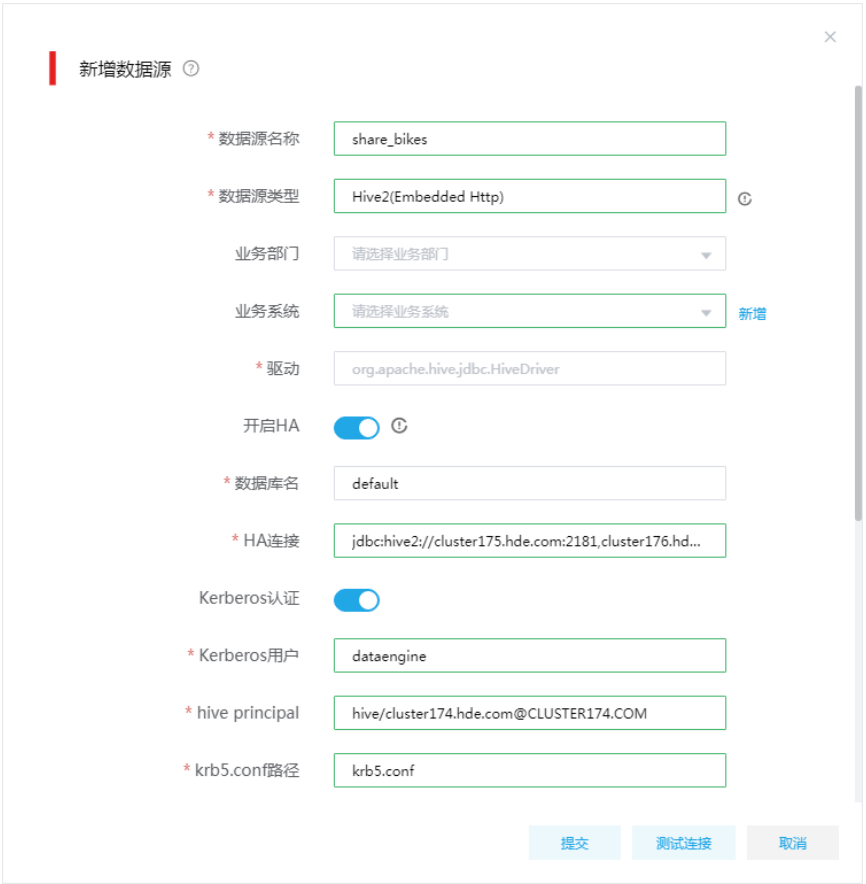
在通过业务流程对需要分析的数据源进行操作之前，需要将待分析的数据源增加到数字平台中。当前数字平台支持丰富的数据源类型，用户可选择合适的数据源并进行配置。本说明文档将以 Hive 存储类型的数据源为例说明数据源的新增过程。

(1) 在[工程配置/数据源管理]模块中，单击左上角<新增>按钮，进行数据源的增加操作，如图 4-4 所示。

图4-4 数据源配置页面

(2) 选择 Hive 数据源，并配置参数，图 4-5 所示为新增 Hive 数据源所需填写的信息模板。

图4-5 新增 Hive 数据源页面



(3) 填写完毕增加数据源所需要的信息之后，可以单击<测试连接>按钮，测试数据源连通性。

(4) 提示“连接测试成功”信息，单击<确定>按钮，执行增加数据源。之后即可在数据源列表中看到增加成功的数据源概要信息。

4.2.3 采集元数据

(1) 成功新增数据源之后，进入[DMP 数据管理/数据治理]模块，按照图 4-6 中所示的 1、2、3 步骤，进入元数据采集的任务创建页面，并创建元数据采集任务。

图4-6 元数据采集配置

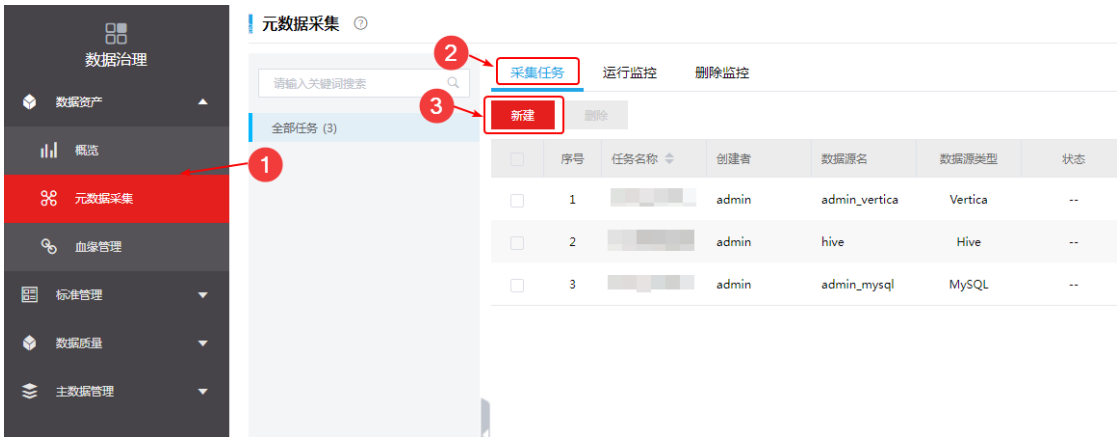


图4-7 新增采集任务

(2) 创建完成后，在列表中单击对应操作列中的<运行>按钮，启动采集任务，如图 4-8 所示，将数据源 share_bikes 中的所有 Hive 表采集到 DMP 数据管理中。

图4-8 启动元数据采集

(3) 采集任务启动之后，选择“运行监控”页签，进入采集任务监控列表查看采集任务执行进度，如图 4-9 所示。

图4-9 元数据任务监控

采集任务也可以设置调度周期为定期或周期性的对某个数据源执行采集操作，更新已采集的数据表信息。

4.2.4 注册离线表

Hive 数据源的表在采集到的元数据后，会被自动注册为离线表，其他类型数据源（MySQL/Oracle/PostgreSQL 等）中的数据表需要在[数据开发/表管理]中执行“离线表注册”操作后才可在业务流程的 SparkSQL 节点中使用。

4.3 构建业务流程

业务流程是按照业务的种类将相关的不同类型的节点任务组织在一起所构成的有向无环图。本章中以 SparkSQL 节点为例介绍离线分析任务在业务流程中的使用步骤。

4.3.1 创建业务流程

(1) 进入[DMP 数据管理/数据开发]模块。

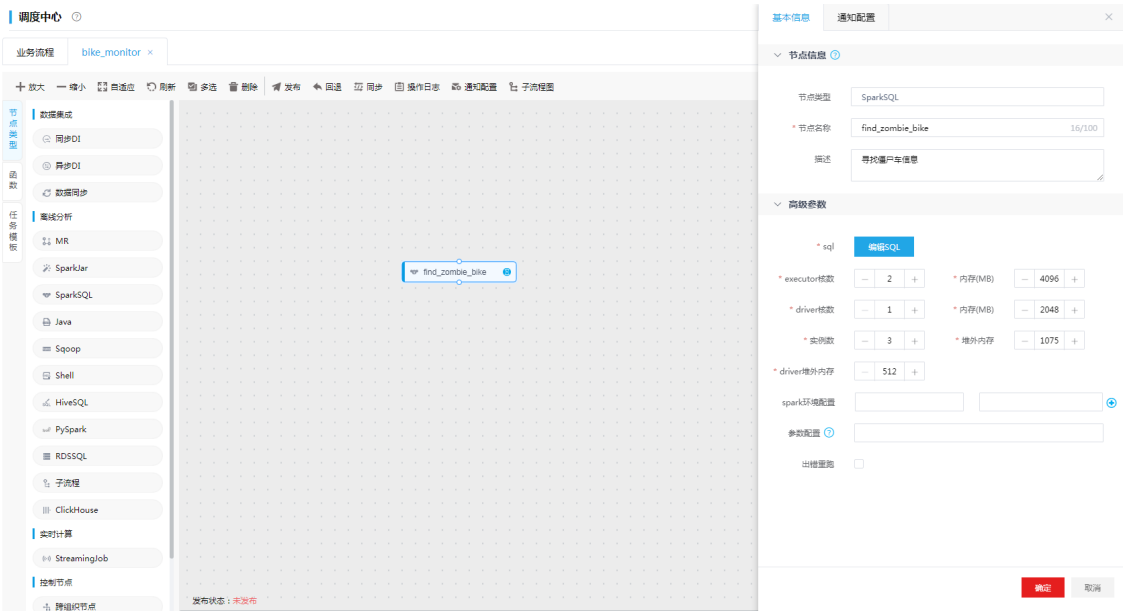
(2) 在调度中心页面中，单击<新建>按钮，弹出新建业务流程窗口，如图 4-10 所示。

图4-10 新建业务流程

(3) 填写业务流程及名称，单击<确定>按钮，新建业务流程。

(4) 进入业务流程画布编辑页面，拖拽一个 SparkSQL 节点到画布中，双击弹出右侧边栏，如下图 4-11 所示。

图4-11 查看 SparkSQL 节点信息



4.3.2 编辑 SparkSQL 节点

(1) 编辑 SparkSQL 节点信息，根据提示填写必要的节点名称等信息。

(2) 单击<编辑 SQL>按钮，弹出 SQL 智能编辑器窗口，如图 4-12 所示。

图4-12 SQL 智能编辑器开发 SQL

(3) 在此窗口中编写符合 SparkSQL/HiveSQL 语法规范的 SQL 语句，所使用的分析表即为通过元数据采集操作注册进来的 Hive 数据表。

本例所述的识别“僵尸车”信息的分析 SQL 语句，其执行的操作是对单车基本数据信息表 share_bikes.bikes_record 和实时更新的数据信息表 share_bikes.bikes_history 做 left join 操作，获取近 30 天内未有实时数据的车辆即推测为“僵尸车”。

一般而言，对于需要通过 SparkSQL 节点任务执行的离线分析操作，建议将分析结果保存至数据表中。将分析结果保存至结果表可以便于后续查看分析结果或作为其他操作的数据源使用。本例中，在保存执行识别僵尸车的分析 SQL 语句之前，对其添加结果表 share_bikes.zombie_bike_info。

将 SQL 语句改写为“create... select...”方式，如图 4-13 所示。

图4-13 通过 SQL 创建结果表

(4) 编辑完成 SQL 语句后，可以通过单击<语法校验>按钮，校验 SQL 语法是否正确规范。在不追加结果表保存的情况下，也可以单击<执行>按钮或<选中执行>按钮，直观地查看结果。

(5) SQL 智能编辑器窗口的各参数填写完毕后，单击<确定>按钮，退出编辑 SQL 窗口，返回业务流程画布编辑页面。

(6) 在页面右侧边栏中配置其他几项可选填的参数，具体说明请参阅表 4-1。

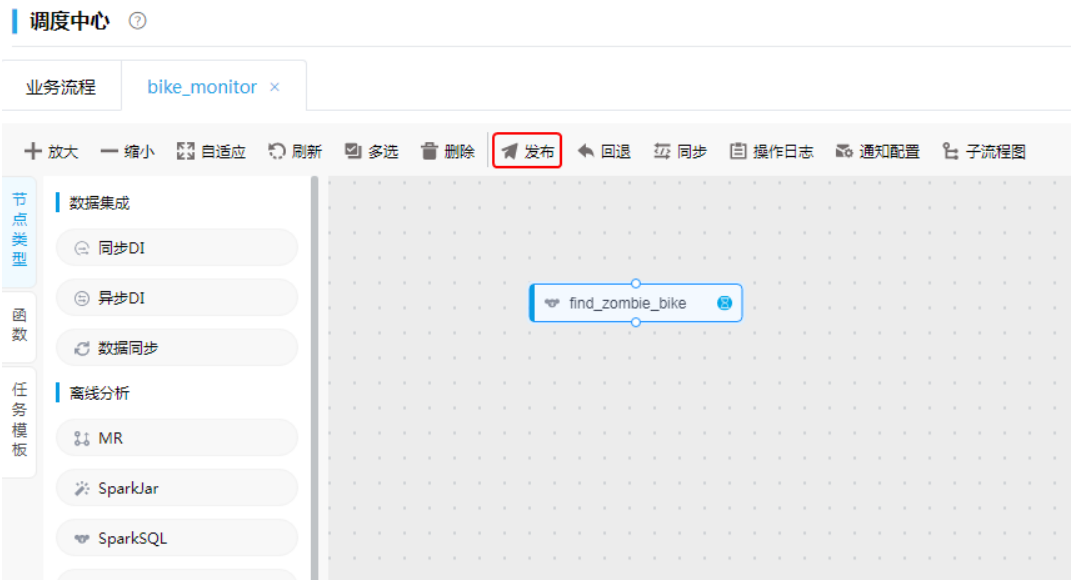
表4-1 SparkSQL 节点选填信息说明汇总

参数名	参数说明	示例值
配置参数	配置执行节点的基本配置参数	• executor 核数：2 • 内存(MB)：4096 • driver 核数：1 • 内存(MB)：2048 • 实例数：3 • 堆外内存：1075 • driver 堆外内存：512
参数配置	如在SparkSQL中需要写入动态参数，如EL表达式等，该参数配置用于传递实际的参数名和参数值	如果SparkSQL代码为“select * from default.table where start_date='$biztime' and end_date='$cyctime'”，其参数配置值为$biztime= 2021-06-10 $cyctime=2021-06-11

