数 据 引 擎

产 品 手 册

汇辰技术服务（贵州）有限公司

目录

1 数据引擎能做什么？

数据采集：

（1）可以采集境内/境外的各类网站数据，是数据采集系统通用性的可视化+开发型爬虫软件；

（2）可以采集互联网上几乎 100%的公开数据。境外网站通过配置境外代理 IP 或 VPN，即可轻松访问；

（3）数据引擎不仅能进行可视化的数据提取，还支持正则表达式操作，更有强大的面向对象的脚本语言系统，定制化程度高；

（4）解决 get 请求与 post 请求灵活切换采集问题；

（5）支持采集关键词检索/含验证码/注册/登陆网站，并且支持验证码 OCR 或第三方 API 接口，支持验证码识别库的训练；

数据清洗：

（1）自动识别各类数据特征，挖掘各种特征文本；

（2）自带各种数据清洗方式，采集干净的数据无需二次清洗；

（3）智能挖掘分析与智能多维度综合分析，并自动生成可视化数据分析报告与专属大数据报告。

（4）支持定制化清洗程序，识别特征文本，一次性完成采集-清洗-发布，无需二次清洗；

（5）支持 AI 建模进行识别特征文本；

数据发布：

（1）数据发布：自动将生成的数据分析报告与专属大数据报告等发布到腾讯、微博、新浪、今日头条等新闻资讯平台；

（2）支持导出多种数据库：可直接将数据实时采集/清洗导出到所关联的数据库中，支持数据库类型有：SQL Sever、MySQL、Access、Oracle 等；

（3）支持定制化导出方案；

数据可视化：

（1）采集数据实时刷新；

（2）接入各个数据库，实时处理海量大数据；

（3）支持常见 40+种图表，原始数据直接生成图表；

（4）拥有海量的常用分析模板，零编码拖拽式操作；

配套服务：

（1）定制采集/清洗/发布规则模板；

（2）规则在线交易，支持在线沟通，平台交易等功能，保障后期规则维护；

（3）数据代采：一站式为用户提供采集、清洗、发布、运维服务；

（4）采集托管：通过远程控制用户服务器，实现入场数据采集的运维工作；

（5）数据定制：根据自身需求，定制数据导出内容；

（6）“图即代码”开发工程师培养；

（7）专属客服服务；

2 、基本介绍

2.1简介

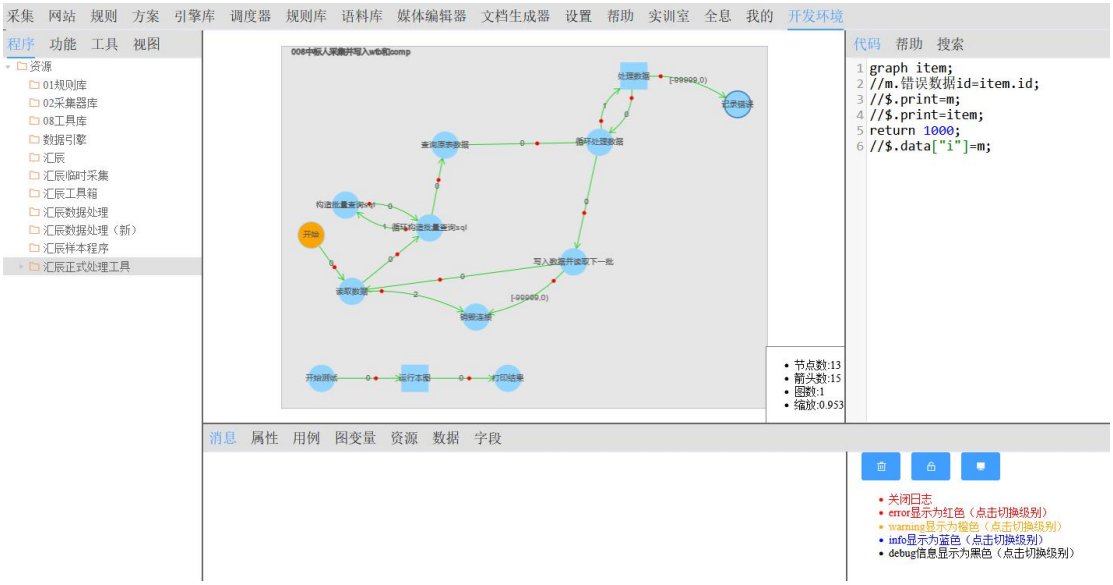
数据引擎是一款数据全流程一站式处理工具，它不仅大幅缩短数据多轮处理的工作量与数据处理的时间，还能大幅减少开发人员的开发工作量，数据引擎不但为企业减少用工成本、时间成本、开发成本、维护成本等，而且还可以为企业提供自定义的数据分析统计价值，为企业将零开发经验员工培养成为一名合格的开发工程师；

