云海佳信科技
MacroData Technology

MacroData Data Integration Platform

云海睿治数据集成平台

白皮书

北京云海佳信科技有限公司

目录

背景及定位

背景

随着企业信息化建设的发展，企业内部积累了大量的业务数据。而企业的业务数据通常分布在相互独立的各业务系统，各业务系统的数据信息都能很好的满足自身业务需求，但是随着信息化的发展，单个业务系统的独立运行已经不满足当前信息化的发展需求，数据的高度整合，深度应用成为了一个重要的发展方向。而在数据整合和应用过程中，数据源种类繁多、接口复杂、数据难互通；数据质量不过关，存在大量冗余、数据不一致等数据质量问题。如何快速高效进行数据整合、互联互通，稳定提高数据质量，是数据集成、治理过程的重中之重。

产品定位

睿治数据集成平台（简称云海睿智）是云海佳信公司在多年行业研究及应用实践基础上，参考国际先进产品设计、结合国内数据特点研发的一款分布式部署、可动态扩展、集数据采集和数据处理于一体的数据集成产品。睿治数据集成平台可从分布式、多样化异构数据源（如关系数据库、数据文件、Rest 接口）进行数据抽取，并按照数据处理需求进行清洗、转换、融合，最后加载到数据仓库或数据集市中，实现跨部门、跨系统的数据集成，为基于数据仓库的数据应用、分析决策提供高质量的数据支撑。

产品概览

系统架构

产品由管理中心和节点群两部分组成，管理中心以图形化方式完成数据集成任务的配置管理、监控运维、统计分析、以及系统管理等功能；节点群提供可分布式部署、动态扩展的数据采集处理引擎，完成数据采集、数据处理、数据加载。

管理中心

数据集成管理中心提供图形化用户接口，提供集设计、开发、发布、执行、监控、统计分析的一体化管理平台。可图形化方式完成数据集成需求到实现的快速转换，并对数据任务进行部署和可视化监控、统计分析，实现数据任务的全生命周期管理。

节点群

节点群是数据集成的执行引擎，负责执行数据集成任务，完成数据集成逻辑，根据负载情况可横向扩展、弹性伸缩；

主要功能介绍

平台概览

提供平台管理视角的统计概览，以便用户直观了解平台规模，数据处理数据量、趋势，以及这些数据都是由哪些机构提供。以及多维统计分析（数据任务按照数据源、数据来源机构统计分析）等。