4.3.3 提交执行

(1) 保存 SparkSQL 节点之后，单击业务流程画布左上方的<发布>按钮，发布业务流程。

图4-14 发布业务流程

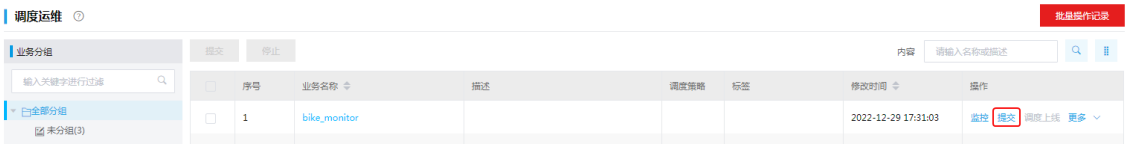


(2) 单击<确定>按钮，业务流程发布完成。

(3) 在左侧导航树中选择[运维管理/调度运维]菜单项，进入调度运维页面。

(4) 在业务流程列表中，单击业务流程对应操作列的<提交>按钮，如图 4-15 所示。

图4-15 提交业务流程



4.3.4 任务监控

提交业务流程之后，可在调度运维页面业务流程列表中，单击业务流程对应操作列的<监控>按钮，如图 4-16 所示。

图4-16 业务流程任务列表

单击<监控>按钮，进入该业务流程的监控画布页面，提交的 SparkSQL 节点上显示监控属性的实时变化如图 4-17 示。在画布中右键单击该节点，弹出菜单中会显示该节点相关的“运行详情”、“查看日志”等监控属性，可查看该节点的具体监控信息，如图 4-18 所示。

图4-17 业务流程执行状态信息展示

图4-18 业务流程节点执行信息查询

4.3.5 调度配置（可选）

发布后的业务流程实例，支持配置调度策略。

为业务流程实例配置调度策略的步骤如下：

(1) 在[运维管理/调度运维]页面中，单击业务流程实例操作列<更多>按钮，并在下拉菜单中选择[调度配置]菜单项，如图 4-19 所示。

图4-19 配置调度

(2) 在弹出的调度策略配置窗口中，配置调度参数，如图 4-20 所示。

图4-20 节点配置调度信息

(3) 根据提示信息填写调度策略参数。

(4) 调度策略配置完成后，单击<确定>按钮，调度配置完成。

(5) 单击窗口中的<上线>按钮，调度配置上线生效。

(6) 单击业务流程实例操作列<监控>按钮，进入业务流程实例的监控界面。

(7) 在监控页面中，左侧边条会展示任务的调度运行实例信息。双击节点，可在页面右侧弹出节点的信息，其中会展示节点的调度运行信息，如图 4-21 和图 4-22 所示。

图4-21 运行实例查询

图4-22 运行信息

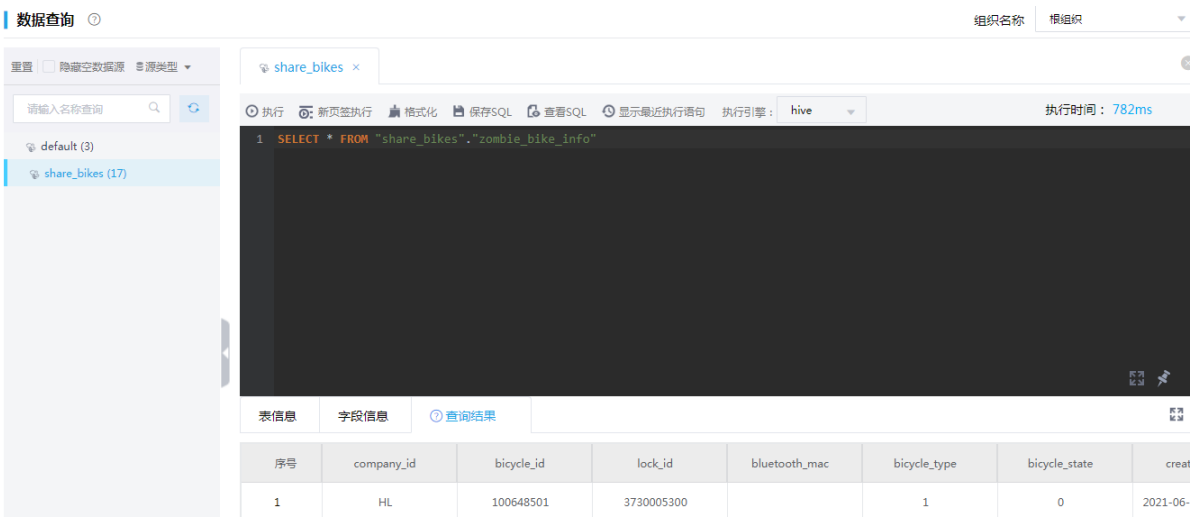
4.4 结果查看

对于未设置“结果导出”操作但带有结果表的 SparkSQL 节点（即 SQL 语句以“create...”或“insert...”等开头）执行成功之后，可以通过[DMP 数据管理/数据探查]模块中的数据查询页面直接查询，如图4-23 所示。

由于本例所使用的案例的结果表存于数据源 share_bikes 中，所以可以再执行一次元数据采集任务，将该表采集到元数据信息中，从而可以在[数据查询]中引用到该表名，再对其执行如下命令同样可以查询结果。

```
select * from "share_bikes"."zombie_bike_info"
```

图4-23 任务结果查询



5 疫苗接种监控案例

5.1 案例说明

在疫情防控中，接种疫苗作为重要的一环，是控制病毒传染扩散的重要手段。随着疫苗的推广，接种人员越来越多，为准确迅速地掌握疫苗接种情况，需要对登记的人员信息、人员类别信息、辖区信息、人员接种信息等进行汇总计算，并将结果提供给大屏展示。

为提供大屏展示所需的数据，需在数据库建立业务数据表（创建业务数据表的参考 SQL 语句请参见 7.2 疫苗接种案例业务数据库建表语句示例），记录原始数据，并对这些数据进行处理和计算，得出如下 4 个指标：

- 行业接种统计数据
- 社区街道接种统计数据
- 各年龄段接种统计数据
- 一针接种至今各个时间段接种人数统计

本案例涉及使用 MySQL（已有的业务数据库）、HDFS（存储 Hive 数据文件）、Hive（存储待处理的数据）、Greenplum 数据库（存储处理后的数据），请提前部署好相关的数据库，部署操作请参见各数据库产品的相关文档。

主要步骤及说明如下：

(1) 本例中的数据来源于业务数据库中的原始数据，而为了保证这些原始数据信息不受影响，需要通过 iPaaS 集成平台的 DI 将业务库中的原始数据（存量和增量）抽取至数字平台的 ODS（Operation Data Store）层中，作为基础数据。

(2) 将 ODS 层中保存的基础数据，在 DMP 数据管理中注册为数据源，以便后续步骤中使用。

(3) 在 DMP 数据管理中，创建对应基础数据的表，表的结构需要与基础数据存储表的结构一致，使系统能够正确读取识别基础数据。此外，还需创建存放清洗后数据的表和统计结果的表，以便在后续步骤中使用。

(4) 在 DMP 数据管理中，对基础数据进行清洗处理，并将清洗后的数据存放至专用的数据表中。

(5) 在 DMP 数据管理中，对清洗后的数据进行计算等处理，得出统计结果数据，并存放至预先准备好的表中。

(6) 对统计结果数据进行查询验证，无误后即可通过 iPaaS 集成平台的服务集成进行发布和授权。第三方应用可以调用数据结果用于大屏展示等。

说明

本例中各配置均使用默认的根组织下的“defaultWorkspace”工作空间。

5.2 准备操作

在进行数据处前，需要先进行基础准备工作，包括：准备工作环境、新增数据源（含数据抽取使用的数据源和数据处理使用的数据源）、抽取基础数据、新建数据处理使用的数据表。

5.2.1 配置运行环境

本例使用根组织下的默认工作空间 defaultWorkspace，在进行数据处理的配置前，需要给改工作空间配置集群资源。

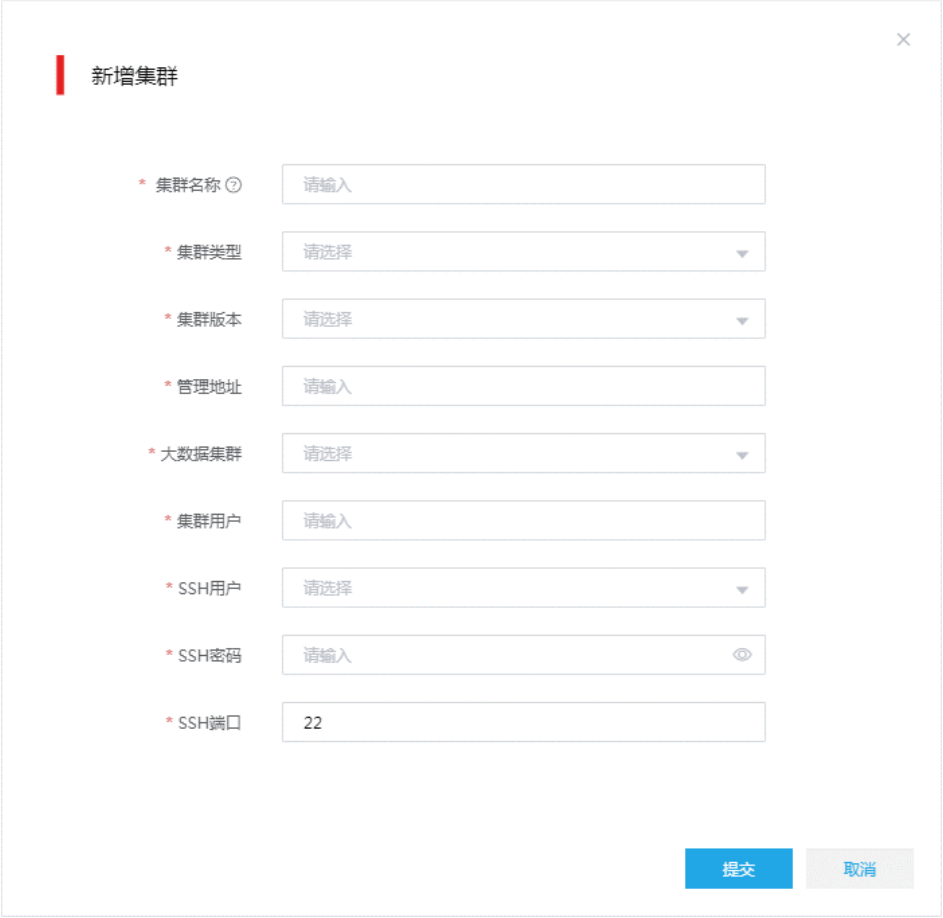
(1) 在顶部导航栏中选择[系统]，进入系统模块。

(2) 在左侧导航树中，选择[大数据环境管理/大数据集群管理]菜单项，进入大数据集群管理页面。

(3) 在页面中，确认是否已存在集群。如已存在，请直接执行下一步骤。如不存在，请增加集群，步骤如下：

a. 单击列表上方的<配置>按钮，弹出新增集群窗口。

图5-1 新增集群



b. 配置集群的参数信息。

- 集群名称：配置大数据集群在本系统中的名称，不可使用 default 关键字作为集群名称。
- 集群类型：选择集群的类型，即对接的大数据集群类型。
- 集群版本：选择集群的版本，对于 BDP 大数据平台类型集群，支持 E6105 版本。
- 管理地址：输入大数据集群所在大数据平台的管理地址。输入完成后，系统会自动获取该地址对应大数据平台中的信息。

– 大数据集群：从下拉列表中选取需要的大数据集群。列表中的集群均为系统从大数据平台中自动获取，如无可选值，说明大数据平台中为创建符合要求的集群，或本系统与大数据平台连接故障。
– 集群用户：系统自动根据所选的大数据集群填充。
– SSH 用户：配置大数据平台的 SSH 用户名，数字平台部分功能需要该 SSH 用户名和密码连接大数据平台节点，修改配置文件，进行功能适配。
– SSH 密码：配置 SSH 用户的登录密码。
– SSH 端口：指定 SSH 连接使用的端口，默认为 22。

c. 单击<提交>按钮，集群配置完成。

(4) 在左侧导航树中，选择[大数据环境管理/运行环境管理]菜单项，进入运行环境页面。

(5) 在“组织”页签中，确认是否已为根组织配置了集群资源。如已配置，请直接执行下一步骤。如未配置，请为根组织配置集群资源：

a. 单击列表上方的<配置>按钮，弹出分配运行环境窗口。

图5-2 配置组织运行环境

b. 配置组织的运行环境参数。
– 集群名称：从下拉列表中选择大数据集群。列表中的集群即之前配置的大数据集群。
– 集群用户：从下拉列表中选择大数据集群的用户。
– 集群度列：从下拉列表中选择使用的队列，本例中选择 root.default。
– 组织：选择集群资源分配给的目标组织。本例中选择根组织。

c. 单击<提交>按钮，配置完成。

(6) 切换至“工作空间”页签，确认是否已为根组织的默认工作空间 defaultWorkspace 配置了集群资源。如已配置，请直接执行下一步骤。如未配置，请为工作空间配置资源：

a. 单击列表上方的<配置>按钮，弹出分配运行环境窗口。

图5-3 配置工作空间运行环境

b. 配置工作空间的运行环境参数。

- 集群名称：从下拉列表中选择大数据集群，可选择的集群为之前分配给工作空间所属组织的集群。
- 集群用户：选择集群后会自动填充。
- 集群度列：选择集群后会自动填充。
- 任务并行度：配置工作空间中，任务执行的并行度，本例中使用默认值。
- 工作空间：选择集群资源分配给的目标工作空间。本例中选择根组织下的默认工作空间defaultWorkspace。

c. 单击<提交>按钮，配置完成。

(7) 至此，运行环境配置完成。

5.2.2 新增数据源

说明

新增数据源需要使用组织管理员级别的用户账号。

本节介绍了将数据源新增至数字平台的操作步骤。新增数据源包括抽取基础数据使用的数据源和数据处理业务流程使用的数据源。

- 抽取基础数据使用的数据源

  包括 MySQL 数据源（用户业务库，存储基础数据）和 HDFS 数据源（承载从用户业务库中抽取的基础数据）。

  业务数据库为 MySQL，从业务数据库中抽取的数据需要存储至数字平台的 ODS 层数据源中。本例 ODS 层数据源使用 Hive 数据源，由于 Hive 数据文件存储在 HDFS 中，为提高性能，抽

