

汇通大数据平台白皮书

2019 年

目录

1	汇通大数据平台概览.....	6
1.1	概述	6
1.2	框架	6
2	统一认证平台	10
2.1	总体介绍	10
2.2	产品功能	11
2.2.1	统一登录页面	11
2.2.2	权限控制	15
3	数据服务平台	19
3.1	总体介绍	19
3.2	产品功能	21
3.2.1	接口注册.....	21
3.2.2	接口发布.....	24
3.2.3	接口管理.....	26
3.2.4	负载均衡.....	29
3.2.5	路由管理.....	33
3.2.6	故障隔离.....	36
3.3	功能展示	39
4	数据集成平台	45
4.1	总体介绍	45

4.2	产品功能	47
4.2.1	数据采集转换与加载	47
4.2.2	流程开发和任务调度	50
4.3	功能展示	51
4.4	应用场景	54
5	数据计算存储平台	56
5.1	总体介绍	56
5.2	产品功能	61
5.2.1	分布式存储系统	62
5.2.2	多样的计算引擎	64
5.2.3	NOSQL 数据库	66
5.2.4	统一资源调度管理	68
5.2.5	分布式应用协调	69
5.2.6	内存计算增强 Spark	70
5.2.7	任务调度 Task Scheduler	71
5.2.8	用户管理 User Admin	74
5.2.9	运维监控管理 Manager	76
5.3	功能展示	78
5.3.1	Spark 内存计算引擎	78
5.3.2	任务调度使用	81
5.3.3	用户管理使用	83

5.3.4	运维监控使用	86
5.4	应用场景	89
5.4.1	场景 1: 企业级数据中心/数据湖系统	89
5.4.2	场景 2: 实时数据处理场景	92
5.4.3	场景 3: 明细清单实时查询	93
6	智能分析平台	95
6.1	总体介绍	95
6.2	产品功能	98
6.2.1	深度学习 DeepNEX	98
6.2.2	机器学习 HyperMiner	102
6.2.3	图分析	104
6.2.4	自助查询分析器 SQL Editor	107
6.3	功能展示	110
6.3.1	深度学习 DeepNEX	110
6.3.2	机器学习 HyperMiner	112
6.3.3	图分析 HyperGraph	115
6.3.4	自助查询分析器 SQL Editor	116
6.4	应用场景	121
7	数据治理平台	123
7.1	总体介绍	123
7.2	产品功能	126

7.2.1	元数据管理	126
7.2.2	数据质量管理	128
7.2.3	数据生命周期管理	129
7.2.4	数据安全性与隐私保护	130
7.3	功能展示	132
7.4	应用场景	137
8	前端开发框架	139
8.1	总体介绍	139
8.2	主要功能	139
8.2.1	PC 组件库	139
8.2.2	移动端组件库	140
8.2.3	微信小程序组件库	141
8.2.4	PC 端可视化布局	141
8.2.5	移动端可视化布局	141
8.2.6	微信小程序可视化布局	142
8.2.7	样式风格	143
8.2.8	组件搜索	143
8.3	功能展示	144
8.3.1	前端开发框架	144
8.3.2	组件库	144
8.3.3	可视化布局	147

8.3.4	样式风格	151
8.3.5	插件搜索	152
8.3.6	帮助文档	154
8.3.7	新手起步	154

1 汇通大数据平台概览

1.1 概述

当前,以大数据、云计算、移动互联网等为代表的新一轮科技革命席卷全球,正在构筑信息互通、资源共享、能力协同、开放合作的制造业新体系,极大地扩展了制造业创新与发展空间。新一代信息通信技术的发展驱动制造业迈向转型级的新阶段——数据驱动的新阶段,这是新的技术条件下制造业生产全流程、全产业链、产品全生命周期的数据可获取、可分析、可执行的必然结果,也是隐性知识显性化不断取得突破的内在要求。

