



云海佳信科技
MacroData Technology

MacroData Data Middle Platform

云睿数据中台

白皮书

北京云海佳信科技有限公司

版权声明 © 2020 北京云海佳信科技有限公司 保留一切权利

任何单位或个人未经云海佳信书面许可，不得擅自摘抄、复制本文件中的内容，不得以任何形式传播。

商标声明

本文件展示、提及或使用的所有商标归云海佳信或者其他商标持有人所有。本文件内容不视为以明示、暗示、默许或者其他形式授予任何单位或个人商标使用权。未经云海佳信书面许可，任何单位或个人不得以任何形式使用云海佳信的商标或标记。

安全港声明

您购买的产品、服务或功能等受您与云海佳信所签订的商业合同约束，本文件所描述的产品、服务或功能可能不在您购买或使用范围之内。由于产品版本升级或其他原因，本文件内容会不定期进行更新，对此不会另行通知。除非另有约定，本文件仅作参考作用，所有陈述不构成对合同相对方的任何保、承诺，不视为合同的组成部分或者附件，云海佳信对此保留最终解释权。

目录

背景及定位	3
背景	3
产品定位	3
功能介绍	5
系统架构	5
数据采集	5
多源异构数据采集加载	6
多样化数据采集模式	7
采集状态实时监控	7
自助式多维统计分析	7
数据处理	8
可视化数据处理流程设计	8
丰富数据处理组件	8
高可用、动态扩容	9
灵活调度、实时监控	9
数据资产管理	9
规范的管理流程	10
动态可定义元模型	10
自动化元数据采集	10
智能全文检索	10
数据资产地图	11
数据质量	11
一站式质量管理	11
图形化、零编码	11
丰富便捷的质量规则	12
智能调度、实时告警	12
定制化质量评估体系	12
全方位质量分析报告	13
数据服务	13
便捷服务开发	14
完善数据服务管理流程	14
多重安全管理	14
分级流控策略	15
细粒度服务访问审计日志	15
多维统计分析	15
统一调度	16
统一运维	16
平台安全	16



背景及定位

背景

党的十九大报告提出要「推动互联网、大数据、人工智能和实体经济深度融合」,大数据、云计算、物联网、“互联网+”等新理念、新技术快速崛起,政府、企业信息化建设正处于从分散转向集约、从孤立封闭转向共享开放、从以政府推动为主转向政企合作推进的重要转型期,即将迈入全面联网、业务协同、智能应用的新阶段。信息技术与经济社会的交汇融合引发了数据迅猛增长,数据已成为国家基础性战略资源,正日益对生产、流通、分配、消费活动以及经济运行机制、社会生活方式和国家治理能力产生重要影响。“互联网+”、大数据已上升成为国家层面发展战略,掌握丰富的高价值数据资源日益成为抢占未来发展主动权的前提和保障。

随着大数据时代的来临以及信息技术的发展,数据量正在以指数级的速度在增长,并且数据正以多元结构(结构化与非结构化)趋势发展,如此海量、多元结构的数据需要用有效的方法对齐其管理,挖掘其价值;数据逐步成为一个企业或政府部门的核心资产,数据作为战略资产来管理,需要建立一套从数据采集、处理到共享应用的管理机制,以整合数据为基础、提高数据质量的手段,达到广泛的数据共享目标,最终实现数据价值最大化,数据中台正是在以上背景的情况下产生的一套数据管理机制,数据中台是将数据转化为智慧,挖掘出价值,最终服务于社会的重要途径。