取基础数据步骤直接使用 HDFS 数据源，抽取到的数据会直接写入 HDFS 中的 Hive 数据文件中，该方式比通过 JDBC 方式写入 Hive 数据表中的速度更快。

- 数据处理业务流程使用的数据源

  包括 Hive 数据源（存储抽取到的基础数据）和 Greenplum 数据源（存储处理后的数据）。

  将 Hive 数据文件（即 HDFS 中承载抽取数据的 Hive 数据文件）所属的 Hive 数据源作为数字平台的 ODS 层数据源。Hive 中的数据经过业务流程计算处理后，结果数据存放在 Greenplum 数据源中，作为 DWS 层数据源。

1. 新增 MySQL 数据源

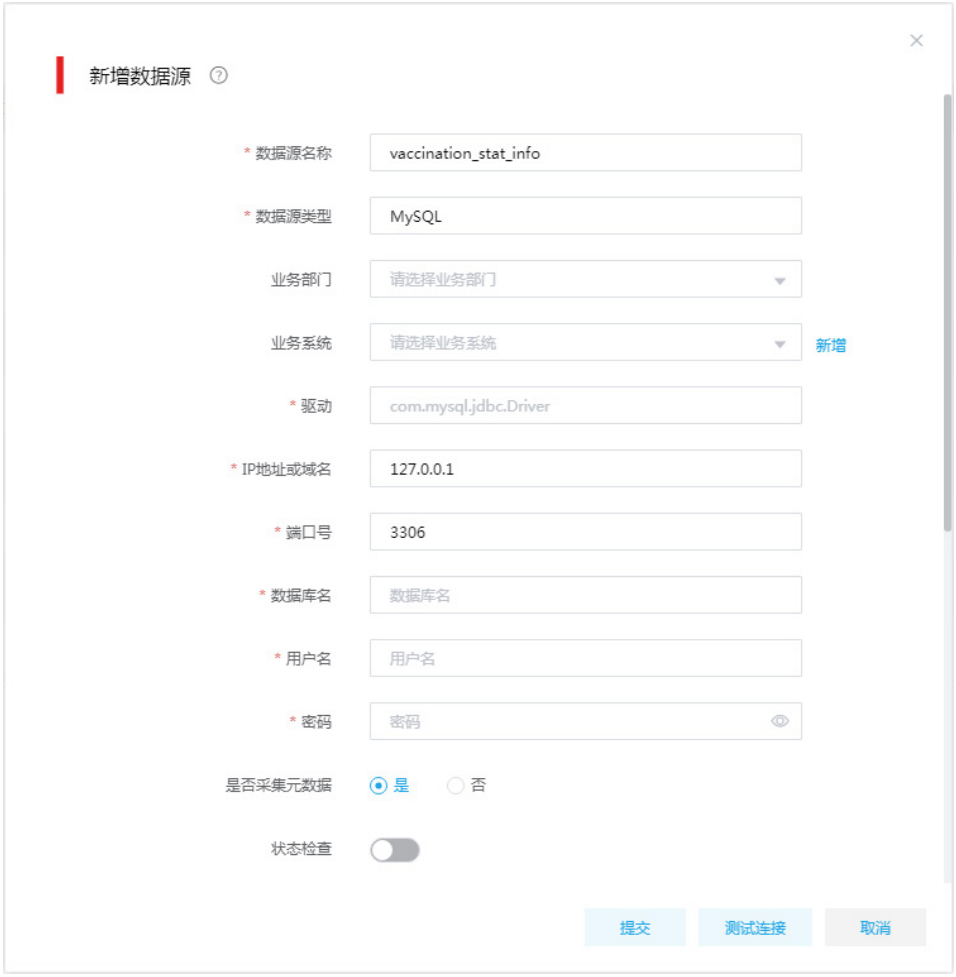
将业务数据库 MySQL 增加至数字平台中。

(1) 在[工程配置/数据源管理]页面中，单击左上角<新增>按钮，进行数据源的创建操作，如图 5-4 所示。

图5-4 数据源配置页面

(2) 如图 5-5 所示，在弹出窗口中配置业务数据库 MySQL 的信息，如 IP 地址，用户名和密码等信息。

图5-5 新增 MySQL 数据源



新增 MySQL 数据源时，部分参数说明如下：

- 业务部门：选填，指定数据源的归属部门，也即本系统中的组织。如不需与部门关联，可为空。
- 业务系统：选填，指定数据源所属的业务系统。如不需与业务系统关联，可为空。
- 驱动：缺省填入，不可修改。
- IP 地址或域名：必填，目标数据库所在的 IP 地址或域名。
- 端口号：必填，MySQL 数据库使用的端口号，缺省为 3306。
- 数据库名：必填，待连接的已存在的数据库名称。
- 用户名：必填，能够访问对应数据库的用户名。
- 密码：必填，用户名对应的登录密码。
- 是否采集元数据：必选，如果置为启用，则会在添加数据源成功后，自动在资产中心中创建相应的元数据采集任务并执行；否则不创建采集任务。
- 状态检查：如果置为启用，则系统会定时检测数据源的连接状态，并会在数据源列表中展示数据源的最新状态（正常、异常）；如果置为不启用，则系统不会检测，并将数据源的状态展示为“未知”。默认为不启用。
- 描述信息：选填，自定义的描述信息。

- 属性列表：选填，数据源的扩展属性，本案例不配置。关于详细属性请参考数据库相关官方文档。

(3) 配置完成后，单击<测试连接>按钮，可检查所填写的信息是否无误，如果测试通过，即证明数据库信息可用。

(4) 提示“连接测试成功”信息后，单击<提交>按钮，完成数据源新增。之后即可在 DI 作业中使用该数据源。

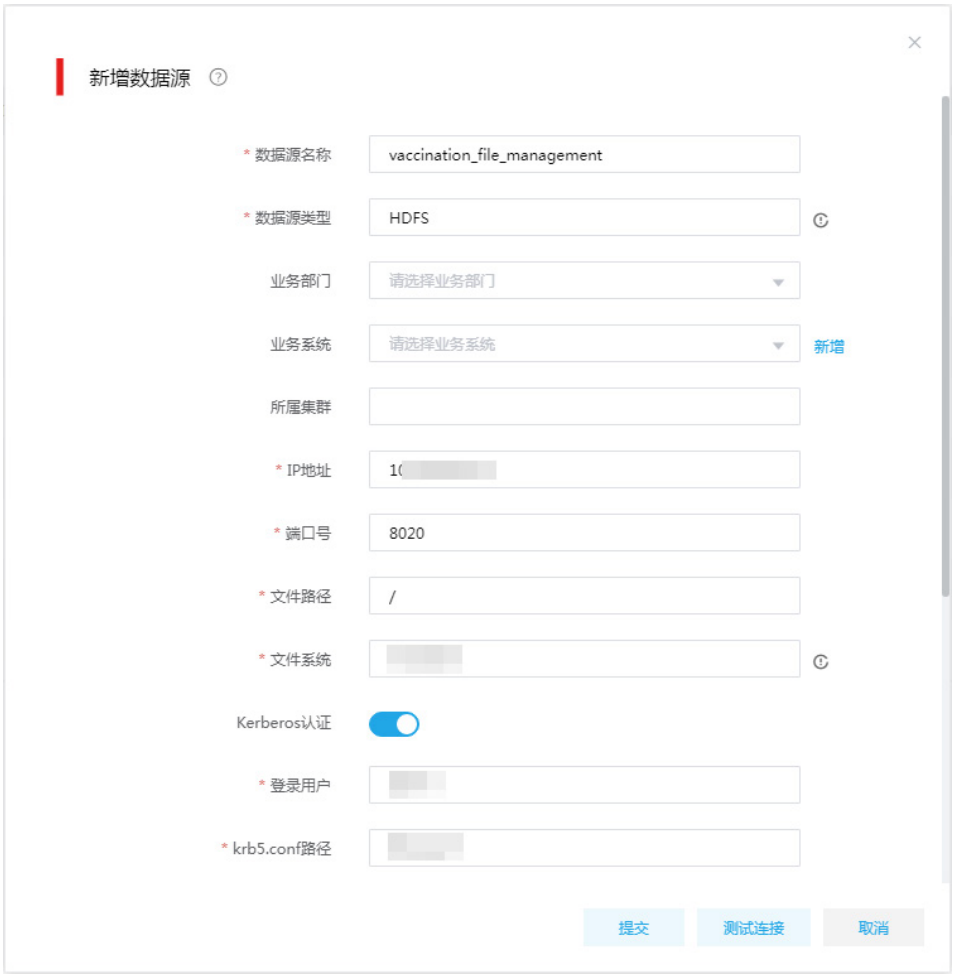
2. 新增 HDFS 数据源

将 HDFS 作为数据源增加至数字平台中，以便将抽取到的数据直接写入 HDFS 中的 Hive 数据文件中。通过将 HDFS 作为数据源，直接将数据写入其中的 Hive 数据文件，具有更高的速度，该操作相当于写入 Hive 的外部表。

(1) 在[工程配置/数据源管理]模块中，单击左上角<新增>按钮，进行数据源的创建操作。

(2) 在弹出窗口中配置 HDFS 的信息，如 IP 地址，端口号，文件路径，登录用户等信息。

图5-6 新增 HDFS 数据源



(3) 配置完成后，单击<测试连接>按钮，可检查所填写的信息是否无误，如果测试通过，即证明数据库信息可用。

(4) 提示“连接测试成功”信息后，单击<提交>按钮，完成数据源新增。之后即可在 DI 作业中使用该数据源。

3. 新增 Hive 数据源

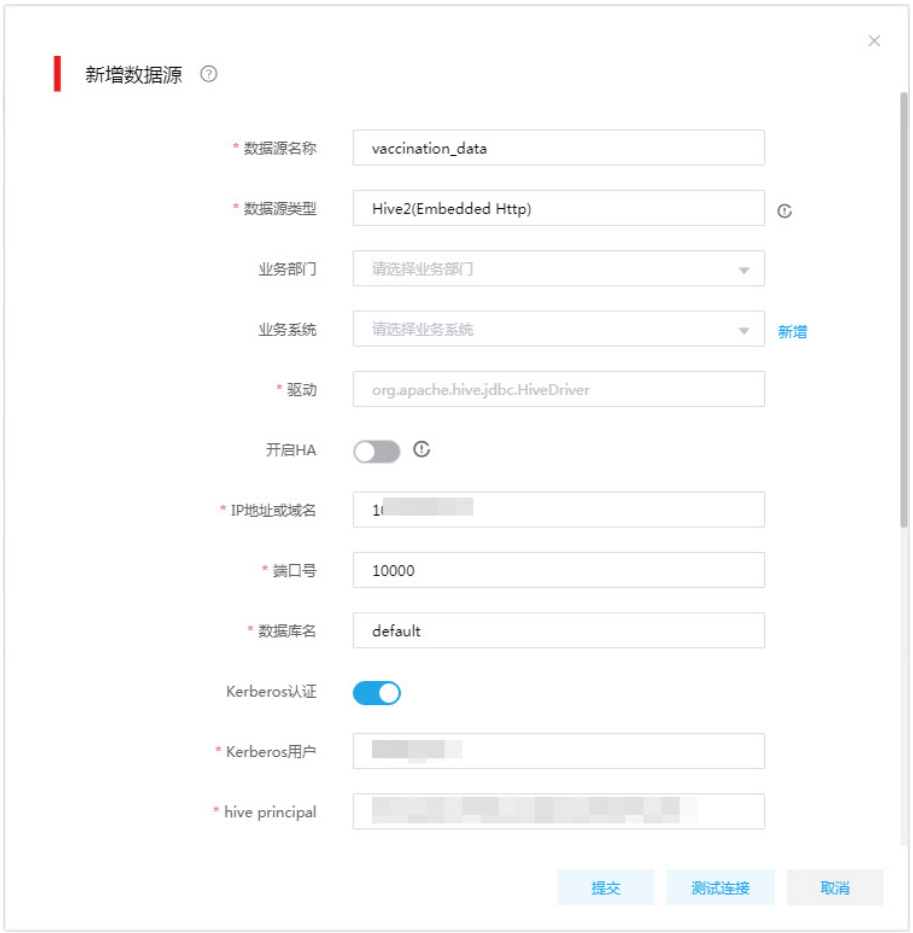
(1) 在[工程配置/数据源管理]模块中，单击左上角<新增>按钮，进行数据源的创建操作。