习近平总书记强调,“要着力推动互联网与实体经济深度融合发展,以信息流带动技术流、资金流、人才流、物资流,促进资源配置优化,促进全要素生产率提升”。习总书记这段话深刻阐释了互联网与实体经济的关系,阐释了以互联网为代表的新一代信息通信技术融合创新推动实体经济转型升级的内在机理,也充分体现了工业大数据作为一种新的资产、资源和生产要素,在创新发展中的作用。

1.2 框架

汇通大数据平台是自主研发的一站式大数据服务平台。帮助企业快速搭建私有化的数据分析、数据服务与可视化平台,完成多数据整合,建立统一数据口径,并提供灵活、易用、高效的可视化探索式分析能力,多场景的应用展示和数据服务能力。提升企业数据洞察能力,并将数据决策快速覆盖各层员工及应用场景,

使企业成功转型为“数据驱动”的智慧型企业。

汇通大数据平台是一站式大数据协作开发平台，解决结构化、半结构化和非结构化数据的采集融合、数据治理、数据元标准管理、数据管理、数据服务等问题。多样化的开发组件适应不同场景，搭配高度自由的 DAG 工作流和强大的作业调度、运维面板，助力大数据项目的快速实施交付。平台采用 B/S 架构，兼容不同的底层技术平台，并分为七个大模块：数据集成平台、数据治理平台、数据存储计算平台、智能分析平台、数据服务平台、统一认证平台、前端开发框架等。



1) 数据集成平台

针对大数据类型多，数据量大，实时性要求高等特点，数据集成平台提供了多种格式的数据采集能力，进行数据的实时采集、实时处理，通过相关的分析规则和分析算法，可以进行实时分析、实时预警，帮助用户从大数据分析中获取最大业务价值。满足企业在复杂业务场景下的各类数据整合要求。

数据集成支持数据源和格式：

- 数据仓库和关系数据库管理系统 (RDBM) 包含事务性客户档案数据。
- 日志文件和事件数据，包括网络日志、应用日志等。
- 半结构化格式的数据，包括 XML、JSON 和 Avro。
- 平面文件，例如 CSV 格式的文件。
- NoSQL 数据库中的文件，例如 HBase、MongoDB 和 Cassandra。
- 从基于 Web 的 API 以及 FTP 服务器中提取的数据。
- 云和本地应用数据，例如客户关系管理 (CRM) 和企业资源规划 (ERP) 数据。
- 分析数据库，例如 HPE Vertica、Amazon Redshift 和 SAP HANA 平台。

2) 数据计算存储平台

数据计算存储平台构建了企业统一数据汇集与分析方案，实现自动化和信息化系统数据的“智能汇聚”，以私有云、公有云或混合云的方式实现数据互通和与外部关联企业间的知识。

数据计算采用分布式架构，支持流式数据计算和离线数据计算。数据存储支持 Hadoop 存储、内存数据库存储、图数据库存储、时序数据库存储。

采用数据计算存储平台可快速部署搭建数据仓库、数据湖。

3) 智能分析平台

智能分析平台算法和模型库包含了特征工程、机器学习、深度学习、运筹优化等专业数据建模工具，经过深度优化后的算法性能比开源算法库提速 3 — 10 倍，可加速进行数据挖掘探索，节省大量时间，为数据价值的发现提供了有力保

障。在此基础上，商业智能报表分析、深入挖掘、自服务式的数据分析等功能为企业应用提供了便捷的接口，开放了平台能力。

4) 数据治理平台

通过数据标准体系建设方法论+数据指标系统，统一数据指标口径，消除数据二义性。主要包含数据统一入口登记、数据规范定义、数据模型规范设计等功能，解决指标定义不一致、计算口径不一致、数据来源不一致带来的指标数据可信度低、严重影响分析决策的问题。