数据引擎是集数据采集、清洗、转换、存储、统计、智能挖掘分析、定制化分析结果输出、数据流程可视化等于一体的数据全流程处理引擎，可使全流程处理自动化且全流程处理中支持逻辑插入，并支持用户之间进行采集/清洗/发布等规则交易，成品数据交易，定制化程序交易等多维度数据交易应用场景。

整合了网页数据采集、移动互联网数据及 API 接口服务（包括数据爬虫、数据优化、数据挖掘、数据存储、数据备份），支持导出到表格、文本、数据库、网站 API 等各种格式，定制计划任务（可随机启动和可固定时间启动），实现自动采集、自动发布，无需人工操作，通过智能算法和定制化规则，自动识别并提取文本中的某一个关键内容，一键采集数据。

数据引擎采用全新的逻辑表达界面和独有的定制开发简易语言，将程序运行逻辑用一张运行指示图展示出来，清晰透明的表达出运行逻辑（即：图即代码），图即代码不仅吸收了 C++、JS 的语言风格，而且实现了逻辑与代码分离，自带框架于语言之内，扩展性好，语法简单，可零经验参与开发，无沟通门槛，不需要重复学习各类框架，仅需少量代码即可快速搭建后台系统，开发效率在 JAVA 的 10 倍以上，使得基于 Web 的应用程序的开发变得迅速和容易，它依托达芬奇快速开发工业应用，基于 opc 工业总线通信实现实施工业互联网智能制造技术升级，将计算过程透明化、使用过程简单化、应用场景人性化，通过图中的节点链接关系描述程序计算逻辑，实现处理逻辑与代码分离，使业务的管理者参与到程序逻辑设计中，更好的把控开发方向，并且图、节点、箭头就可以完成程序的逻辑设计，极大的减少了代码量，并降低了编程门槛、开发周期、开发成本等。