产品定位

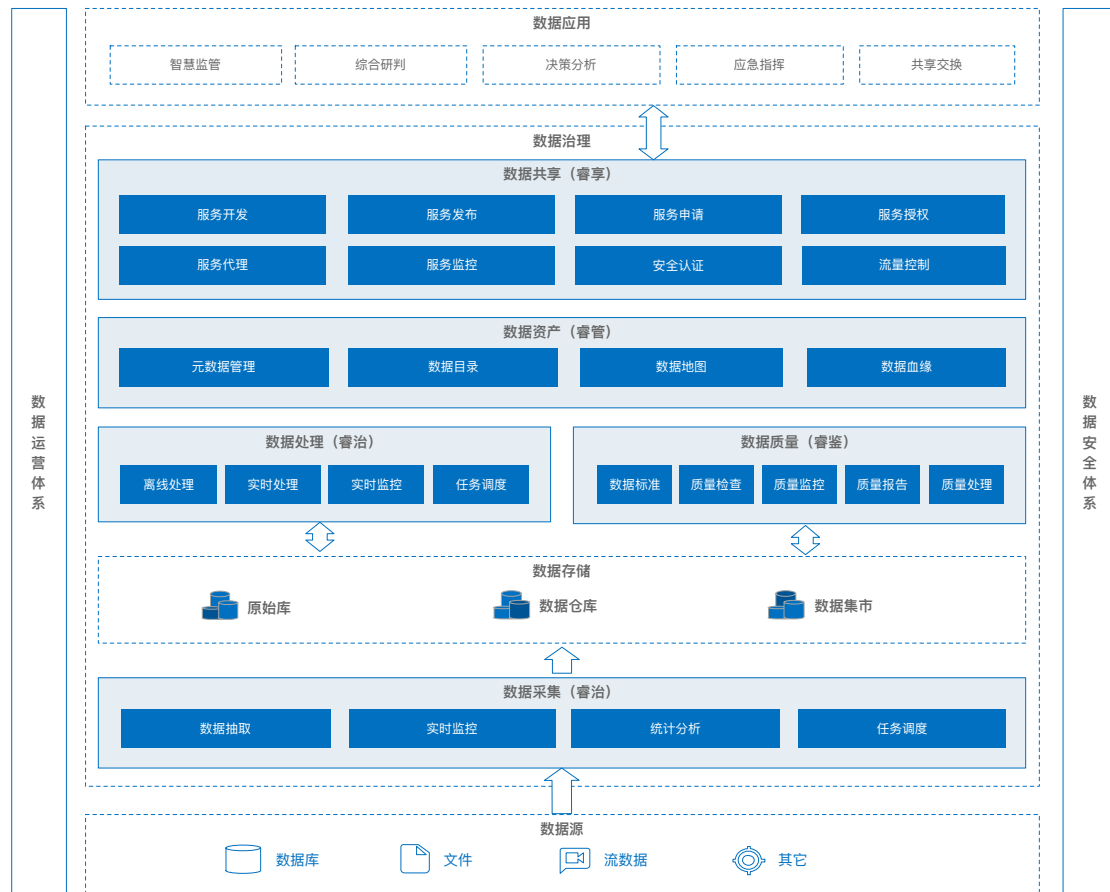
公司自主开发的中台(云睿中台)系列产品以**整数据、管资产、通服务**为基

基础思路，产品以层次化、插件化、体系化的松耦合架构支撑数据中台的全周期建设，实现**数据资产化、资产服务化、服务价值化**的建设目标，其系列支撑产品包括：数据集成（睿治）、数据质量（睿鉴）、数据资产（睿管）、数据共享（睿享）。



功能介绍

系统架构



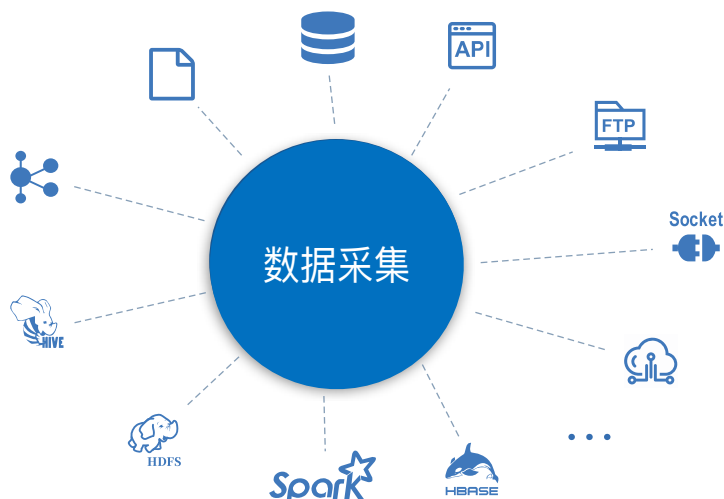
云睿数据中台通过数据技术，对分散的海量数据进行异构数据采集聚合、计算、统一存储、融合加工、统一规范，形成标准数据，并对标准数据提供数据分析，业务化数据服务，支撑个性化数据应用，为客户提供给高效服务。