围绕完整性、准确性、一致性、及时性监控分析数据质量问题、提升企业数据质量。

从数据接入、数据加工、数据导出、指标、数据应用实现全链路血缘跟踪、提前预判数据是否能够准时产出、了解任务失败后影响分析以及快速地修复。

建立数据资产管理目录，统一数据查询入口，通过分层、分主题、业务过程记录、元数据的标签、注释以及血缘链路等元数据信息帮助数据开发、数据分析快速找到数据并理解数据。

5) 统一认证平台

统一认证平台实现大数据平台平台的统一身份认证，大数据平台各子平台共享身份认证功能。只需要登录统一认证平台，就可对被授予访问权限各子平台(数据集成平台、数据计算存储平台、智能分析平台、数据治理平台、数据服务平台、前端开发框架等)进行访问。

为了打通与其他应用系统间的访问，避免用户需要登录多套系统，统一认证平台还提供单点登录功能，其他应用系统集成了统一认证平台单点登录功能后，

就能通过统一认证平台进行统一访问管理。

6) 数据服务平台

数据服务平台提供数据服务接口，通过 API 网关对外提供数据服务。数据服务平台采用分布式部署架构，通过负载均衡管理实现各节点的负载分配。有效解决数据服务平台高可用、高性能、可扩展的问题。同时数据服务平台还提供服务降级、故障隔离等异常处理机制，保障数据服务平台的稳定运行。

7) 前端开发框架

前端开发框架提供 PC 端、移动端、微信小程序等前端开发框架，帮助开发人员采用可视化方式搭建前端框架，让开发人员把精力更多放到逻辑处理上，提升开发人员效率。同时，通过统一的样式库和布局模板，使整个前端框架风格统一，提升用户的体验。

2 统一认证平台

2.1 总体介绍

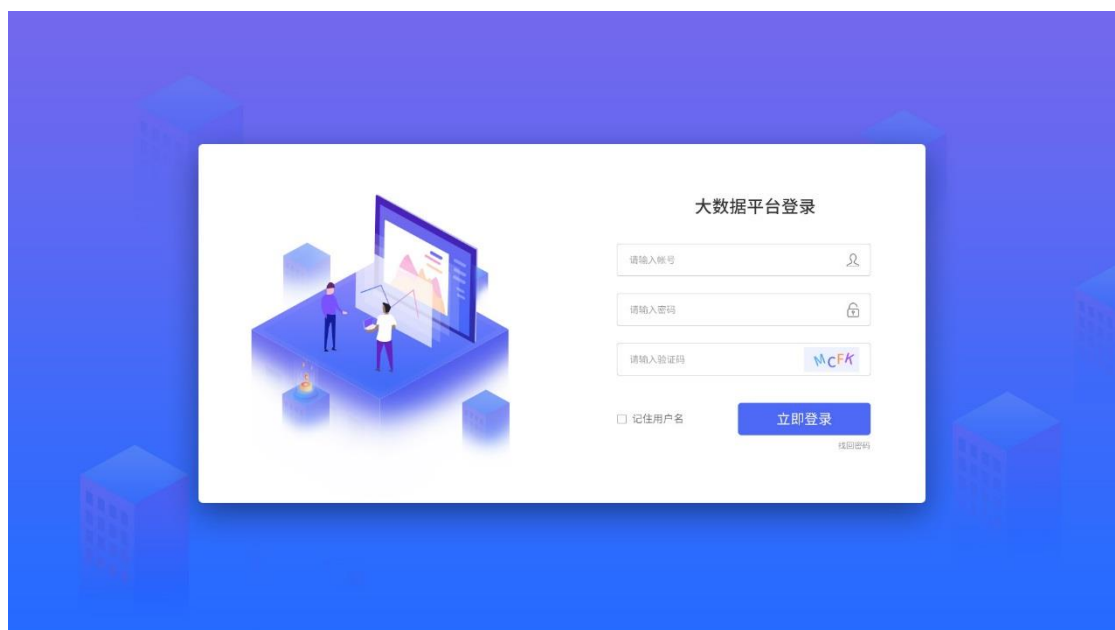
随着信息技术和网络技术的迅猛发展，企业内部的应用系统越来越多。比如在政务系统行业，垂直业务系统与各级网上办事大厅、政务服务中心系统等。由于这些系统互相独立，用户在使用每个应用系统之前都必须按照相应的系统身份进行登录，为此用户必须记住每一个系统的用户名和密码，这给用户带来了不少麻烦。特别是随着系统的增多，出错的可能性就会增加，受到非法截获和破坏的可能性也会增大，安全性就会相应降低。针对于这种情况，统一用户认证、单点登录等概念应运而生，同时不断地被应用到企业应用系统中。