节点群管理

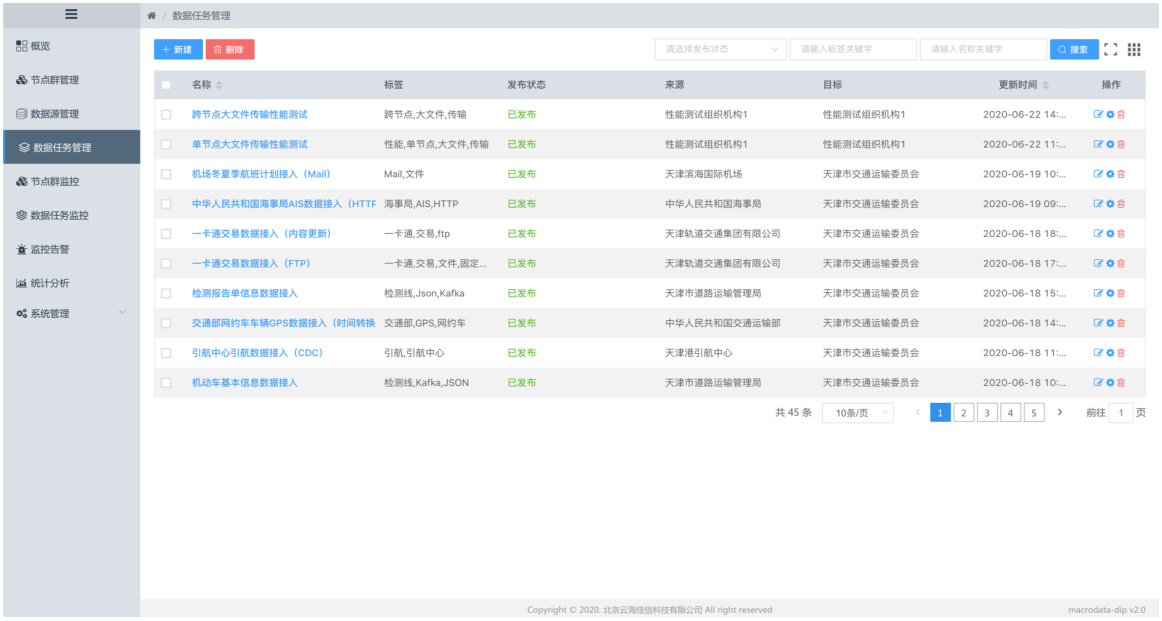
用于对所管理的节点群进行添加、修改、删除等管理。

数据源管理

对数据集成处理过程中用到的数据源，进行添加、修改、删除管理。

数据任务管理

可根据集成需求建立、修改、删除、发布数据任务，以及数据任务内的流程进行管理配置，启动执行。



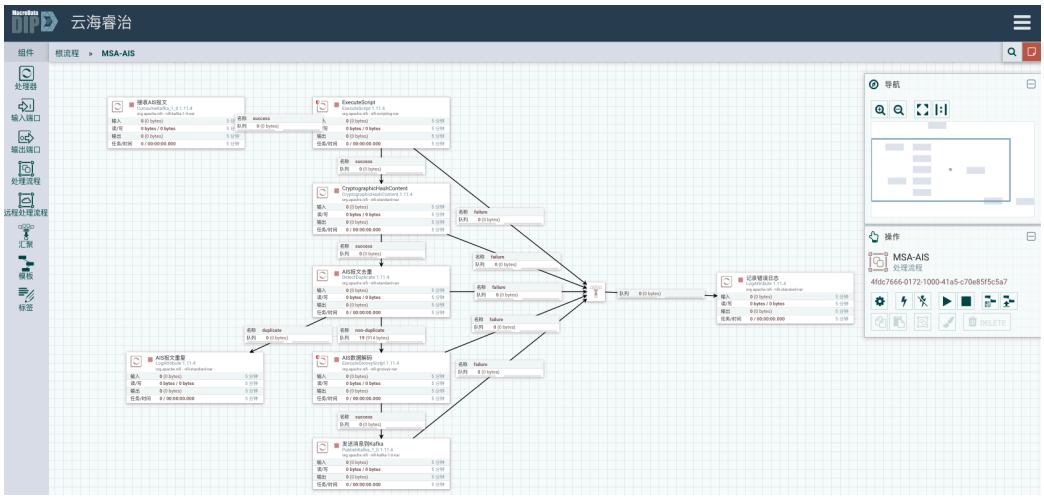
流程配置

依据数据集成计划，可视化、拖拽式进行数据流程配置管理。

流程配置总览

流程配置

提供丰富的流程组件，并以有向拓扑图形式进行数据流程配置。



节点群监控

监控节点群、及运行在节点上的流程的状态，当节点或流程有运行异常时，系统自动产生告警，并在监控界面中标注告警状态，提醒用户进行关注处理。

用户可查看告警信息，进行进一步分析告警原因，并进行处理。

数据任务监控

数据任务监控模块以列表形式展示各个数据任务该任务下数据流程的运行状态，数据任务下的流程有异常告警时，对应的数据流程和数据任务以红色告警状态提醒用户，以便用户及时关注并处理。