数据引擎内置数据库、知识库管理引擎和 IT 数据系统管理工具，既可以快速的搭建 Web 服务系统，也能够为企业内部构建 IT 信息系统提供了较完整的解决方案。

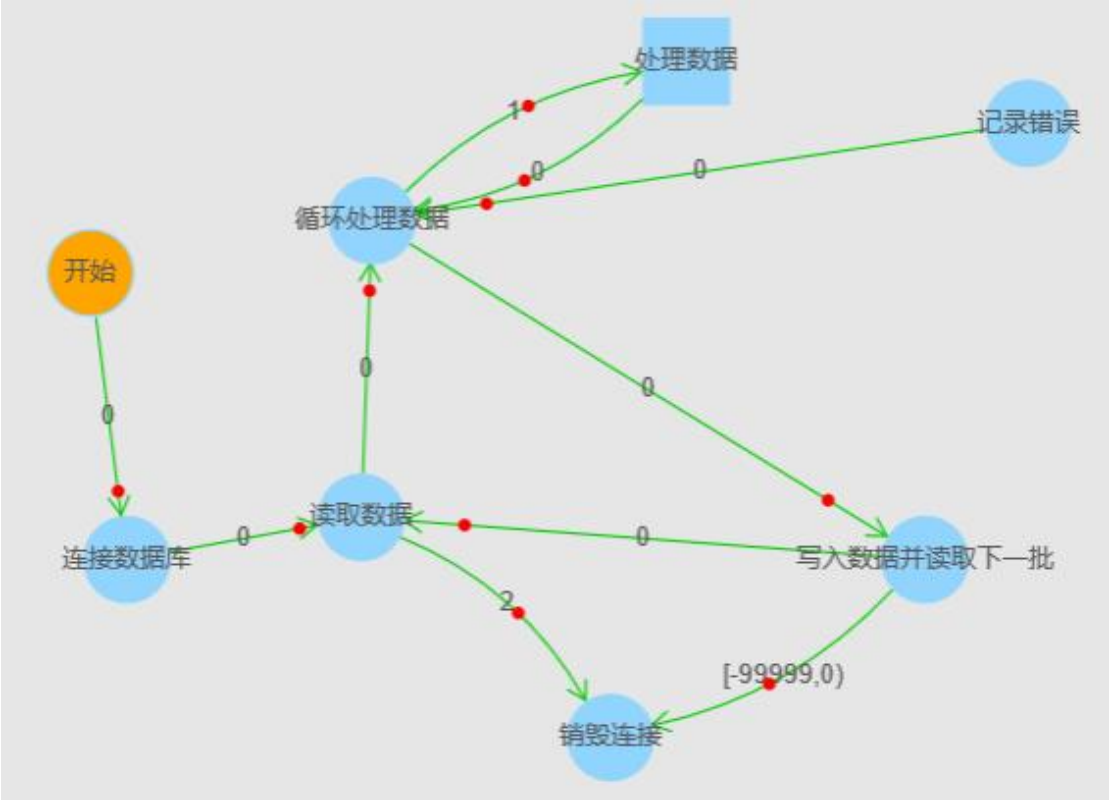


数据引擎展示图

2.2数据引擎的核心能力

2.2.1 图即代码

图即代码是一门简单易学的定制开发语言，它采用全新的逻辑表达界面，将程序运行逻辑用一张运行指示图展示出来（如：图即代码逻辑图），清晰透明的表达出运行逻辑，图即代码不仅吸收了 C++、JS 的语言风格，而且通过图中的节点链接关系描述程序计算逻辑，实现了逻辑与代码分离，使业务的管理者参与到程序逻辑设计中，自带框架于语言之内，扩展性好，语法简单，可零经验参与开发，无沟通门槛，不需要重复学习各类框架，仅需少量代码即可快速搭建后台系统，开发效率在 JAVA 的 10 倍以上，使得基于 Web 的应用程序的开发变得迅速和容易。



图即代码逻辑图

2.2.2 数据全流程（一站式）处理

数据引擎是一款数据全流程一站式处理工具，支持采集、清洗、转换、存储、统计、智能挖掘分析、定制化分析结果输出、数据流程可视化等全流程一站式处理。

采集引擎：

（1）满足任何数据采集场景；

（2）精准采集，轻松挖掘；

（3）多规则自动匹配；

清洗引擎：

（1）根据特征挖掘文本数据；

（2）智能定位字段取值区域；

（3）智能识别表格列表多值；

（4）自动识别文本语义；

（5）自定义主题识别过滤；

（6）自由选择清洗规则，且可自定义添加“图即代码”作为清洗规则；

发布引擎：

（1）支持多种数据库：可实时将数据存储到所关联的数据库中，如：SQL Sever、MySQL、Access、Oracle 等。

（2）内置亿万级分析型数据库：对已有数据进行定制化分析并导出相应结果；

（3）支持分布式数据存储：实现多机集群采集，满足不同企业不同部门数据采集场景，为企业提供任何场景的分布式部署方案。

（4）本地存储，安全部署：绝对保证用户数据的私有性和安全性。安全性远高于市场上的云采集器及浏览器采集等爬虫软件。

数据分析导出引擎：

（1）支持数据统计分析，且支持统计维度可视化：具有对海量数据的分析计算的能力，融合独创的深度大数据分析理论，可将数据进行文本的语义识别，从多维度模块化进行分析。

（2）自定义分析程序：可自定义添加“图即代码”作为分析工具，如：分析市场、行业、竞争对手的情报信息等；

（3）支持数据分组批量分析结果；

2.2.3 引擎交易、规则交易

用户可选择将自己建立的采集/清洗/发布/分析/提取等规则上传至数据引擎云端服务器，数据引擎同样支持整体交易，设置金额进行在线交易规则，可以将数据资产进行等价交换，促使大众人人参与。