2.2 产品功能

2.2.1 统一登录页面

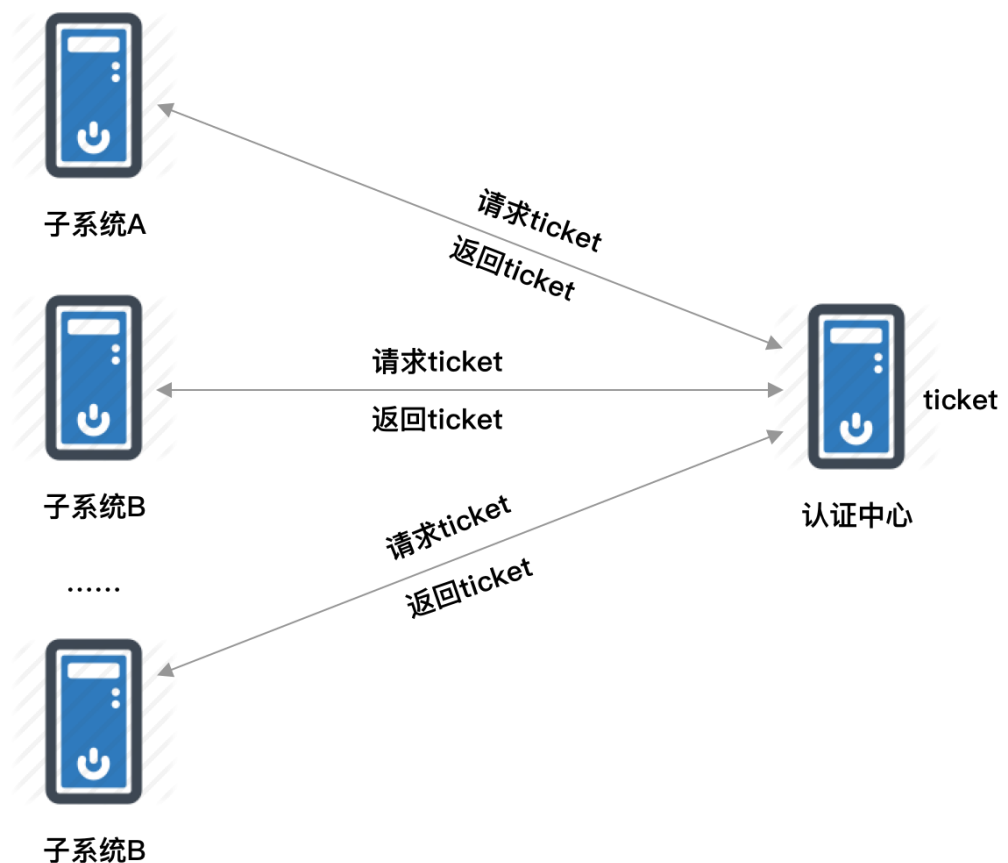
2.2.1.1 主要功能

实现共享统一身份认证平台，实现大数据平台的用户统一身份认证。如图所示：



2.2.1.2 产品特性

- 单点登录框架



如上图所示：在系统存在一个认证中心以及多个站点。用户信息统一由认证中心管理，在其它子站的登录均会跳转到认证中心来登录。

● 单点登录原理

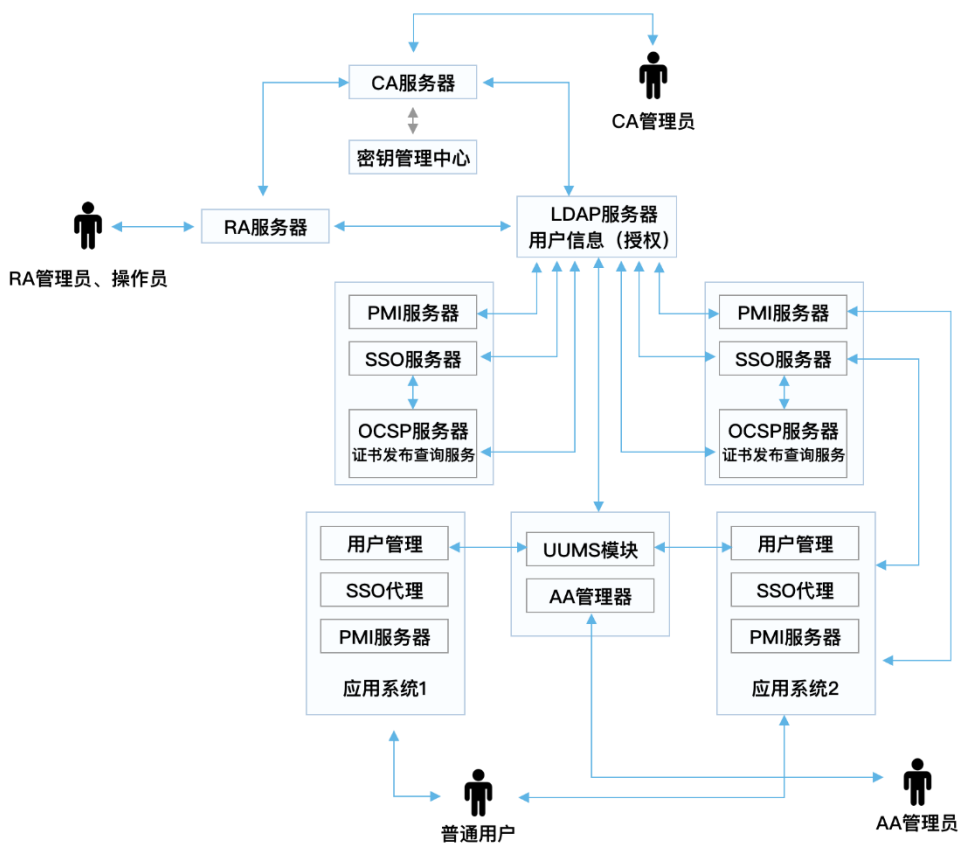
一般来说，每个应用系统都拥有独立的用户信息管理功能，用户信息的格式、命名与存储方式也多种多样。当用户需要使用多个应用系统时就会带来用户信息同步问题。用户信息同步会增加系统的复杂性，增加管理的成本。

例如，用户 X 需要同时使用 A 系统与 B 系统，就必须在 A 系统与 B 系统中都创建用户 X，这样在 A、B 任一系统中用户 X 的信息更改后就必须同步至另一系统。如果用户 X 需要同时使用 10 个应用系统，用户信息在任何一个系统中做出更改后就必须同步至其他 9 个系统。用户同步时如果系统出现意外，还要保证

数据的完整性，因而同步用户的程序可能会非常复杂。解决用户同步问题的根本办法是建立统一用户管理，实现单点登录。

用户认证中心具备以下功能：

- 统一用户管理。实现用户信息的集中管理，并提供标准接口。
- 统一认证。用户认证是集中统一的，支持 PKI、用户名/密码、B/S 和 C/S 等多种身份认证方式。