数据采集

云睿中台依托云海睿治数据集成平台组件完成分布式数据采集汇聚。支持批量数据迁移、实时数据集成和数据库实时同步，支持多种异构数据源，全向导式配置和管理，支持单表、整库、增量、周期性数据集成。

多源异构数据采集加载

云海睿治数据集成平台提供多源异构数据采集能力，广泛适配多种数据协议，支持结构化、半结构化和非结构化数据抽取采集和加载：



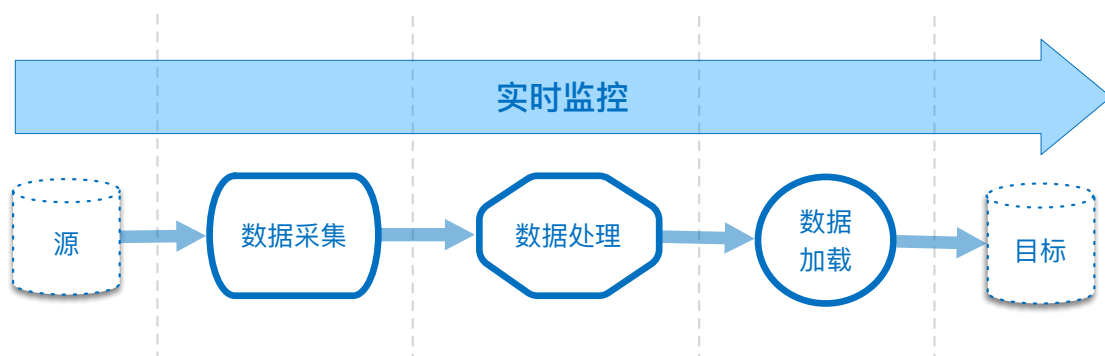
- 数据库：适配主流的关系型数据库的数据采集和加载，如 Oracle、MS SQL Server、Sybase、IBM DB/2、InfoMix 等商用数据库，MySQL、PostgreSQL 等开源数据库，神通、达梦、Kingbase、GBase 8t 等国产数据库；
- 数据文件：CSV、JSON、XML、AVRO、EXCEL 等；
- 通讯协议：JMS、KAFKA、MQTT、HTTP(S)、WebSocket、FTP、SFTP、MAIL、SNMP、AMQP、UDP、TCP 等；
- 大数据：HDFS、Hive、HBase、Spark、MongoDB、InfluxDB、Elasticsearch、Cassandra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

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。

采集状态实时监控

云海睿治理提供对采集和分发过程中数据的生产、消费、以及采集链路进行实时监控，提供完善的系统状态、集群状态、数据任务状态实时监控，帮助管理员及时发现系统故障和异常并及时恢复。



自助式多维统计分析

系统提供按照数据来源系统、数据来源部门、数据采集任务等多维自助式统计分析，分析结果提供图表化直观展现。

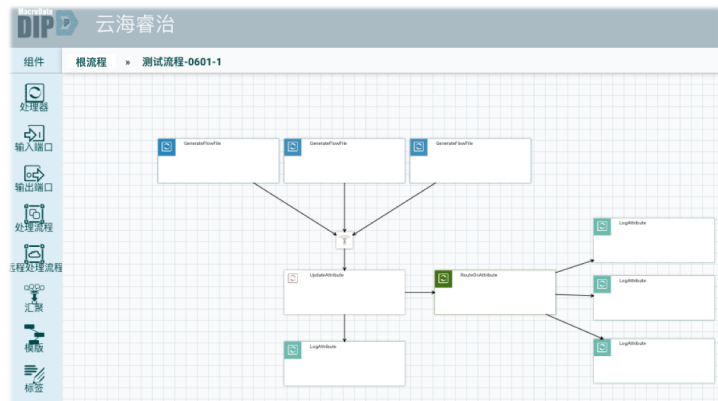


数据处理

云睿中台依托云海睿治数据集成平台组件完成对数据的清洗、转换、融合处理。

可视化数据处理流程设计

平台提供可视化拖拽式方式设计数据处理流程，通过可视化操作界面完成数据处理流程的设计、部署、调度、监控。



丰富数据处理组件

系统提供接近 300 个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、数据路由、数据去重、数据合并、数据验证、数据计算、脚本编辑.....