（1）支持线上付费；

（2）支持在线联系；

（3）支持规则维护；

（4）在线指导；

2.2.4 定制化解决方案服务体系

基于“汇辰一家”平台海量数据定制数据分析程序：

（1）用户可在数据引擎中选择平台的成品数据分析程序，为用户输出相应的数据分析结果；

（2）用户可根据需求自行开发数据分析程序，向平台提出请求调用平台数据，为用户输出相应的数据分析结果；

（3）用户可联系平台定制开发数据分析程序；

（4）支持分析结果导出；

（5）支持分析结果可视化；

基于用户自有数据定制分析程序：

（1）用户可将自有数据配置到平台的成品数据分析程序中，为用户输出相应的数据分析结果，分析自有数据资产；

（2）用户可根据需求自行开发数据分析程序，为用户输出相应的数据分析结果；

（3）用户可联系平台定制开发数据分析程序；

（4）支持分析结果导出；

（5）支持分析结果可视化；

基于“汇辰一家”平台与用户自有数据结合定制数据分析程序：

（1）用户可根据需求自行开发数据分析程序，平台提供数据查询接口，为用户输出相应的数据分析结果；

（2）用户可联系平台定制开发数据分析程序；

（3）支持分析结果导出；

（4）支持分析结果可视化；

3 、独特优势

3.1规则自动匹配与规则互用

规则自动匹配与规则互用是一项技术创新，不但可以自动识别采集规则，并且还能自行判断互用规则，达到一个规则、多个网站共同使用的目的，减少采集人员的大量建立规则的时间，从而达到降本增效的目的；

3.2丰富的底层数据

（1）拥有大量采集规则、清洗规则、智能挖掘分析与智能多维度综合分析规则；

（2）拥有大量企业数据：

企业基本信息、建筑资质、企业注册资格人员信息、施工许可、施工图审查、合同登记信息、项目关联企业信息、竣工验收备案信息法院公告、年报、风险、发票抬头、分支机构、资质证书、企业历史变更、著作权、人员、融资、对外投资、招聘、裁判文书、行政许可、历史名称、舆情、投资人及出资信息、专利、行政处罚、股权出质、主要产品、失信、商标、网站、执行、个人信息、个人与企业关系、企业信用评级、业绩、荣誉、上市公司财务数据、企业纳税等多维度数据及企业图谱画像；

（3）拥有大量招标采购数据：全国招采项目、拟在建项目、投资项目库、民间资本项目、PPP 项目、招商引资项目等超过 20 年历史数据+日更新数据。具体包括招标人信息、中标人信息、代理人信息、项目名称、中标金额、联系人、招标文件、项目概况等多维数据。

3.3独特的数据采集能力

3.3.1 采集范围广

可以采集境内/境外的各类网站数据，数据采集系统通用性的可视化+开发型爬虫软件，可以采集互联网上几乎 100%的公开数据。境外网站通过配置境外代理 IP 或 VPN，即可轻松访问；数据引擎不仅能进行可视化的数据提取，又支持正则表达式操作，更有强大的面向对象的脚本语言系统，定制化程度高。