- 单点登录。支持不同域内多个应用系统间的单点登录。
- 单点登录流程

单点登录系统的关键是，当用户从一个子站登录后，它会在认证中心生成 ticket 标识，只要该标识存在就代表用户已经成功登陆了。每个子站在进入之前，无论本地 ticket 标识是否存在，均会去请求一次认证中心，要做的事情如下：如果子站 ticket 不存在，并不一定代表用户没有从别的子站登录，但是如果用户在别的子站登录过，则在认证中心一定会有 ticket 存在，则该次请求的目的就是去判断认证中心 ticket 是否存在，如果存在，则说明用户已经在别的子站登录了，要做的只是请求回来后在自己站点的 cookie 中种上 ticket，如果不存在则说明用户根本没有登录，则直接跳转至登录页面即可；如果子站 ticket 存在，则它去认证中心请求的目的在于，如果认证中心 ticket 已经不存在了则说明用户已经退出了，则请求返回后需要将本地 cookie 中的 ticket 清掉，达到“一登全登，一退全退”的效果，如果认证中心 ticket 还存在，则需要验证该 ticket 的真伪，是不是该客户端的（需借助一些其他手段）。详细流程如下。

先假设用户从未登录过系统，当他访问 A 站点时，会去认证中心请求一次 (jsonp 跨域的请求)，结果发现认证中心也没有 ticket，则跳转至认证中心的登录页面，当用户在认证中心登录后，则认证中心会跳转至 A 站点，此时 A 站点还会去认证中心请求一次，而这次会拿到 ticket 返回，就将 ticket 种植在本站的 cookie 中，代表用户在本站已经登录了。

当用户此时又去访问 B 站点时，同样会先去认证中心请求一次，请求的结果发现 ticket 存在，则它将 ticket 也种在自己站点的 cookie 中，这样用户就不需登录就可以在 B 站点随意访问了。

当用户在 B 站点操作完退出时，会去认证中心请求一次，该次请求会将认证中心保存的 ticket 值清掉。当在别的站点操作，去认证中心请求时，发现 ticket 已经没有了，同时也会将自己站点的 ticket 清掉，这样，又会跳转至登录页面。

- 单点登录安全

在 cookie 传递账号密码等信息是不安全的，虽然我们的 ticket 值是加密后的子串，但倘若别人截获了我的 ticket，然后直接利用 ticket 给服务端发送请求信息，则它同样可以通过验证。基于此，稍微安全一点的方法是，登陆时，在认证中心端，同时将客户端的 IP 地址记录下来，在验证 ticket 的环节同时对比此时请求验证的客户端的 IP 地址是否与之前保存的 IP 地址一致，只有一致才可以通过验证，否则返回失败。

2.2.2 权限控制

2.2.2.1 主要功能

系统都会有多个角色，不同角色可访问的系统功能不同，通过给用户分配不同角色，决定用户可访问的系统功能。如下所示：

功能权限管理

新增批量删除

功能权限名称: 请输入功能权限名称 功能权限编码: 请输入功能权限编码 状态: 请选择状态

查询重置

☐	功能权限名称	功能权限编码	状态	创建时间	操作
☐	测试	test	☐ 停用	2019-05-31 01:02:24	编辑 删除
☐	下载	download	☒ 启用	2019-04-29 02:02:41	编辑 删除
☐	查询	query	☒ 启用	2019-04-28 03:43:59	编辑 删除
☐	新增	add	☒ 启用	2019-04-19 06:32:06	编辑 删除
☐	上传	upload	☒ 启用	2019-04-19 04:32:48	编辑 删除
☐	删除	delete	☒ 启用	2019-04-19 04:32:25	编辑 删除
☐	编辑	update	☒ 启用	2019-04-19 04:32:15	编辑 删除