组件分类

全部 (289)

关系数据库 (13)

+ 基础协议 (65)

+ 大数据 (65)

+ 云平台 (34)

数据转换 (16)

数据验证 (3)

数据路由 (8)

聚合拆分 (11)

加密压缩 (9)

+ 工具 (37)

+ 其它 (27)

+ 监听 (1)

展示 13 中 289

过滤

类型	版本	标签
ExecuteSQL	1.11.4	关系数据库
ExecuteSQLRecord	1.11.4	关系数据库
GenerateTableFetch	1.11.4	关系数据库
GenerateTimeBasedCdcSqlProcessor	2.0.0	关系数据库
ListDatabaseTables	1.11.4	关系数据库
PutDatabaseRecord	1.11.4	关系数据库
PutSQL	1.11.4	关系数据库
QueryDatabaseTable	1.11.4	关系数据库
QueryDatabaseTableRecord	1.11.4	关系数据库
QueryRecord	1.11.4	关系数据库
QueryTriggerCdc	2.0.0	关系数据库
UpdateDatabaseRecord	2.0.0	关系数据库

UpdateDatabaseRecord 2.0.0

com.macrodata.dip.core - dip-database-nar

该处理器使用指定的“RecordReader”输入来自传入数据包的记录(可能有多条)。这些记录被转换成SQL语句并作为单个批处理执行。如果发生任何错误，数据包将被路由到失败或重试。如果成功传输了记录，则传入数据包将被路由到成功。处理器执行的语句类型通过“语句类型/属性指定，该属性接受一些硬编码的值。比如“INSERT”、“UPDATE”和“DELETE”，以及“Use Record operationType”。“Use Record operationType”可以与<QueryTriggerCdc>处理器配合使用，将根据该处理器数据包中operationType的值完成SQL语句的“INSERT”、“UPDATE”和“DELETE”执行操作。“更新...

高可用、动态扩容

数据处理引擎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），实现数据任务流程高可用；

同时根据实际负载情况，在不中断数据任务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，提高数据处理效率。

灵活调度、实时监控

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。并提供实时监控告警，及时掌握数据处理状态。

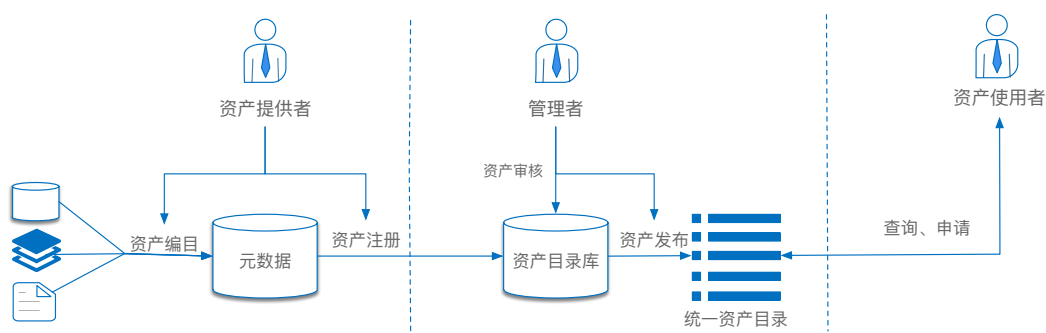


数据资产管理

云海睿管数据资产管理组件提供对数据资产的规范化、精细化、可视化管理。辅助管理者全面掌握数据资产分布、存储和流动情况，厘清数据资产、血缘关系，总览数据全局，全景可视，提供数据资产统一智能检索和运营监测。

规范的管理流程

云海睿管为企业提供统一数据资产管理，为数据资产提供者提供在线编目、注册，资产管理者在线审核、发布，数据资产使用者在线浏览、申请使用的科学管理里流程，保证数据资产注册发布、审核管理、使用的规范合理性。