(2) 选择 Hive 数据源，并配置参数，如图 5-7 所示。其中：

- Kerberos 用户等信息可以在 DMP 数据管理所使用的大数据集群页面中查看。
- hive principal 参数格式为：hive/*IP 地址对应节点的主机名@集群名称大写*.COM。
- krb5.conf 和 Keytab 文件为 Kerberos 认证文件，需要从 BDP 大数据平台中的集群管理页面下载。

(3) 在 BDP 大数据平台的集群中获得 Kerberos 的相关信息和认证文件后，返回 DMP 数据管理的新建数据源页面，如图 5-7 所示，填写各个 Kerberos 参数并上传所需的认证文件即可。

图5-7 新增 Hive 数据源



(4) 填写完毕注册数据源所需要的信息之后，可以单击<测试连接>按钮，测试数据源连通性。

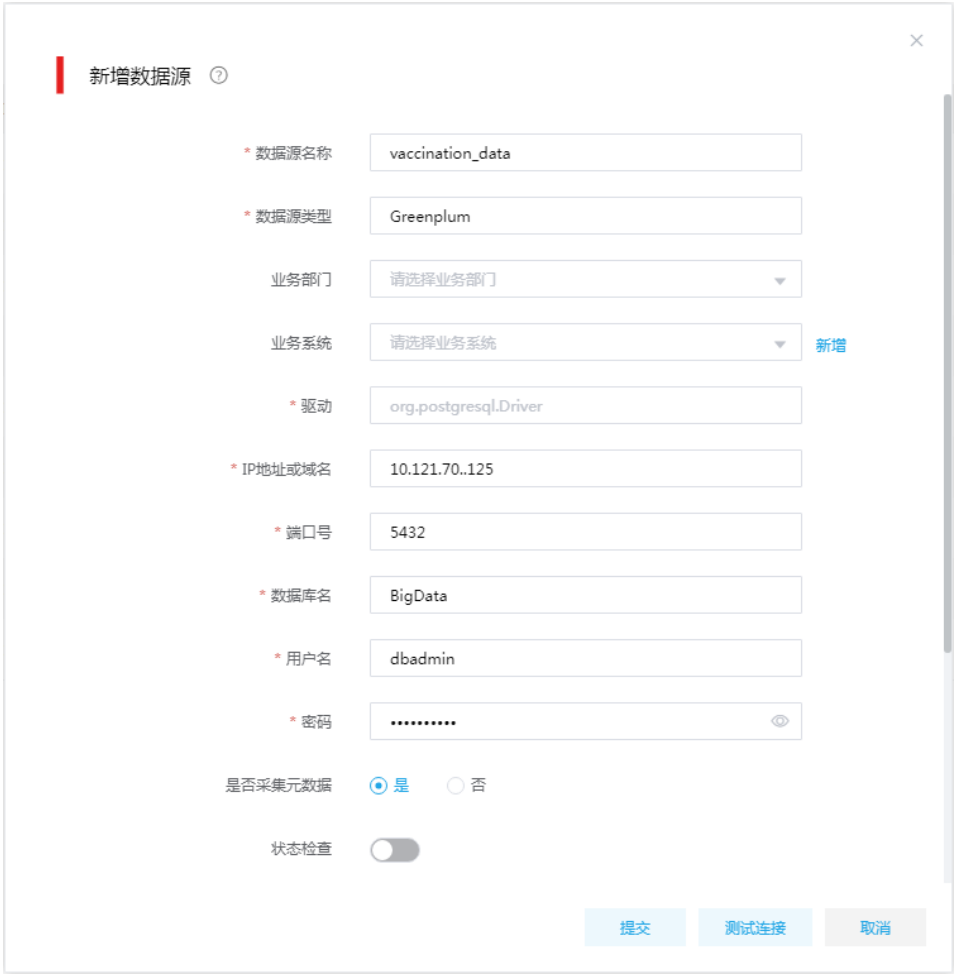
(5) 提示“连接测试成功”信息后，单击<提交>按钮，执行注册数据源。之后即可在数据源列表中看到注册成功的数据源概要信息。

4. 创建 Greenplum 数据源

(1) 在[数据源管理]模块中，单击左上角<新增>按钮，进行数据源的创建操作。

(2) 选择 Greenplum 数据源，并配置参数，如图 5-8 所示。

图5-8 新增 Greenplum 数据源



(3) 填写完毕注册数据源所需要的信息之后，可以单击<测试连接>按钮，测试数据源连通性。

(4) 提示“连接测试成功”信息后，单击<提交>按钮，执行注册数据源。之后即可在数据源列表中看到注册成功的数据源概要信息。

5.2.3 抽取基础数据

基础数据的抽取需要通过 iPaaS 集成平台的数据集成服务完成，通过在 iPaaS 集成平台-数据集成中创建 ETL 作业，实现将用户业务库（MySQL）中的数据抽取至 HDFS 中的 Hive 数据文件中（ODS 层数据）。

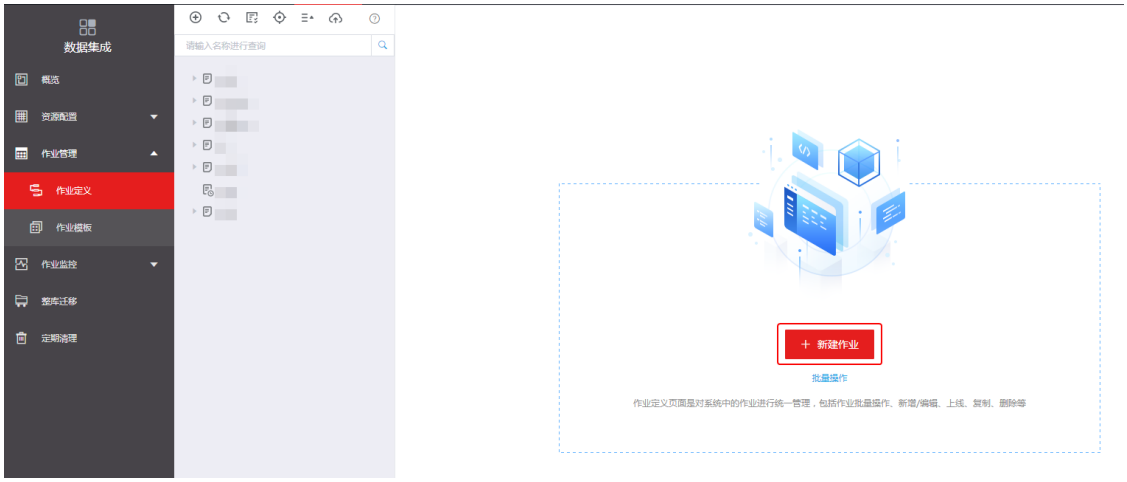
本例中涉及建立如下作业抽取对应的基础数据：

- 抽取人员信息数据的 DI 作业
- 抽取人员接种信息数据的 DI 作业
- 抽取人员分类信息数据的 DI 作业
- 抽取街道信息数据的 DI 作业

作业的创建方式相同，以建立抽取人员信息数据的 DI 作业为例介绍操作步骤。

(1) 进入[iPaaS 集成平台/数据集成]模块的[作业管理/作业定义]页面，单击<新建作业>按钮。

图5-9 作业定义



(2) 在窗口中配置作业的基本参数。

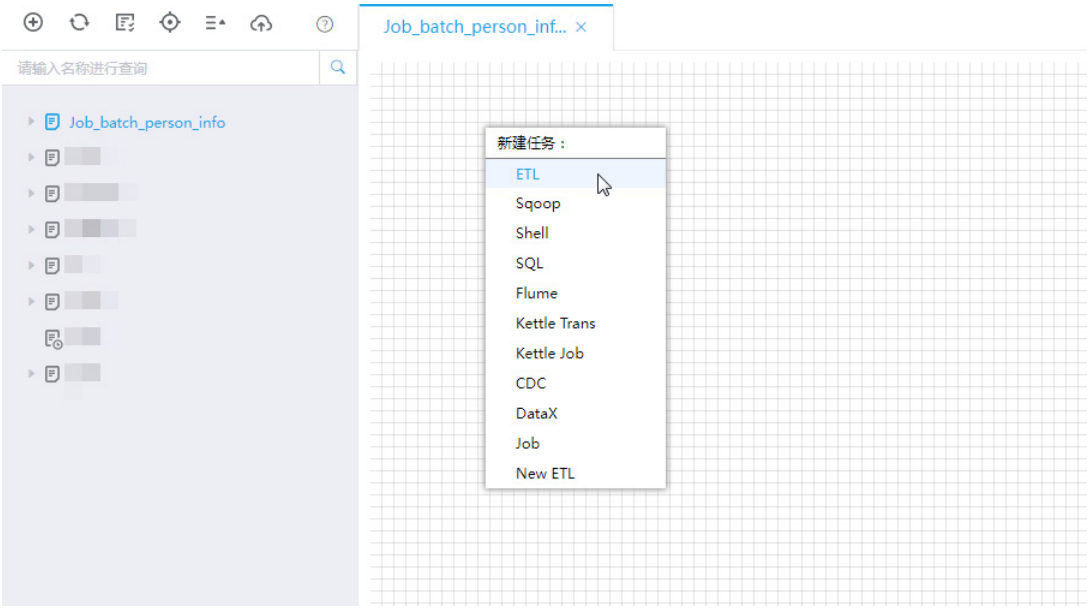
图5-10 抽取人员信息数据

(3) 单击<确定>按钮，作业新增成功。

(4) 在左侧目录中，左键双击作业，进入作业编辑画布页面。

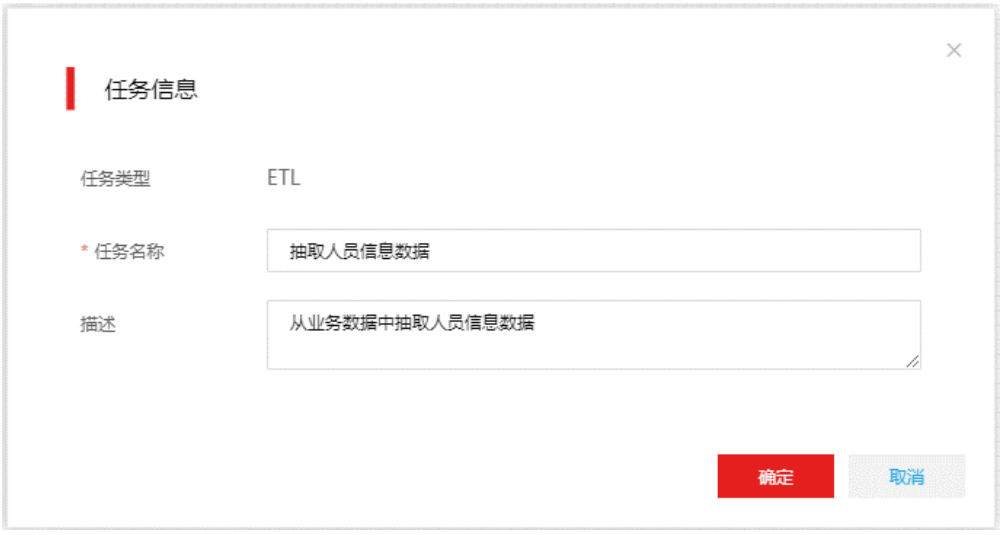
(5) 在画布中，单击右键，在弹出菜单中选择 ETL，弹出任务信息窗口。

图5-11 选择 ETL 任务



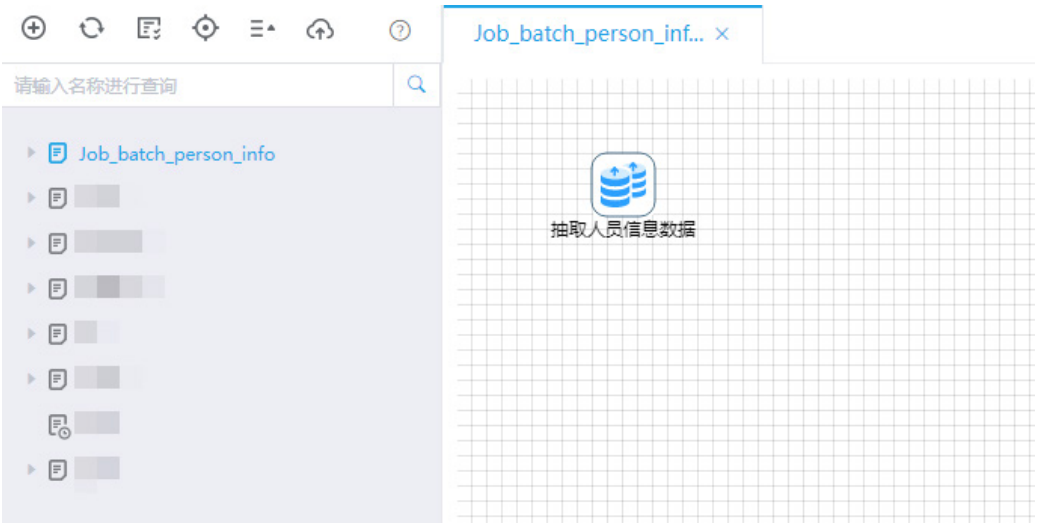
(6) 在任务信息窗口中，配置任务的名称和描述。

图5-12 配置 ETL 任务信息



(7) 单击<确定>按钮，任务基本信息配置完成。

图5-13 ETL 任务基本信息配置完成



(8) 双击任务图标，进入任务编辑画布页签。

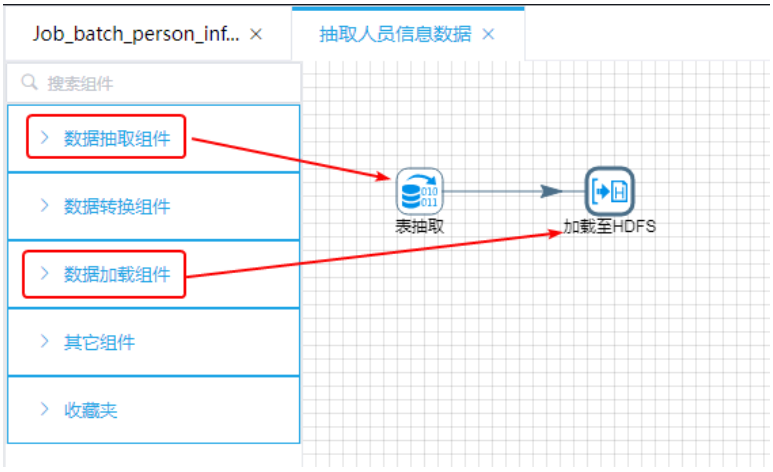
(9) 选取数据抽取组件中的“表抽取”组件，然后选取数据加载组件列表中的“加载至 HDFS”组件。