菜单管理

请输入菜单名称

新增加量删除

请输入菜单名称

全部

管理

体系

志

采集

日志

监控

☐	菜单名称	类型	请求路径	上级菜单	描述	创建时间	操作
☐		目录		-	-	2019-04-20 05:20:05	编辑 删除
☐		目录	/inModel	-	算	2019-05-07 07:31:31	编辑 删除
☐		目录	/	-	-	2019-04-20 03:03:03	编辑 删除
☐		目录		-	-	2019-04-20 05:59:59	编辑 删除
☐		目录		-	-	2019-05-07 06:06:06	编辑 删除
☐	会	目录	ju	-	A	2019-05-07 05:05:05	编辑 删除
☐	日志	目录	log	-	-	2019-05-06 06:06:06	编辑 删除
☐	采集	目录	/collec	-	-	2019-05-07 07:07:07	编辑 删除
☐	任务日志	目录	task	-	-	2019-05-07 07:21:21	编辑 删除
☐	告警	目录	/alarm	-	-	2019-05-07 11:22:22	编辑 删除

组织架构管理

请输入部门名称

新增批量删除

请输入部门名称

全部

人事部

产品中心

技术中心

☐	部门名称	上级部门	描述	创建时间	操作
☐		-	-	2019-05-13 20:56:56	编辑 删除
☐		-	-	2019-05-13 20:53:53	编辑 删除
☐	人事部	-	行政人事部	2019-04-19 19:19:19	编辑 删除
☐	产品中心	-	软件产品中心	2019-04-17 17:12:12	编辑 删除
☐		-	总经理办	2019-05-17 17:01:01	编辑 删除
☐	技术中心	-	工程技术中心	2019-05-17 17:00:00	编辑 删除

用户管理

新增

账号名称:

所属部门:

请选择所属部门

岗位:

请选择岗位

查询

重置

账号	账号名称	部门	岗位	手机号	邮箱	状态	操作
	张	开发部	前端开发	13611174	sl@huitone.com	启用	编辑 重置密码 删除
	王	开发部	前端	18721166	bw@huitone.com	启用	编辑 重置密码 删除
	李	开发部	后端开发	15811116	hd@huitone.com	启用	编辑 重置密码 删除
	赵	开发部	后端开发	15361158	qz@huitone.com	启用	编辑 重置密码 删除
	杨	交付部	测试人员	13411284	yjk@huitone.com	启用	编辑 重置密码 删除
	王	交付部	后端开发	17170208	wz@huitone.com	启用	编辑 重置密码 删除
	陈	开发部	后端开发	1554182	cl@huitone.com	启用	编辑 重置密码 删除
	陈	开发部	前端开发	13111555	ll@huitone.com	启用	编辑 重置密码 删除

岗位管理

新增

批量删除

部门:

请选择部门

岗位:

查询

重置

☐	岗位	部门	状态	创建时间	操作
☐	总	开发部	启用	2016-18:47:56	编辑 人员 删除
☐	产品经理	开发部	启用	2016-18:52:22	编辑 人员 删除
☐	前端	开发部	启用	2016-18:58:04	编辑 人员 删除
☐	后端	开发部	启用	2016-19:00:14	编辑 人员 删除
☐	大数据	交付部	启用	2016-19:02:39	编辑 人员 删除
☐	UI	交付部	启用	2016-19:04:35	编辑 人员 删除
☐	测试人员	交付部	启用	2016-19:06:06	编辑 人员 删除
☐	运维人员	交付部	启用	2016-19:07:46	编辑 人员 删除

2.2.2.2 产品特性

- 用户：访问系统的主体；
- 角色：分配权限的最小单元，通过角色给用户分配权限；
- 权限菜单：权限的最小单元，一个角色配置多个权限菜单；
- 功能权限授权

对于一个可扩展的系统来说，意味着功能是不不断变化的。为了适应这种不能事先确定的变化，必须将功能权限进行分散管理。分散管理的好处如下：

天然地支持业务的动态变化，系统实现简单。

权限的调整范围可控制在局部范围，方便权限的管理和操作。

➤ 一般权限的授予

功能权限是单维度的，可以通过简单的在功能列表或功能树上进行勾选来确定一个“角色”所允许的功能操作。然后，让相应用户成为该“角色”的“扮演者”。这样就可以把该角色所允许的功能操作授权给指定的用户了。