用户也可通过本功能查看当前数据任务或流程的累计和当日数据处理量和趋势。

监控告警

本功能可查看平台内的全局告警事件，并可按照告警时间、告警级别、关键字等对告警事件进行检索过滤等，快速发现问题并分析出处理。

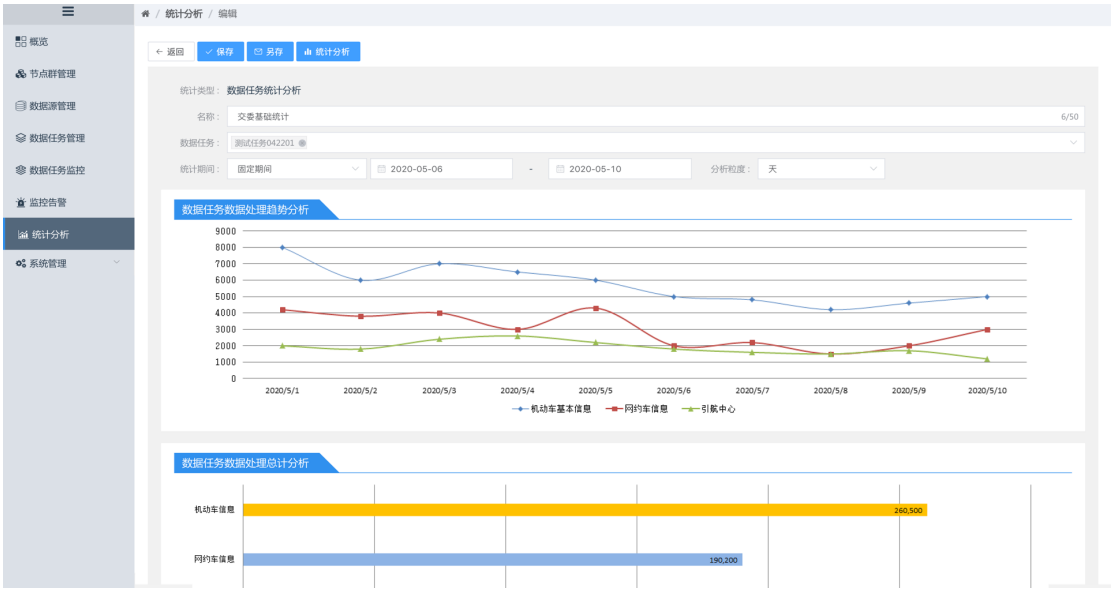
统计分析

通过本功能，用户可自定义三类统计分析：

数据任务统计分析：一定时间段内、一定分析粒度，统计分析一个或多个数据任务的数据处理总量和趋势，多个数据任务的趋势可进行叠加分析。

组织机构统计分析：一定时间段内、一定分析粒度，统计分析两个组织机构间的数据往来总量和趋势。

数据源统计分析：一定时间段内、一定分析粒度，统计分析两个数据源间的数据往来总量和趋势。



系统管理

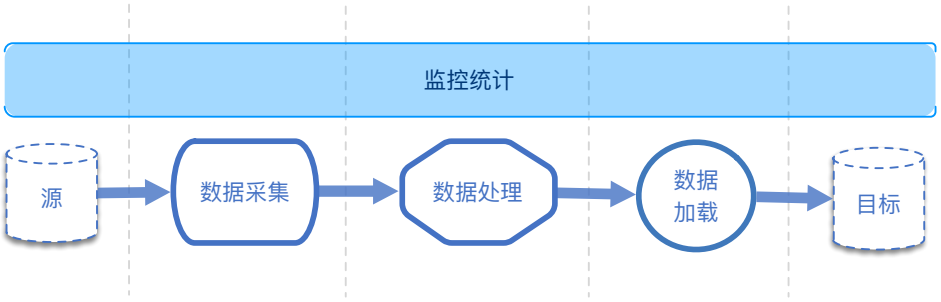
通过本功能可进行系统级的管理操作，如：机构管理、用户管理、角色管理、安全审计等。

产品特点

适应多平台

数据集成平台基于 Java 技术开发，受益于 Java 的跨平台特性，数据集成平台支持跨平台，可部署运行在 Windows、Linux、MacOS、AIX 等操作系统中。