(10) 在“表抽取”组件中选择数据库连接（数据源），并输入检索 SQL 语句，从 MySQL 数据库中抽取数据。

(11) 在“加载至 HDFS”组件中设置 HDFS 路径等参数。

(12) 右键单击“表抽取”组件图标，在弹出菜单中选择[建立连接]菜单项，并将连线连至“加载至 HDFS”组件，建立数据抽取组件和数据加载组件之间的联系。

图5-14 编辑业务画布



(13) 任务配置完成后，单击右上角的<保存>按钮，任务保存完成。

(14) 关闭任务页签，返回作业画布页面，单击右上角的<保存>按钮，作业新建完成。

(15) 重复步骤 1-14，分别建立抽取人员接种信息数据的 DI 作业、抽取人员分类信息数据的 DI 作业、抽取街道信息数据的 DI 作业。

(16) 在作业定义页面中左侧的作业列表中，右键单击各作业，在弹出菜单中选择[上线]菜单项，上线 DI 作业。

(17) 双击作业进入作业画布，然后双击任务图标，进入任务画布。

(18) 依次选中各作业，并单击 图标，执行 DI 作业任务，抽取基础数据。为实现每天抽取增量数据，可以单击右侧的“调度配置”页签，配置 DI 作业定时执行，获取最新基础数据，定时配置可参见相关联机帮助。

5.2.4 新建业务流程中使用的数据表

本例中，需要根据 Hive 数据源中的基础数据新建对应的数据表，并提前创建针对人员接种信息数据的清洗表以及后续存储数据处理结果的各结果表。

1. 新建基础信息表

为使系统能够正确识别 Hive 数据源中的基础数据，并检测到数据源的表结构，方便后续业务流程中作业的 SQL 处理，需要对应 Hive 数据源中的基础信息数据表在[DMP 数据管理/数据开发]的表管理中新建数据表，这些表为 ODS 层的表。

(1) 在[DMP 数据管理/数据开发]模块中，选择左侧导航树中的[表管理]菜单项，进入表管理页面。

(2) 单击左上角的<新建>按钮，进入新建表页面。

(3) 选择 Hive 数据源类型，并选择 5.2.2　3. 新增 Hive 数据源中创建的数据源。

(4) 配置表名等基本属性参数和物理模型设计参数。其中，表名根据实际情况配置，本例中为“ods_d_inter_person_inoculation_d”（人员接种信息）；物理模型设计的“外部表”参数需设为 状态，并指定 Hive 数据源中数据文件存放的 HDFS 路径。

图5-15 基本属性配置

图5-16 物理模型设计配置

(5) 单击<下一步>按钮，进入数据结构配置页面。

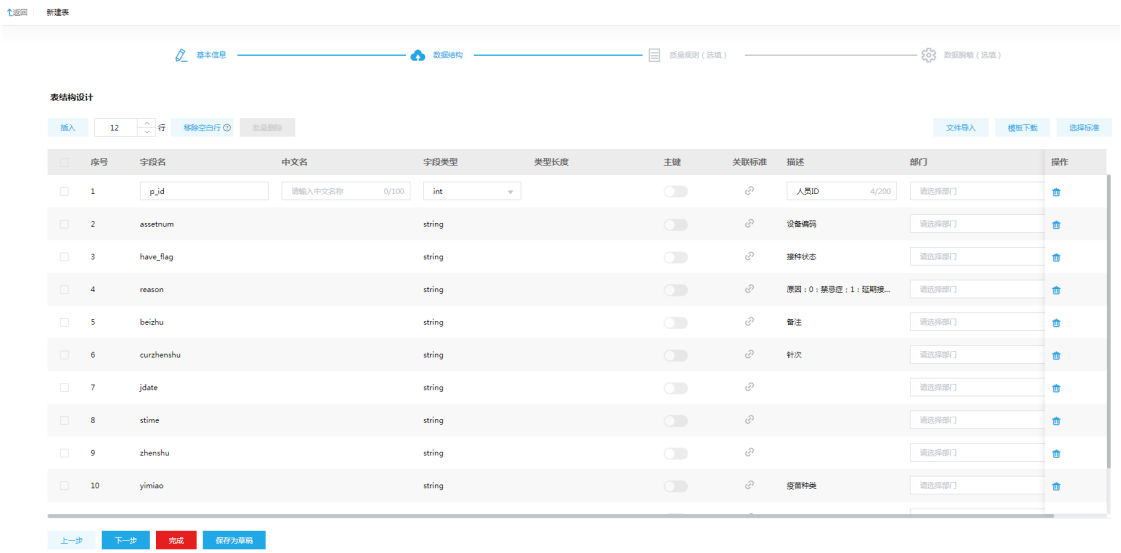
(6) 在数据结构配置页面中，可通过<模板下载>按钮下载模板，然后填入字段等信息，再通过<文件导入>按钮进行导入。人员接种信息表示例字段如表 5-1 所示，通过文件导入后如图 5-17 所示。

表5-1 人员接种信息表示例字段信息

字段名称	字段类型	描述
p_id	int	人员ID
assetsnum	string	设备编码
have_flag	string	接种状态
reason	string	原因：0：禁忌症；1：延期接种；2：需退回上级重新分配；3：不符合接种人群范围
beizhu	string	备注
curzhenshu	string	针次
jdate	string	-

字段名称	字段类型	描述
stime	string	-
zhenshu	string	-
yimiao	string	疫苗种类
jinjizheng	string	禁忌症
created	string	接种时间

图5-17 表结构设计



(7) 单击<完成>按钮，表新建完成。

(8) 重复步骤(2)-步骤(7)，依次新建人员信息表（字段如表 5-2 所示）、辖区（街道）字典表（字段如表 5-3 所示）、人员分类字典表（字段如表 5-4 所示）。

表5-2 人员信息表示例字段信息

字段名称	字段类型	描述
id	int	序号
name	string	姓名
sex	string	性别
age	int	年龄
mobile	string	手机号
cardno	string	身份证号
classfication_id	int	人员分类一级

字段名称	字段类型	描述
content	string	备注
company_id	int	单位ID
region_id	int	辖区ID对应属地的摸底工作部门
declare_department_id	int	申报部门ID
created_time	string	创建时间
modified_time	string	修改时间
uuid	string	UUID
addr	string	现住址
streetId	string	街道
provinceid	string	省ID
cityid	string	市ID
disctrictid	string	区域ID
flag	string	是否本市住户
area	string	小区名字
subclass_id	int	人群分类二级

表5-3 辖区（街道）字典表示例字段信息

字段名称	字段类型	描述
id	int	序号
region	string	辖区
streetId	string	街道
userid	string	User_id
category	string	有没有二级分类（1表示有）
sort	int	顺序编号

表5-4 人员分类字典表示例字段信息

字段名称	字段类型	描述
id	int	序号
classsfication	string	人员分类
created_time	string	创建时间
modified_time	string	修改时间
category	string	级别

2. 新建人员接种信息数据清洗表

在基础的人员接种信息表中，可能存在错误或不完整的数据，为保证后续的数据处理可以正常进行，需要对基础信息表中的人员接种信息表进行清洗处理。人员接种信息数据清洗表即用于存放清洗后的人员接种信息数据，需要在数据清洗操作执行前，先新建该表。该表结构与基础信息表中的人员接种信息表相同。该表为 DWB 层的表。

(1) 在[DMP 数据管理/数据开发]模块中，选择左侧导航树中的[表管理]菜单项，进入表管理页面。

(2) 在页面右上角选择组织，本例中选择“根组织”。

(3) 单击左上角的<新建>按钮，进入新建表页面。

(4) 选择 Hive 数据源类型，并选择 5.2.2　3. 新增 Hive 数据源中创建的的数据源。

(5) 配置表名等基本属性参数和物理模型设计参数。其中，表名根据实际情况配置，本例中为“dwb_filtered_person_inoculation_d”（人员接种信息数据清洗表）；物理模型设计的“外部表”参数需设为  状态，将该表设置为内部表，以便于管理和使用。

(6) 单击<下一步>按钮，进入数据结构配置页面。

(7) 在数据结构配置页面中，表的字段信息与人员接种信息表示例字段一致，如表 5-1 所示，可通过文件导入，导入后如图 5-17 所示。

(8) 单击<完成>按钮，表新建完成。

3. 新建结果表

为便于存储数据处理后的结果数据，需要先新建各结果数据表，这些表为 DWS 层的表。

(1) 在[DMP 数据管理/数据开发]模块中，选择左侧导航树中的[表管理]菜单项，进入表管理页面。

(2) 在页面右上角选择组织，本例中选择“根组织”。

(3) 单击左上角的<新建>按钮，进入新建表页面。

(4) 选择 Greenplum 数据源类型，并选择 5.2.2　4. 创建 Greenplum 数据源中创建的数据源。

(5) 配置表名等基本属性参数和物理模型设计参数。其中，表名根据实际情况配置，本例中为“dws_region_inoculation_day_statistics”（社区街道疫苗接种按天统计结果表）；物理模型设计部分，存储模式使用默认值“row”，模式选择 public。

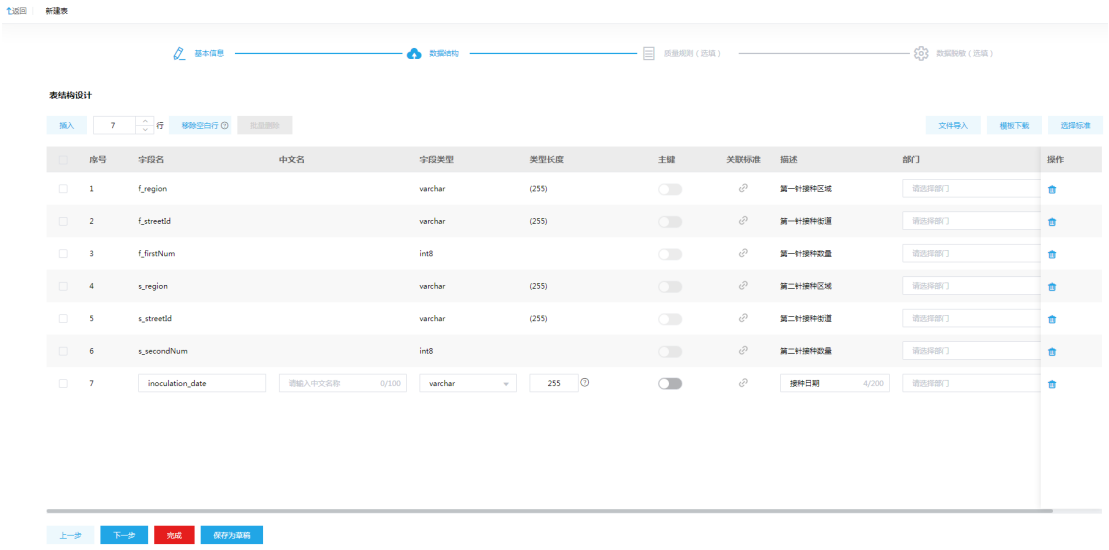
(6) 单击<下一步>按钮，进入数据结构配置页面。

(7) 在数据结构配置页面中，表的字段信息与人员接种信息表示例字段一致，如所示，可通过文件导入后如所示。

表5-5 社区街道疫苗接种按天统计结果表字段信息

字段名称	字段类型	描述
f_region	varchar(255)	第一针接种区域
f_streetId	varchar(255)	第一针接种街道
f_firstNum	int8	第一针接种数量
s_region	varchar(255)	第二针接种区域
s_streetId	varchar(255)	第二针接种街道
s_secondNum	int8	第二针接种数量
inoculation_date	varchar(255)	接种日期