3.3.2 支持采集各种数据类型

A.采集 Ajax 网页数据

数据引擎支持 Ajax 技术,可采集 Ajax 网页中的内容。

B.采集基于 js 页面数据

数据引擎抓取工具可自动解析 JS，采集基于 js 页面中的数据，即可采集页面中包含 JS 的数据。

C.采集 post 请求数据

数据引擎可采集数据在 post 请求中的网页内容，即采集 post 请求中的数据。

D.采集 get 请求数据

数据引擎可采集数据在 get 请求中的网页内容，即可抓取 get 请求中的数据。

E.采集需要 Cookie 的网站数据

数据引擎采集分析引擎可设置 cookie 来模拟登陆，从而采集需要用到 cookie 的网站内容。

F.采集需要 OAuth 认证的网页数据

数据引擎支持 OAuth 认证，可以采集需要 OAuth 认证的页面中的数据。

G.采集本地数据

数据引擎采集软件支持采集本地数据，可以采集本地文件中的数据。

H.采集内网网站

数据引擎抓取软件，是私有化部署，可安装在本地服务器中，采集内网网站数据。

I.登录采集数据

数据引擎采集器，可使用 Cookie 模拟登录网站，也可直接配置登录网站，从而采集到需要登录的网站、APP 中的数据。

J.采集关键词搜索的数据

数据引擎采集分析引擎可批量导入、修改关键词，从而采集到在页面中搜索关键词出来的数据内容。

K.采集带有翻页的数据

数据引擎可采集带有翻页的网页中的数据，例如：数字翻页、瀑布流翻页、下一页、更多等翻页均可采集。

L.采集批量数据源网站

数据引擎，配置一个模板采集成千上万个网站，可批量采集海量网站数据。

M.采集 HTTPS、HTTP 协议网页数据

数据引擎采集分析系统，支持采集 HTTPS、HTTP 协议网页中的数据。

N.采集具有反爬的网站

数据引擎采集引擎可智能模拟浏览器和用户行为，自带 IP 代理优化加速功能，突破封锁限制，可以有效采集具有反爬的网站数据

3.3.3 支持采集各种格式的数据

针对不同用户的采集需求，数据引擎可提供自动生成爬虫的自定义模式，可精准批量识别各种网页元素，还有翻页、下拉、AJAX、页面滚动、条件判断等多种功能，支持不同网页结构的复杂网站采集，满足多种采集应用场景。数据引擎内置自主研发的脚本语言，几乎能够 100%采集浏览器上公开可见的数据。

①支持基于 http/https 协议的网页采集

②支持插入 cookie 的网页采集

③支持关键词搜索采集

④支持添加基于同一个根域名下无法直接链接的网址

⑤支持 get/post 传输

⑥支持基于 js/ajax 加载页面的采集

⑦支持需要登录页面的采集

⑧支持输入验证码页面的采集

⑨支持图片、视频、直播页面的采集

⑩支持 utf-8 与 GBK 相互转码

3.3.4 采集附件、图片、音频、视频、直播数据

可采集网页、APP 中的视频及直播视频数据，还可以设置视频文件大小，筛选采集指定大小范围内的视频数据。自带前嗅国产千万量级数据库，支持多种数据库和 Excel，海量视频数据存储毫无压力。

（1）采集音频

可抓取网页、APP 中的音频数据。

（2）采集图片

可采集网页、APP 中各种格式(bmp、jpg、tiff、gif、pcx、tga、exif、fpx、svg、psd、cdr、pcd、dxf、ufo、eps、ai、raw 等)的图片数据。