动态可定义元模型

系统提供动态的数据资产元模型管理，支持动态建模，适应不同行业资产管理相关标准，可动态可视化编辑定义数据资产元数据模型，生成符合不同行业的数据资产元模型模板。

自动化元数据采集

系统不仅支持传统人工交互式元数据采集，同时提供自动化数据资产元数据自动采集和导入。有效降低数据资产复杂度，提高数据资产编目效率。

智能全文检索

智能全文检索技术允许用户简单快速检索数据资产目录，并支持智能排序，便于用户便捷发现自己感兴趣数据资产。

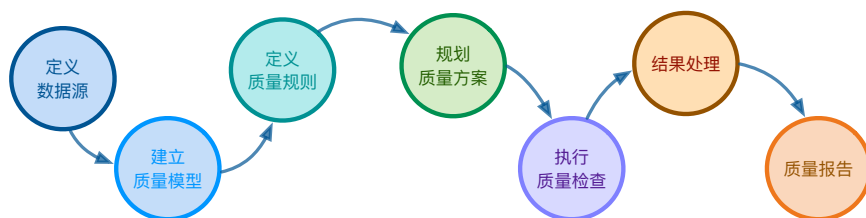
数据资产地图

多维度、多层次、图像化 360°全景数据资产查看，全面掌握数据资产；明确数据资产分布、存储和流动情况，掌握资产元数据详情、血缘关系；帮助数据管理者总览数据全局，实现数据资产统一检索、监测和维护等全方位管理。

数据质量

云海睿鉴数据质量平台围绕数据真实性、数据准确性、数据唯一性、数据完整性、数据一致性、数据关联性、数据及时性等一系列数据质量维度，通过有效的数据质量控制手段，进行数据质量的管理和控制，消除数据质量问题，提升企业数据变现的能力。

一站式质量管理



系统提供数据质量管理的数据范围定义、质量模型建立、质量规则配置、质量方案规划、质量检查、结果处理、质量分析报告的全周期管理。

图形化、零编码

平台提供交互式图形化界面，操作简单，无技术门槛，用户无需编码基础即可完成质量管理的定义和检查监控，质量分析。

丰富便捷的质量规则

▼ 规则(7)

- ▢ 重复数据检查(1)
- ▢ 引用完整性检查(1)
- ▢ 规范检查(1)
- ▢ 逻辑检查(1)
- ▢ 空值检查(2)
- ▢ 值域检查(1)

平台提供了引用完整性检查、规范性检查、逻辑性检查、重复性检查、值域、空值检查等多种数据检查规则定义方法，为用户提供全方位数据质检监测保障。

智能调度、实时告警

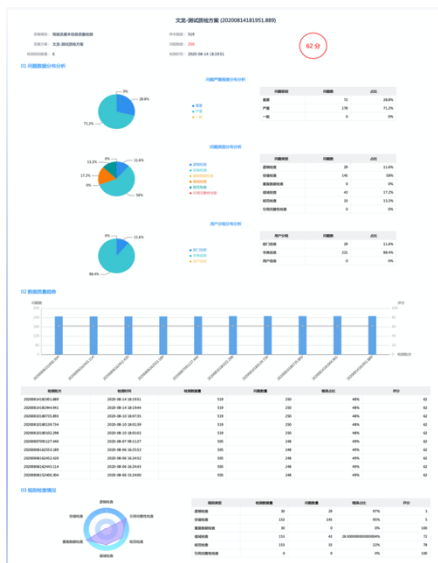
系统支持灵活调度规则，允许按照调度方案的规划的质量规则，完成质量检查，依据检查结果针对问题数据提供给你多方式及时告警，完成全方位质量分析，形成质量分析报告，推进数据质量改进。