图5-18 表结构设计



(8) 单击<完成>按钮，表新建完成。

(9) 重复步骤(2)-步骤(8)，依次新建社区街道疫苗接种全量统计结果表（字段信息如表 5-6 所示）、行业疫苗接种按天统计结果表（字段信息如表 5-7 所示）、行业疫苗接种全量统计结果表（字段信息如表 5-8 所示）、各年龄段疫苗接种按天统计结果表（字段信息如表 5-9 所示）、各年龄段疫苗接种全量统计结果表（字段信息如表 5-10 所示）、第一针接种至今各个时间间隔人数统计结果表（字段信息如表 5-11 所示）。

表5-6 社区街道疫苗接种全量统计结果表字段信息

字段名称	字段类型	描述
f_region	varchar(255)	第一针接种区域
f_streetId	varchar(255)	第一针接种街道
first_total_num	int8	第一针接种数量
s_region	varchar(255)	第二针接种区域
s_streetId	varchar(255)	第二针接种街道
second_total_num	int8	第二针接种数量

表5-7 行业疫苗接种按天统计结果表字段信息

字段名称	字段类型	描述
f_classification	varchar(255)	第一针行业分类
firstNum	int8	第一针接种数量
s_classification	varchar(255)	第二针行业分类
secondNum	int8	第二针接种数量
inoculation_date	varchar(255)	接种日期

表5-8 行业疫苗接种全量统计结果表字段信息

字段名称	字段类型	描述
f_classification	varchar(255)	第一针行业分类
firstNum	int8	第一针接种数量
s_classification	varchar(255)	第二针行业分类
secondNum	int8	第二针接种数量

表5-9 各年龄段疫苗接种按天统计结果表字段信息

字段名称	字段类型	描述
f_ageRange	varchar(255)	第一针年龄段
firstNum	int8	第一针接种数量
s_ageRange	varchar(255)	第二针年龄段

字段名称	字段类型	描述
secondNum	int8	第二针接种数量
inoculation_date	varchar(255)	接种日期

表5-10 各年龄段疫苗接种全量统计结果表字段信息

字段名称	字段类型	描述
f_ageRange	varchar(255)	第一针年龄段
first_total_num	int8	第一针接种数量
s_ageRange	varchar(255)	第二针年龄段
second_total_num	int8	第二针接种数量

表5-11 第一针接种至今各个时间间隔人数统计结果表字段信息

字段名称	字段类型	描述
interval_period	varchar(255)	时间间隔
person_num	int8	人数

5.3 构建业务流程

准备工作完成后，即可开始构建业务流程，包括创建业务流程，并在业务流程画布中增加数据清洗作业和各类数据计算作业。

5.3.1 创建业务流程

(1) 在[DMP 数据管理/数据开发]模块中，选择左侧导航树中的[调度中心]菜单项，进入调度中心页面。

(2) 在页面右上角选择组织，本例中选择“根组织”。

(3) 单击左上角的<新建>按钮，进入新建业务流程。

(4) 输入业务流程名称和描述信息，本例中名称为“疫苗接种数据统计”。

(5) 单击<确定>按钮，业务流程创建成功，页面进入该业务流程的画布编辑页签。

(6) 将左侧的作业组件拖入画布中，生成业务流程中的作业节点。双击该作业节点，可在弹窗中配置节点参数。

本例中需要增加人员接种数据清洗作业和各数据计算作业。

5.3.2 添加数据清洗作业

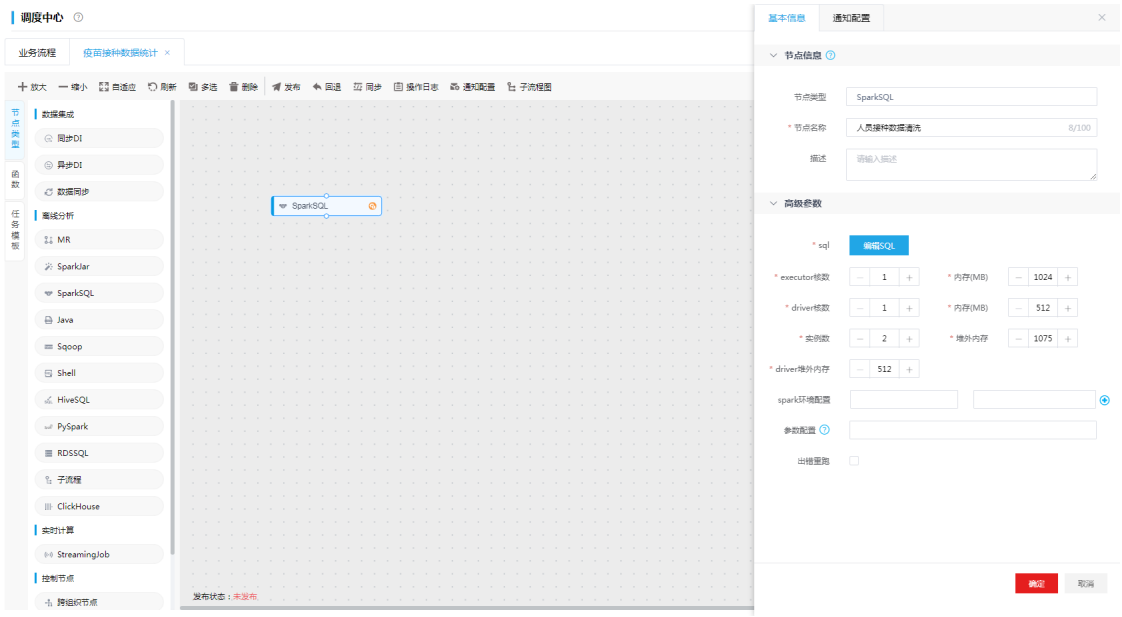
业务流程创建后，需要在业务流程的画布中增加人员接种数据清洗作业。

(1) 在业务流程的画布编辑页签中，选择左侧离线分析下的 SparkSQL 组件，并拖入画布中。

(2) 双击画布中的 SparkSQL 作业节点，弹出作业节点参数编辑窗口。

(3) 本例中，配置节点名称为“人员接种数据清洗”，选择执行队列为缺省队列。

图5-19 配置节点参数



(4) 单击<编辑 SQL>按钮，在弹出框中编写 SQL 语句，示例如下：

```
insert into
    default.dwb_filtered_person_inoculation_d
select
    *
from
    default.ods_d_inter_person_inoculation_d
where
    p_id is not null
    and haveflag is not null
    and curzhenshu is not null
    and created is not null;
```

图5-20 配置 SQL 语句

(5) 编写完成，并通过语法校验后，单击<确定>按钮，保存 SQL 语句。

(6) 单击<确定>按钮，数据清洗作业节点配置完成。

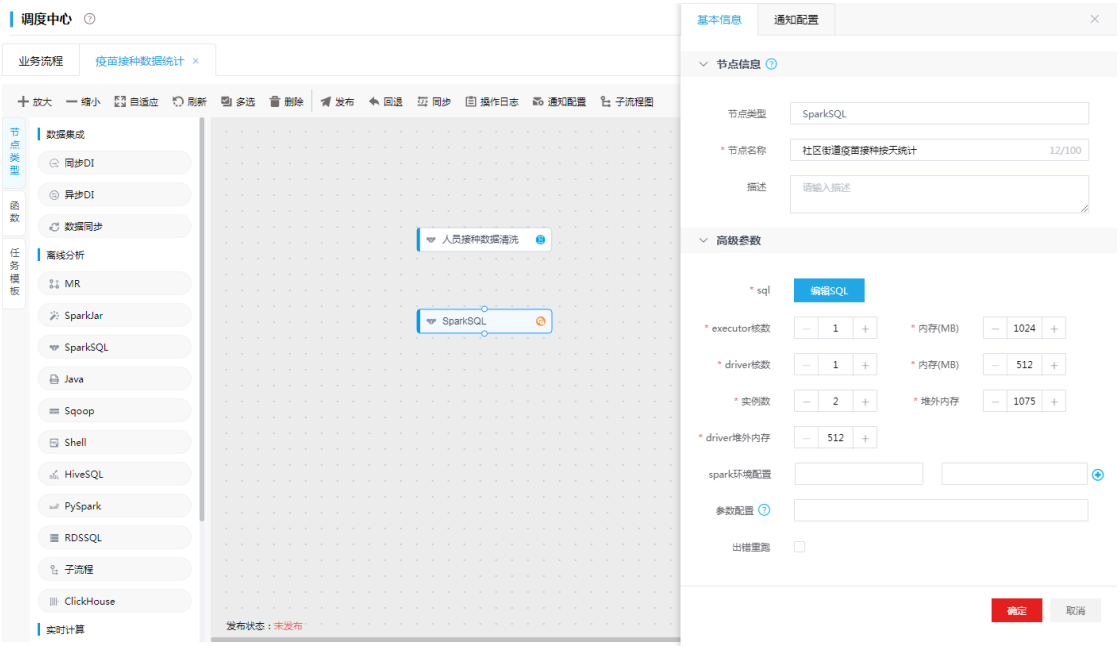
5.3.3 添加数据计算作业

(1) 清洗后的数据即可进行数据的计算处理，生成目标统计数据。在业务流程的画布编辑页签中，选择左侧离线分析下的 SparkSQL 组件，并拖入画布中。

(2) 双击画布中的 SparkSQL 作业节点，弹出作业节点参数编辑窗口。

(3) 本例中，配置节点名称为“社区街道疫苗接种按天统计”，选择执行队列为缺省队列。

图5-21 配置节点参数



(4) 单击<编辑 SQL>按钮，在弹出框中编写 SQL 语句，示例如下：

```
select
   f.region as f_region,
   f.streetId as f_streetId,
   f.firstNum as f_firstNum,
   s.region as s_region,
   s.streetId as s_streetId,
   s.secondNum as s_secondNum,
   f.f_inoculation_date as inoculation_date
from
   (
      select
         case
            when rd.region is not null then rd.region
            else '无区域'
         end as region,
         case
            when p.streetId is not null then p.streetId
            else '无街道'
         end as streetId,
         pi.f_inoculation_date,
         count(pi.p_id) as firstNum
      from
         (select *, substring_index(created, ' ', 1) as f_inoculation_date from
default.dwb_filtered_person_inoculation_d) pi
         left join ods_d_inter_person_d p on pi.p_id = p.id
         left join ods_d_inter_region_dict_d rd on rd.id = p.region_id
      where
         pi.haveflag = 'true'
         and curzhenshu = '0'
      group by
         rd.region,
         p.streetId,
         pi.f_inoculation_date
   ) f full
   join (
      select
         case
```

```
            when rd.region is not null then rd.region
            else '无区域'
        end as region,
        case
            when p.streetId is not null then p.streetId
            else '无街道'
        end as streetId,
        pi.s_inoculation_date,
        count(pi.p_id) as secondNum
    from
        (select *, substring_index(created, ' ', 1) as s_inoculation_date from
default.dwb_filtered_person_inoculation_d) pi
        left join ods_d_inter_person_d p on pi.p_id = p.id
        left join ods_d_inter_region_dict_d rd on rd.id = p.region_id
    where
        pi.haveflag = 'true'
        and curzhenshu = '1'
    group by
        rd.region,
        p.streetId,
        pi.s_inoculation_date
) s on f.region = s.region
and f.streetId = s.streetId and f.f_inoculation_date = s.s_inoculation_date;
```

(5) 编写完成，并通过语法校验后，单击<确定>按钮，保存 SQL 语句。

(6) 单击<确定>按钮，社区街道疫苗接种按天统计结果节点配置完成。

(7) 依次增加其他数据计算作业，各作业使用的 SQL 语句如所示。

表5-12 各数据计算作业使用的 SQL 语句

作业	SQL 语句
社区街道疫苗接种全量统计	select f.f_region as f_region, f.f_streetId as f_streetId, f.first_toal_num as first_total_num, s.s_region as s_region, s.s_streetId as s_streetId, s.second_toal_num as second_total_num from (select f_region, f_streetId, sum(f_firstNum) as first_toal_num from default.dws_region_inoculation_day_statistics group by f_region, f_streetId) as f

作业	SQL 语句
	full join (select s_region, s_streetid, sum(s_secondNum) as second_toal_num from default.dws_region_inoculation_day_statistics group by s_region, s_streetid) as s on f.f_region = s.s_region and f.f_streetId = s.s_streetId ;
行业疫苗接种按天统计	select f.classification as f_classification, f.firstNum as firstNum, s.classification as s_classification, s.secondNum as secondNum, f.f_inoculation_date as inoculation_date from(select case when pc.classification is not null then pc.classification else '未分类人员' end as classification, pi.f_inoculation_date as f_inoculation_date, count(pi.p_id) as firstNum from (select *, substring_index(created, ' ', 1) as f_inoculation_date from default.dwb_filtered_person_inoculation_d) pi left join ods_d_inter_person_d p on pi.p_id = p.id left join ods_d_inter_person_classification_d pc on pc.id = p.classification_id where pi.haveflag = 'true' and pi.curzhenshu = '0' group by pc.classification, pi.f_inoculation_date) f full join (select case

作业	SQL 语句
	when pc.classification is not null then pc.classification else '未分类人员' end as classification, pi.s_inoculation_date as s_inoculation_date, count(pi.p_id) as secondNum from (select *, substring_index(created, ' ', 1) as s_inoculation_date from default.dwb_filtered_person_inoculation_d) pi left join ods_d_inter_person_d p on pi.p_id = p.id left join ods_d_inter_person_classification_d pc on pc.id = p.classification_id where pi.haveflag = 'true' and pi.curzhenshu = '1' group by pc.classification, pi.s_inoculation_date) s on f.classification = s.classification and f.f_inoculation_date = s.s_inoculation_date;
行业疫苗接种全量统计	select f.f_classification as f_classification, f.first_toal_num as first_toal_num, s.s_classification as s_classification, s.second_toal_num as second_toal_num from (select f_classification, sum(firstNum) as first_toal_num from default.dws_classfication_inoculation_day_statistics group by f_classification) as f full