（3）采集 pdf 文件

可采集网页、APP 中的 pdf 文件数据。

（4）采集 word 文件

可采集网页、APP 中的 word 文件。

（5）采集表格文件

可采集网页、APP 中各种格式（xls、csv、xlsx 等）的表格文件。

（6）采集各种附件

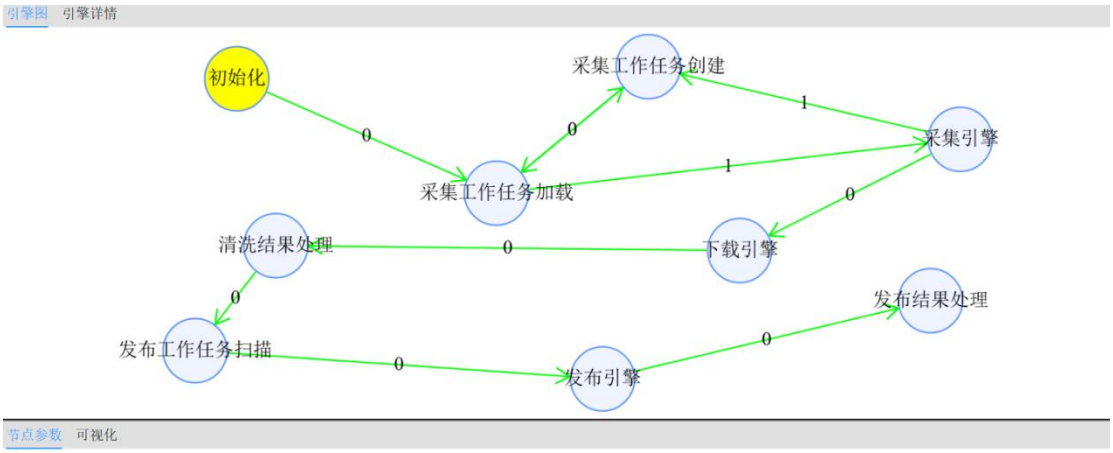
可采集网页、APP 中的各类附件数据。

以上数据导出时，自定义命名字段，可分批导出分割存储到不同的文件夹中。

4 、数据引擎功能介绍

4.1总引擎

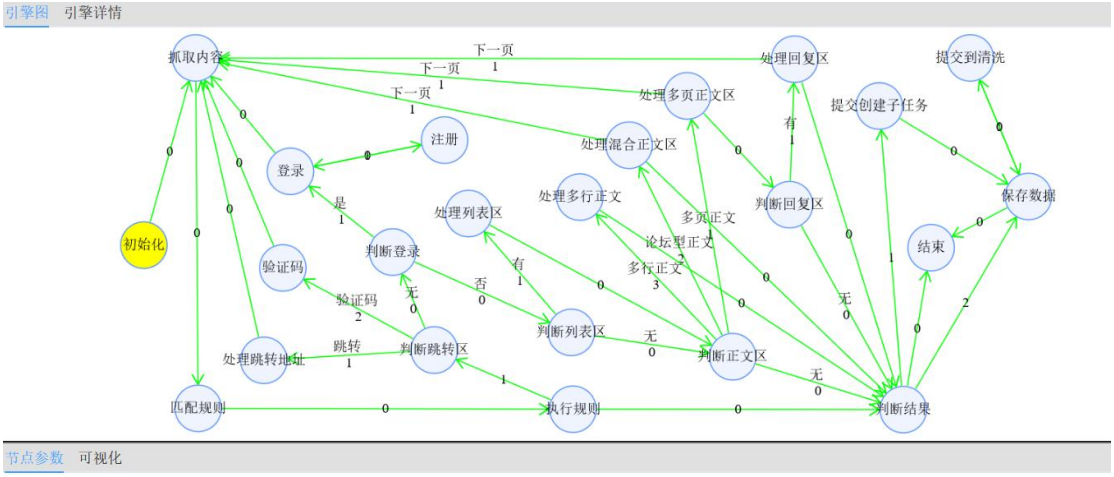
总引擎是数据引擎的总配置图，用户可根据自身需求，配置采集、清洗、发布、下载等任务，总引擎支持单选与多选；



总引擎示意图

4.2采集引擎

采集引擎负责采集器的配置，可根据采集需求，自行配置采集任务，采集引擎还负责自动匹配规则、规则互用的自动智能选取，实现多场景自定义配置、多维度自定义运行设置、多策略反爬设置，支持登录注册、验证码、动静态代理 IP、VPN、智能 AI 识别等功能；



采集引擎示意图

4.3清洗引擎

清洗引擎是负责数据处理的工具，他不仅负责数据的清洗工作，并且还负责数据的智能挖掘分析与智能多维度综合分析，自动生成可视化数据分析报告与专属大数据报告。

清洗引擎可对采集数据或数据库中的数据进行智能挖掘分析与智能多维度综合分析，平台不但提供了较为完善的智能挖掘分析与智能多维度综合分析规则供用户进行选择，而且用户还可以自定义添加“图即代码”来进行数据处理，根据自身需求进行定制化展示与结果输出。

（1）语义识别，智能分类

自动进行内容的语义识别，过滤不需要的主题，实现数据自动智能分类；

（2）多维度可视化图表

多维度对采集的数据进行自动分析，形成丰富的可视化图表，全网商情数据直观呈现；

（3）全自动数据报告

完全机器生成的数据报告，能够精确处理大量数据信息，同时节约大量人力成本；

4.4发布引擎

发布引擎负责数据的发布与导出，可将数据发布到各平台、数据导出多种数据库等。

数据发布：自动将生成的数据分析报告与专属大数据报告等发布到腾讯、微博、新浪、今日头条等新闻资讯平台；

导出多种数据库：可直接将数据实时采集/清洗导出到所关联的数据库中，设置链接数据库，将数据库关联至本地服务器上的第三方数据库中，支持数据库类型有：SQL Sever、MySQL、Access、Oracle 等。

5 、应用场景

（1）可用于需求调研、数据建模、架构设计、流程算法、产品开发、云服务等多个领域。

（2）对于完全没有数据采集经验的使用者，可以通过可视化的操作方式完成数据采集。

（3）对于有一定数据采集经验的使用者，可以自己定义规则来满足采集/清洗/发布一站式处理的需求。

（4）对于有丰富数据采集经验的使用者，可以自定义任务，满足对各种网站采集/清洗/发布一站式处理的需求。