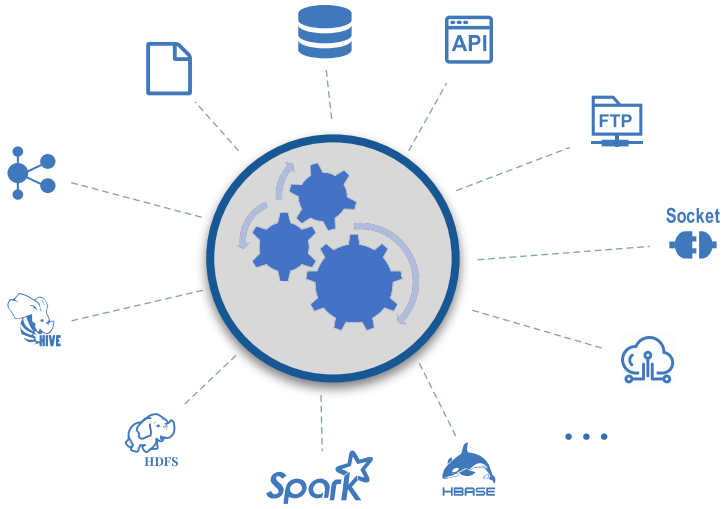
数据集成全生命周期管理



睿治数据集成产品提供数据采集、处理、加载、监控统计全周期一体化管理平台。

广泛的数据协议适配

数据集成平台广泛适配多种数据协议，支持机构化、半结构化和非机构化数据集成：



- 数据库：适配主流的关系型数据库的数据采集和加载，如 Oracle、MS SQL Server、Sybase、IBM DB/2、InfoMix 等商用数据库，MySQL、PostgreSQL 等开源数据库，神通、达梦、Kingbase、GBase 8t 等国产数据库；
- 数据文件：CSV，JSON、XML、AVRO、EXCEL 等；
- 通讯协议 :JMS、KAFKA、MQTT、HTTP(S)、WebSocket、FTP、SFTP、MAIL、SNMP、AMQP、UDP、TCP 等；
- 大数据：HDFS、Hive、HBase、Spark、MongoDB、InfluxDB、Elasticsearch、Cassandra 等；
- 云平台：Amazon、Azure、Google、阿里云。

丰富的数据处理组件

数据集成产品提供丰富的数据处理组件供用户组合使用来配置数据处理流程，组件类型包括：数据转换、加密压缩、聚合拆分、路由分发、数据去重、数据合并、数据验证、数据计算、脚本编辑……

多样化数据采集模式

支持全量采集以及增量数据采集，其中数据库增量采集支持触发器、时间戳、标志位方式、CDC 等；适应数据工程中不同阶段对数据采集处理的需求。

灵活的调度管理

数据处理流程支持定时、实时、手工触发、API 触发、条件触发等多种调度触发模式。

高可用、动态扩容

数据集成平台采用分布式架构设计，支持集群化部署、自动故障转移（当某个节点出现故障时，部署在该节点的数据流程可自动转移到其他集群节点继续运行），实现数据集成任务流程高可用；

同时根据实际负载情况，在不中断数据集成流程的情况下实现对集群节点进行动态扩展，弹性伸缩，有效避免单节点资源瓶颈，提高数据集成处理效率。

并发处理、负载均衡

数据集成流程中的数据处理组件可依据并发配置，进行并发数据处理，对应大量数据处理可通过提高并发数来提升数据处理性能。

在集群节点中部署数据集成流程时，可选择数据集成流程或某些组件运行在主引擎节点，也可选择自动部署。当选择自动部署时，系统会自动选择在负载较低的引擎节点上运行，运行过程中，可在集群节点间进行负载切换，完成负载均衡；

数据处理优先级策略

数据集成平台提供多种数据优先级策略：

- 先进先出
- 旧数据优先
- 新数据优先
- 属性权重优先

允许设置一种或多种优先级分配方案，来确定流程组件间传递数据的优先顺序。默认采用先进先出策略，也可根据实际需求选择其他策略，如实时位置数据采集场景，可采用后进先出策略，以保最新数据优先处理。

数据可靠保证

数据处理过程中，每个处理步骤都进行持久性日志预写，并通过内置内容仓库临时存储处理数据，保证即使突然宕机数据也不会丢失。

数据缓冲和背压

数据集成平台支持所有数据处理步骤的数据缓冲机制，并当缓冲区达到一定限制（数据包数或数据量大小）时，提供背压能力；有效提高数据处理效率的同时，对数据流量消峰处理，避免数据溢出。

全流程数据来源追溯

数据在数据集成平台中流转时，如扇入、扇出、转换过程中，系统会自动记录并索引数据来源 ,通过管理中心以可视化方式查看数据包流转过程、血缘关系，并可对数据进行重放处理，从而支持对流程进行持合规性检查、故障诊断、流程优化。

支持流程模板，简化配置

用户通过管理中心的流程编辑器可通过拖拽配置方式构建非常复杂的数据处理流程。但经常会遇到场景类似的多个若干流程，重复相同的配置过程枯燥而乏味。

为解决这个问题，系统提供流程模板概念，允许用户将一个已有流程保存为模板，新建流程时可选择已有模板来创建流程，只需要简单修改必须参数即可完成一个流程配置，简单而高效。

详细日志审计

数据集成平台，提供用户操作和数据处理两种审计，可有效帮助用户进行操作追溯和数据追溯。

高扩展架构

数据集成平台采用插件式高扩展架构，提供组件开发扩展点，当预置组件无法满足特定数据集成场景需求时，可通过扩展点进行插件式二次开发，扩展用户自定义组件，自定义组件在使用和管理监控上与内置组件相同。

接口开放、易集成

数据集成产品提供配置、监控、统计的 Restful 接口，可与第三方平台便捷的深度集成。