定制化质量评估体系

用户可自定义质量规则的问题级别、评分权重，满足不同领域、不同场景对不同质量检查纬度的侧重，从而给出更切合实际的质量评估报告。

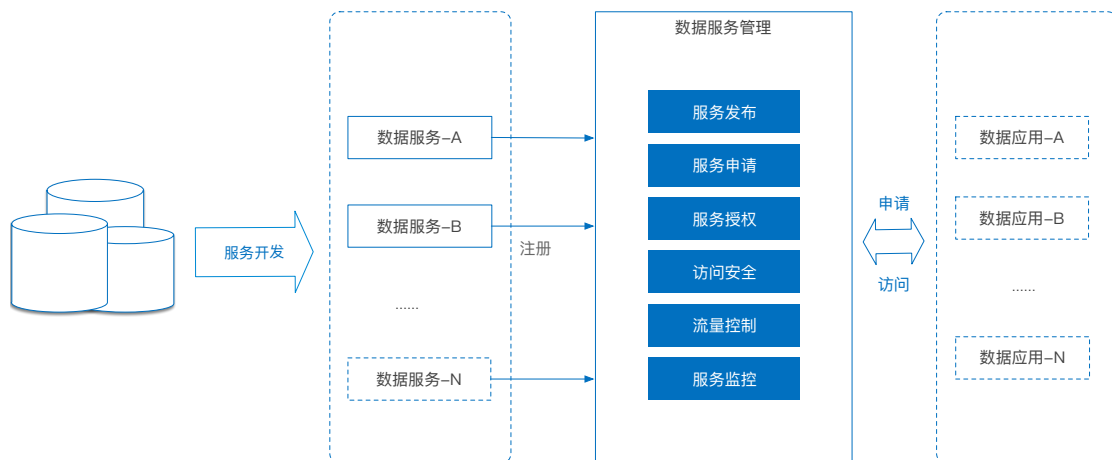
全方位质量分析报告



对每次质量检测都可以从问题严重程度、规则类型分布、质量趋势等多个纬度进行质量分析,并结合质量规则的权重进行综合评分,给出质量分析报告,有效帮助用户提升数据质量。

数据服务

云海睿享数据共享平台提供企业级的数据共享管理,为数据服务方和数据需求方提供服务开发、注册、审核、发布、申请、授权等管理过程的统一管理,运行时提供高性能、高可靠的分布式服务网关完成数据服务统一接入代理、安全控制,辅助用户实现数据服务发布,安全管理、运行监控,帮助用户快速、低成本、低风险的实现服务聚合、能力开放、优化重构;



便捷服务开发

系统提供数据服务开发能力，可通过管理中心的图形化界面，方便的将数据源数据快速开发数据服务接口，实现以数据服务形式提供数据共享。

开发过程中，无需编码，屏蔽服务开发技术细节、简化服务开发，遵循规范标准，降低数据服务开发难度。



完善数据服务管理流程

系统为数据服务方和需求方的服务注册、审核、发布、申请、授权的完善管理流程。

多重安全管理

共享服务的安全是数据共享关注重点，共享服务安全面临多方面的挑战：未知的访问者、未授权的服务访问等。共享服务通过访问控制、黑白名单等手段为数据服务的安全提供了多重安全保障。

■ 访问控制

对共享服务消费者身份进行定义和识别，并在服务授权时控制服务访问权限，授予访问密钥，密钥具有有效期，服务管理员可随时对密钥的有效性

和有效期进行动态调整，有效控制服务访问授权。

■ 黑白名单

针对共享服务的授权用户，考虑不同应用场景，可动态加入访问黑/白名单控制策略，限制授权用户仅可在指定的授权服务器访问数据共享服务，确保数据服务访问的合法和安全。

分级流控策略

流控策略主要是为共享服务提供流量控制管理策略，防止合法授权用户对服务的不合理或者非法超频调用，同时通过流量控制避免了对平台的恶意攻击，保障数据服务的可用性。

系统支持两级流控策略：

- **服务级流控策略**：允许对数据服务设置整体流控策略，可限制指定服务单位时间内的调用次数。
- **应用级流控策略**：允许对某个指定服务的特定消费者应用设置流控策略，限制该消费者应用单位时间内调用该服务的次数。

细粒度服务访问审计日志

系统对数据服务的访问提供实时访问日志记录，包括：受访服务、访问应用、访问时间、响应状态、响应时间、响应流量、访问异常信息等。为数据服务的性能分析，故障诊断分析提供充足依据。

多维统计分析

平台对所有服务请求日志从多个维度进行实时计算及历史数据的统计分析，

以多种图表模式进行直观展示包括，服务访问统计分析、服务响应时间分析、服务响应流量分析、服务健康分析、服务热点分析等。



统一调度

全面托管的调度，支持按时间、事件触发的任务触发机制，支持分钟、小时、天、周和月等多种调度周期。

统一运维

可视化的任务运维中心，监控所有任务的运行，支持配置各类报警通知，便于责任人实时获取任务的情况，保证业务正常运行。

平台安全

统一的安全认证，租户隔离，数据的分级分类管理，数据的全生命周期管理，保证数据的隐私合规、可审计、可回溯。