作业	SQL 语句
	join (select s_classification, sum(secondNum) as second_toal_num from default.dws_classfication_inoculation_day_statistics group by s_classification) as s on f.f_classification = s.s_classification;
各年龄段疫苗接种按天统计	select f.age_range as f_ageRange, f.firstNum as firstNum, s.age_range as s_ageRange, s.secondNum as secondNum, f.f_inoculation_date as inoculation_date from (select p.age_range as age_range, pi.f_inoculation_date as f_inoculation_date, count(pi.p_id) as firstNum from (select *, substring_index(created, ' ', 1) as f_inoculation_date from default.dwb_filtered_person_inoculation_d) pi left join (select case

作业	SQL 语句
	when age <= 18 then '18岁以下' when age <= 59 and age > 18 then '18岁至59岁' when age > 59 then '大于59岁' else '未找到年龄信息' end as age_range, id from ods_d_inter_person_d) p on pi.p_id = p.id where pi.haveflag = 'true' and pi.curzhenshu = '0' group by p.age_range, pi.f_inoculation_date) f full join (select p.age_range as age_range, pi.s_inoculation_date as s_inoculation_date, count(pi.p_id) as secondNum from (select *, substring_index(created, ' ', 1) as s_inoculation_date from default.dwb_filtered_person_inoculation_d) pi left join (select

作业	SQL 语句
	case when age <= 18 then '18岁以下' when age <= 59 and age > 18 then '18岁至59岁' when age > 59 then '大于59岁' else '未找到年龄信息' end as age_range, id from ods_d_inter_person_d) p on pi.p_id = p.id where pi.haveflag = 'true' and pi.curzhenshu = '1' group by p.age_range, pi.s_inoculation_date) s on s.age_range = f.age_range and f.f_inoculation_date = s.s_inoculation_date;
各年龄段疫苗接种全量统计	select f.f_ageRange as f_ageRange, f.first_total_num as first_total_num, s.s_ageRange as s_ageRange, s.second_total_num as second_total_num from (select f_ageRange, sum(firstNum) as first_total_num from default.dws_age_range_inoculation_day_statistics group by f_ageRange) as f full

作业	SQL 语句
	join (select s_ageRange, sum(secondNum) as second_total_num from default.dws_age_range_inoculation_day_statistics group by s_ageRange) as s on f.f_ageRange = s.s_ageRange;
第一针接种至今各个时间间隔人数统计	select case when datediff(date_format(CURRENT_DATE, 'yyyy-MM-dd'), replace(substring_index(created, ' ', 1), '/', '-')) <= 21 then '三周以内' when datediff(date_format(CURRENT_DATE, 'yyyy-MM-dd'), replace(substring_index(created, ' ', 1), '/', '-')) > 21 and datediff(date_format(CURRENT_DATE, 'yyyy-MM-dd'), replace(substring_index(created, ' ', 1), '/', '-')) <= 56 then '三周至八周' else '超过八周' end as interval_period, count(p_id) as person_num from default.dwb_filtered_person_inoculation_d where curzhenshu = 0 and haveflag = 'true' group by interval_period;

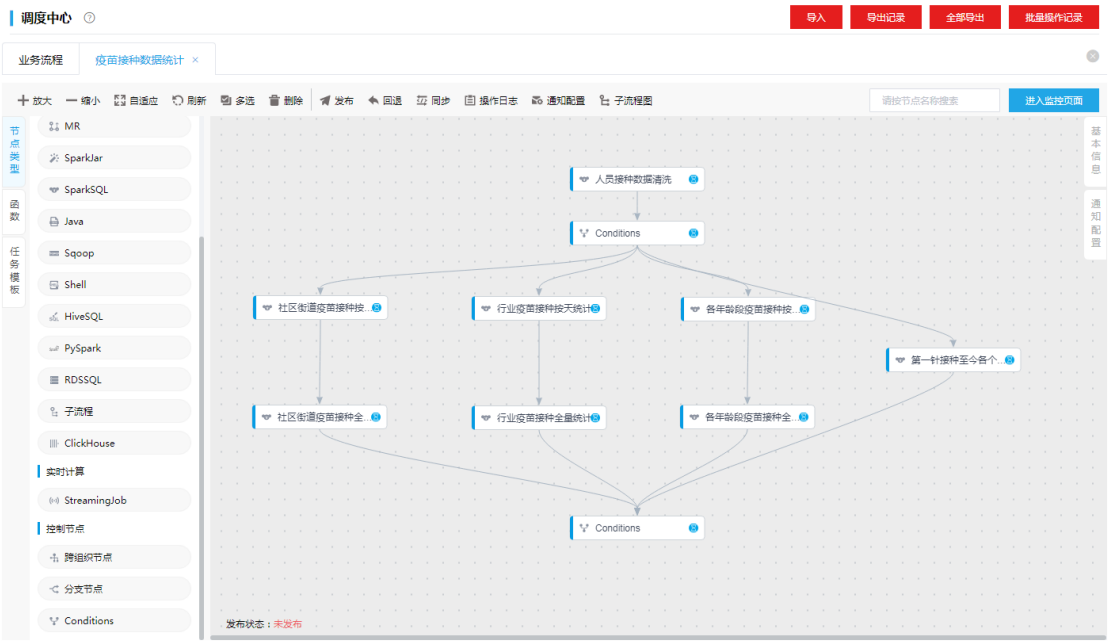
5.3.4 构建完成作业并运行

各作业创建完成后，需要通过控制节点下的 Conditions 组件进行连接，构建完整的业务流程。

(1) 在业务流程的画布编辑页签中，选择左侧控制节点下的 Conditions 组件，并拖入画布中。

(2) 依次连接各作业，连接结果如图 5-22 所示。

图5-22 关联作业



(3) 连接完成后，即可单击画布上方的<发布>按钮，发布业务流程为实例。

(4) 在左侧导航树中选择[运维管理/调度运维]菜单项，进入调度运维页面。

(5) 在业务流程实例列表中，单击业务流程对应操作列的<提交>按钮。针对业务流程实例，可配置调度策略，使其定期自动运行。

5.4 数据查询

当业务流程运行完成后，统计结果数据会存入预先创建的统计结果表中。

DMP 数据管理中提供了数据查询功能，可查询统计结果数据。

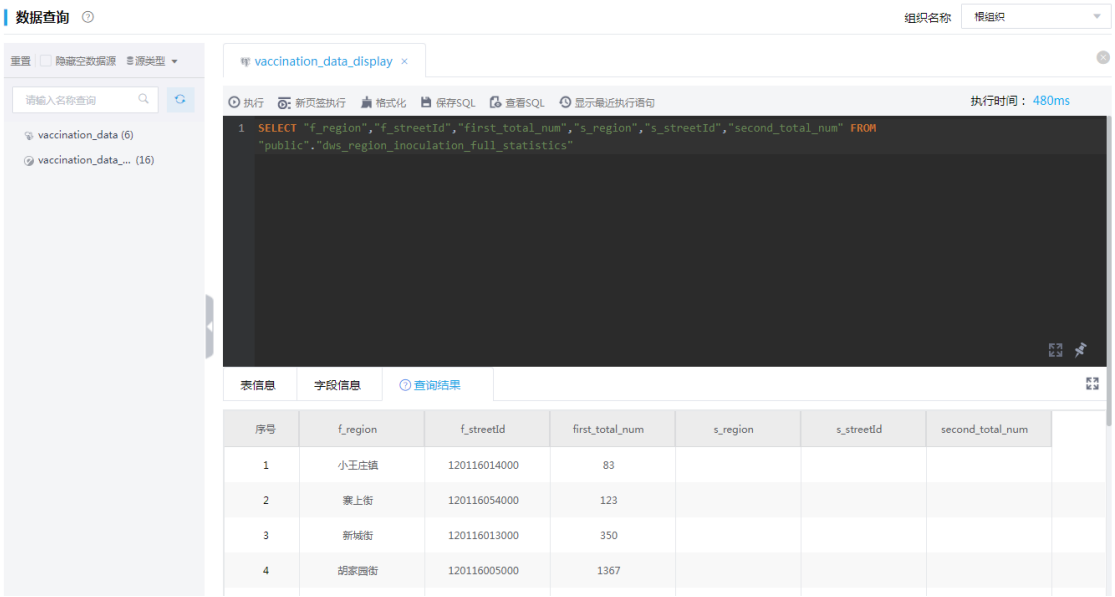
(1) 在[DMP 数据管理/数据探查]模块中的数据查询页面中在左侧目录中选择数据源，右侧出现该数据源的数据查询页签。

(2) 在输入框中输入 SQL 查询语句，查看各统计结果表中的数据。示例语句如下：

```
SELECT "f_region","f_streetId","first_total_num","s_region","s_streetId","second_total_num"
FROM "public"."dws_region_inoculation_full_statistics"
```

(3) 单击<执行>按钮，执行该查询语句，下方的查询结果页签中，会展示该统计结果表中的数据。

图5-23 查询结果



5.5 结果数据发布

通过数据计算得出的统计数据存入了统计结果表中，数字平台支持以表为单位，在 iPaaS 集成平台的服务集成功能中，将 Greenplum 数据源中的统计结果表发布，并授权给特定的工作空间，以便于第三方应用通过 URL 获取数据。

1. API 注册

在[iPaaS 集成平台/服务集成]模块下的[API 工厂/API 管理]页面，单击<API 注册>按钮，选择注册“数据 API”类型。

图5-24 API 注册

选择 5.2.4　3. 新建结果表中的结果表（注意：注册的一个 API 只能发布一张表）作为发布数据对象。配置数据 API 基本属性、选择需要发布的数据表。

图5-25 数据 API 设计

注册完成后，单击新生成的 API 右侧的<测试>按钮，对接口进行测试，如下图所示，测试接口是否可用。

图5-26 API 测试

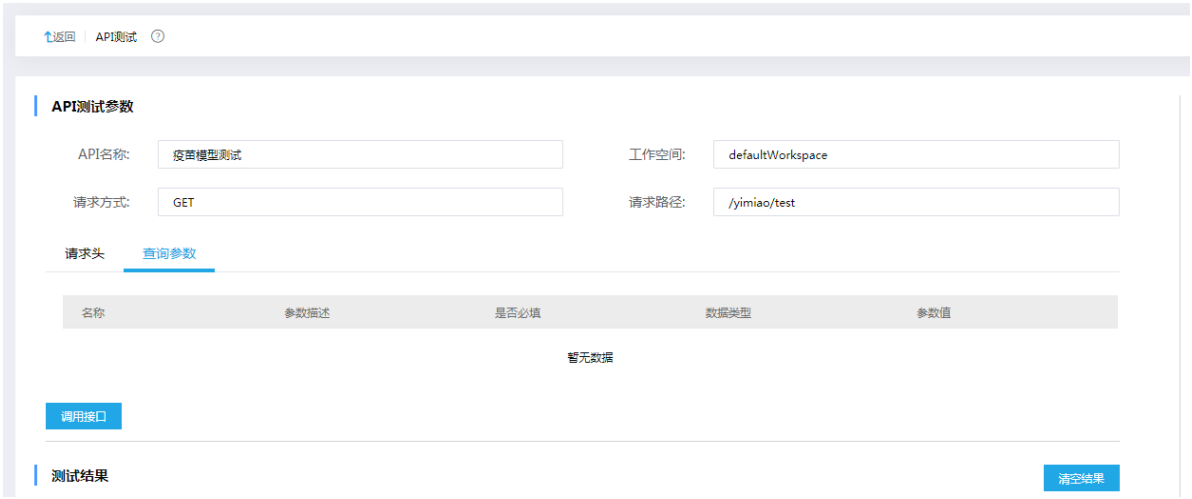
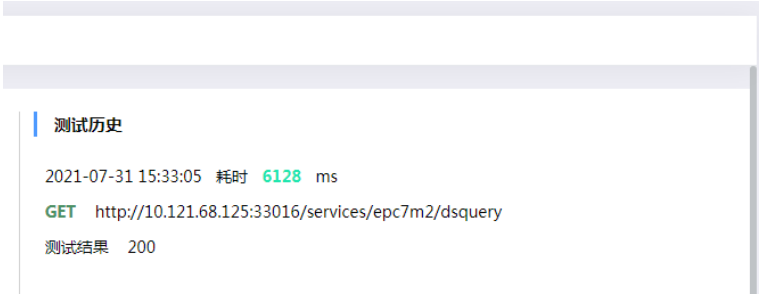


图5-27 测试结果



完成测试后，当前 API 状态即为待部署状态。单击右侧的<部署>按钮，即可在弹窗页面中配置部署节点。部署完成后，即可进行 API 授权操作。

2. API 授权

服务集成模块下的[API 网关/API 列表]页面，在列表中选择上一步注册的 API，然后单击右侧的<授权>按钮，进入 API 授权页面，配置需要授权的工作空间。授权完毕以后，在 API 授权页面下方会出现已授权的工作空间，单击操作列中的<测试>按钮。

图5-28 已授权工作空间

单击<测试>按钮后，进入 API 测试页面，API 测试页面完整显示了包括 IP 地址、端口号在内的完整访问路径，第三方应用只需访问此 URL 即可获取数据，无需在请求中携带任何用户信息。

图5-29 API 测试页面

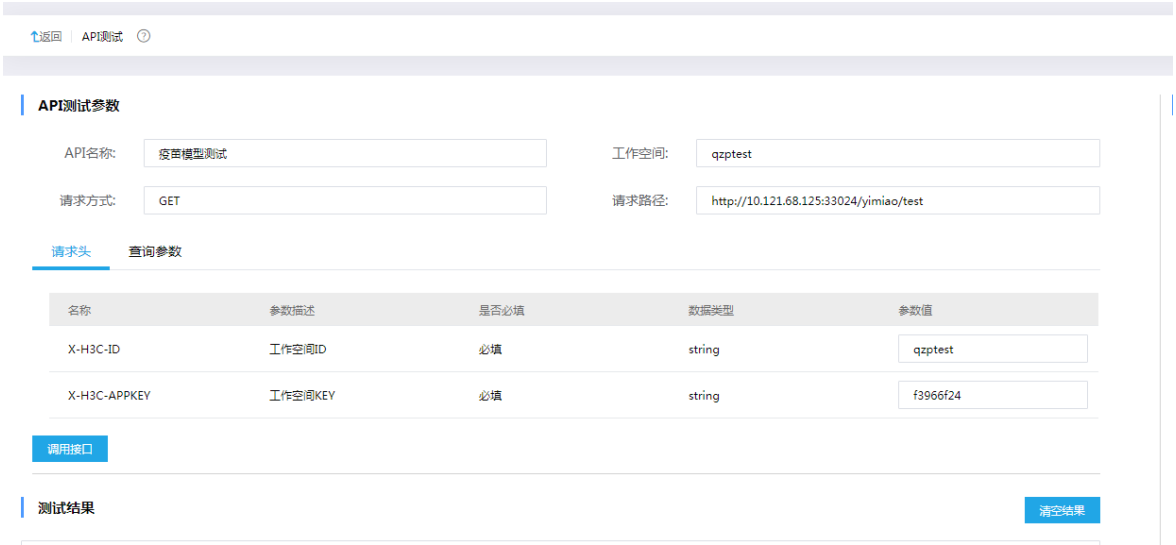


图5-30 测试结果

5.6 数据最终呈现

该案例中的统计数据发布后，支持通过第三方调用展示，如图 5-31 和图 5-32 所示。

图5-31 疫苗接种情况展示（一）

图5-32 疫苗接种情况展示（二）

6 常见问题解答

1. 通过 DI 抽取到 HBase 表的数据支持在数据查询中访问吗

由于 DI 是一套独立完整的数据抽取工具，不和具体的业务强相关，HBase 作为 NoSQL 数据库，在表结构设计、查询等方面比较灵活，因此需要根据具体的业务进行设计。

数据查询组件针对 HBase 数据的访问有一套自己的业务逻辑，因此不支持数据查询对 DI 抽取到 HBase 表数据的访问。

如果存在此需求，可以先将数据通过 DI 抽取到管道，然后通过管道将数据写入到 HBase 中，管道针对 HBase 的表结构设计，索引、以及入数据的过程是和数据查询的业务逻辑保持一致的。

2. 将数据通过 DI 抽取到管道后，新建数据同步任务运行报错以及数据采样存在数据缺失

若存在需求通过 DI 将数据抽取至管道中，针对 CSV 及 JSON 类型管道，请确保抽取过程中，数据格式满足目标管道预设数据结构。

7 附录

7.1 数据同步作业字段映射规则

表7-1 字段映射规则

数据源类型	字段类型	Kafka 字段类型
ClickHouse	int64	bigint
	uint64	
	decimal	double
	float64	
	float32	float
	int8	integer
	int16	
	int32	
	uint8	
	uint16	
	uint32	
	date	string
	datetime	
	datetime64	
	string	
DLH	bigint	bigint
	boolean	boolean
	decimal	double
	double	
	float	float
	int	integer
	smallint	

数据源类型	字段类型	Kafka 字段类型
	tinyint	
	char	string
	string	
	varchar	
	timestamp	timestamp
DRDS	bigint	bigint
	bit	boolean
	decimal	double
	int	integer
	mediumint	
	smallint	
	serial	
	tinyint	
	binary	string
	blob	
	char	
	date	
	datetime	
	longblob	
	longtext	
	mediumblob	
	mediumtext	
	text	
	time	
	tinyblob	
	tinytext	

数据源类型	字段类型	Kafka 字段类型
	varbinary	
	varchar	
	timestamp	timestamp
Elasticsearch	boolean	boolean
	double	double
	integer	integer
	long	long/timestamp
	array_date	string
	array_double	
	array_integer	
	array_keyword	
	array_long	
	array_text	
	attachment	
	geo_point	
	ip	
	keyword	
	text	
	date	timestamp
Greenplum	bit	boolean/string
	bool	boolean
	float8	double
	numeric	
	float4	float
	int2	integer
	int4	

数据源类型	字段类型	Kafka 字段类型
	serial	
	int8	long
	serial8	
	bpchar	string
	bytea	
	char	
	date	
	interval	
	text	
	time	
	varchar	
	uuid	
	timestamp	timestamp
HBase	double	double
	integer	integer
	long	long/timestamp
	array	string
	string	
Hive	bigint	bigint
	boolean	boolean
	decimal	double
	double	
	float	float
	int	integer
	smallint	
	tinyint	
	array<string>	string

数据源类型	字段类型	Kafka 字段类型
	binary	
	char	
	date	
	map<string,string>	
	string	
	varchar	
	timestamp	timestamp
Kafka	boolean	boolean
	bigint	bigint
	double	double
	float	float
	integer	integer
	long	long/timestamp
	object	object
	string	string
	timestamp	timestamp
	array[bigint]	array[bigint]
	array[double]	array[double]
	array[boolean]	array[boolean]
	array[float]	array[float]
	array[integer]	array[integer]
	array[long]	array[long]
	array[object]	array[object]
	array[string]	array[string]
	array[timestamp]	array[timestamp]
MySQL	bigint	bigint
	bit	boolean

数据源类型	字段类型	Kafka 字段类型
	decimal	double
	double	
	float	float
	int	integer
	mediumint	
	serial	
	smallint	
	tinyint	
	binary	string
	blob	
	char	
	date	
	datetime	
	longblob	
	longtext	
	mediumblob	
	mediumtext	
	text	
	time	
	tinyblob	
	tinytext	
	varbinary	
	varchar	
	timestamp	timestamp
PostgreSQL	bit	boolean/string
	bool	boolean
	float8	double

数据源类型	字段类型	Kafka 字段类型
	numeric	
	float4	float
	int2	integer
	int4	
	serial	
	int8	long
	serial8	
	bpchar	string
	bytea	
	char	
	date	
	interval	
	text	
	time	
	varchar	
	uuid	
	timestamp	timestamp
STDB	boolean	boolean
	double	double
	float	float
	integer	integer
	long	long/timestamp
	bytes	string
	date	
	geometry	
	geometrycollection	
	linestring	

数据源类型	字段类型	Kafka 字段类型
	list[a]	
	map[a,b]	
	multipoint	
	multilinestring	
	multipolygon	
	point	
	polygon	
	string	
	uuid	
	timestamp	timestamp
Vertica	boolean	boolean
	numeric	double
	float	float
	integer	integer
	binary	string
	char	
	date	
	geography	
	geometry	
	long varbinary	
	long varchar	
	time	
	uuid	
	varchar	
	varbinary	
	timestamp	timestamp
达梦	bigint	bigint

数据源类型	字段类型	Kafka 字段类型
	bit	boolean
	datetime	double
	decimal	
	double	
	double precision	
	number	
	numeric	
	float	float
	int	integer
	integer	
	smallint	
	tinyint	
	bfile	string
	binary	
	blob	
	byte	
	char	
	character	
	clob	
	date	
	image	
	text	
	time	
	varbinary	
	varchar	
	varchar2	
	timestamp	timestamp

7.2 疫苗接种案例业务数据库建表语句示例

在业务数据库中，可通过 SQL 语句创建记录原始数据的表，本节提供了参考示例。

1. 创建人员信息表的 SQL 语句示例

```
CREATE TABLE 'person' (
  'id' int(11) NOT NULL AUTO_INCREMENT COMMENT '序号',
  'name' varchar(255) DEFAULT NULL COMMENT '姓名',
  'sex' varchar(255) DEFAULT NULL COMMENT '性别',
  'age' int(11) DEFAULT NULL COMMENT '年龄',
  'mobile' varchar(255) DEFAULT NULL COMMENT '手机号',
  'cardno' varchar(20) DEFAULT NULL COMMENT '身份证号',
  'classification_id' int(11) DEFAULT NULL COMMENT '人员分类(一级)',
  'content' varchar(255) DEFAULT NULL COMMENT '备注',
  'company_id' int(11) DEFAULT NULL COMMENT '单位名称',
  'region_id' int(11) DEFAULT NULL COMMENT '辖区 id，对应属地的摸底工作部门',
  'declare_department_id' int(11) DEFAULT NULL COMMENT '申报部门 id',
  'created_time' datetime DEFAULT NULL COMMENT '创建时间',
  'modified_time' varchar(255) DEFAULT NULL COMMENT '修改时间',
  'uuid' varchar(50) DEFAULT NULL COMMENT 'UUID',
  'addr' varchar(500) DEFAULT NULL COMMENT '现住址',
  'streetId' varchar(20) DEFAULT NULL COMMENT '街道',
  'provinceId' varchar(20) DEFAULT NULL COMMENT '省 ID',
  'cityId' varchar(20) DEFAULT NULL COMMENT '市 ID',
  'districtId' varchar(20) DEFAULT NULL COMMENT '区域 ID',
  'flag' varchar(10) DEFAULT NULL COMMENT '是否本市住户',
  'area' varchar(255) DEFAULT NULL COMMENT '小区名字',
  'subclass_id' int(11) DEFAULT NULL COMMENT '人群分类（二级)',
  PRIMARY KEY ('id') USING BTREE,
  UNIQUE KEY 'class_index' ('id','classification_id') USING BTREE,
  KEY 'compant_index' ('company_id') USING BTREE,
  KEY 'region_id_index' ('region_id') USING BTREE
) ENGINE=InnoDB AUTO_INCREMENT=1649514 DEFAULT CHARSET=utf8 ROW_FORMAT=DYNAMIC
COMMENT='人员信息'
```

2. 创建人员分类字典表的 SQL 语句示例

```
CREATE TABLE 'person_classification' (
```

```
  'id' int(11) NOT NULL COMMENT 'id',
  'classification' text COMMENT '人员分类',
  'created_time' datetime DEFAULT NULL COMMENT '创建时间',
  'modified_time' datetime DEFAULT NULL COMMENT '修改时间',
  'category' varchar(255) DEFAULT NULL COMMENT '级别',
  PRIMARY KEY ('id') USING BTREE
) ENGINE=InnoDB DEFAULT CHARSET=utf8 ROW_FORMAT=DYNAMIC COMMENT='人员分类字典表'
```

3. 创建辖区字典表的 SQL 语句示例

```
CREATE TABLE 'region_dict' (
  'id' int(11) NOT NULL AUTO_INCREMENT COMMENT '序号',
  'region' varchar(255) DEFAULT NULL COMMENT '辖区',
  'userid' varchar(50) DEFAULT NULL COMMENT 'userid',
  'category' varchar(255) DEFAULT NULL COMMENT '有没有二级分类(1 是有)',
  'sort' int(11) DEFAULT NULL COMMENT '顺序编号',
  PRIMARY KEY ('id') USING BTREE,
  UNIQUE KEY 'region' ('region') USING HASH COMMENT 'regoion 索引'
) ENGINE=InnoDB AUTO_INCREMENT=27 DEFAULT CHARSET=utf8 ROW_FORMAT=DYNAMIC COMMENT='
辖区字典表'
```

4. 创建人员接种信息表的 SQL 语句示例

```
CREATE TABLE 'person_inoculation' (
  'p_id' int(11) NOT NULL COMMENT 'ID',
  'assetsNum' varchar(100) DEFAULT NULL COMMENT '设备编码',
  'haveflag' varchar(10) DEFAULT NULL COMMENT '接种状态  true/false',
  'reason' varchar(800) DEFAULT NULL COMMENT '状态：0:禁忌症,1:延期接种,2:需退回上级重新分
配,3:不符合接种人群范围',
  'beizhu' text COMMENT '备注',
  'curzhenshu' varchar(20) DEFAULT NULL COMMENT '针次',
  'jdate' varchar(20) DEFAULT NULL,
  'stime' varchar(20) DEFAULT NULL,
  'zhenshu' varchar(20) DEFAULT NULL,
  'yimiao' varchar(10) DEFAULT NULL COMMENT '疫苗种类  0:北京生物,1:北京科兴,2:武汉生物,3:康希
诺',
  'jinjizheng' varchar(255) DEFAULT NULL COMMENT '禁忌症',
  'created' datetime DEFAULT NULL COMMENT '接种时间',
  KEY 'class_index' ('p_id') USING BTREE
) ENGINE=InnoDB DEFAULT CHARSET=utf8 ROW_FORMAT=DYNAMIC
```