如果有需要更多的不同角色，那么新建角色，勾选不同的被允许的功能操作，并分别让相应的用户成为新角色的成员即可。

➤ 例外权限的授予

如果某个特定用户张三需要额外的一个权限，那么新建一个允许该功能操作的角色，并让张三成为该角色成员即可使张三拥有额外的权限。

如果某个特定用户李四需要比同一项目组的其他人少一个权限，那么新建一个拒绝该功能操作的角色，并让李四成为该角色成员即可使李四不能进行该项操作。因为在鉴权过程中，拒绝的优先级要高于允许。

● 数据权限的授权

由于数据的特殊性质，实际上在有限范围内的授权比功能权限的授予更加方便。因为生产部门和生产者具有天然的分类属性，所以象“本机构”、“本部门”、“本人”这些对生产者维度的进一步抽象就有了用武之地。

在不对业务维度做限定的情况下，就可以配置例如“允许本部门的成员管理（增删改查）本部门的数据”这样的权限规则。那么对于不同部门的用户，这条共同的规则所产生的效果并不相同，具体的效果是与用户所在的部门的业务和产生的数据相对应的。

3 数据服务平台

3.1 总体介绍

大数据服务平台为进一步推动各业务系统之间的对接，加快推进互联互通，规范统一标准化数据，以信息化促进各业务系统之间可以更好的实现信息资源共享，实现信息资源的合理化利用。

通过大数据服务对接平台，自主生成各个业务系统功能所需的数据接口 API，再通过数据接口进行对接，来对后端数据接口服务平台的服务接口 API 进行调用，实现数据流程流转，以达到能快捷、安全、有效的提供数据服务平台对接应用以及统一、便捷的服务管理，以满足多种业务系统数据及业务能力对接的需求。

系统架构图如下所示：



- 支持集群部署

API 网关由系统控制台和若干网关节点组成，节点支持水平拓展，有效避免服务器单点故障等问题，提高系统的可用性。

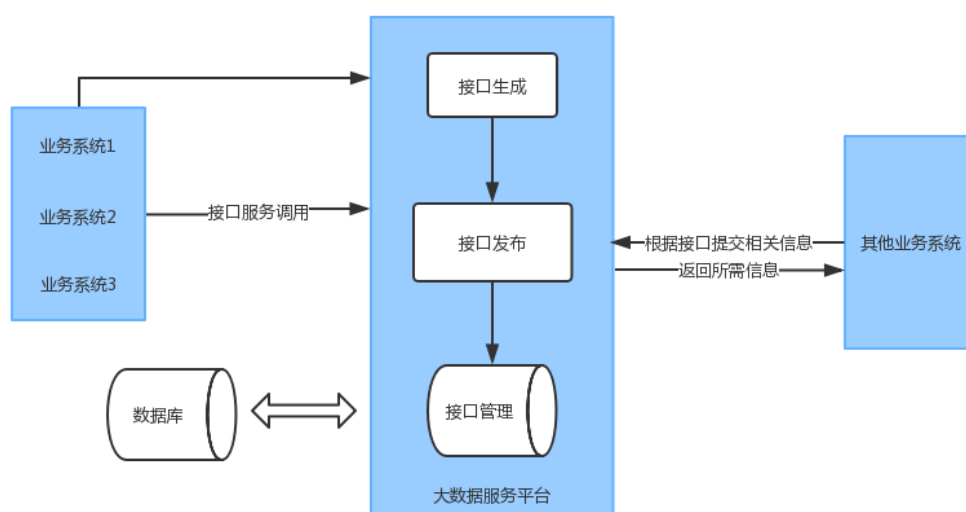
- 后端负载均衡

API 网关节点动态分配流量到后端业务服务器，降低后端服务器压力。

- 拥有独立的控制中心

API 接口注册、参数配置、API 监控等均可在管理中心进行。配置支持热更新，无需重启数据服务平台即可实时生效。

系统业务结构图如下所示：



大数据服务平台主要由接口注册、接口发布、接口管理、负载均衡、容错处理、接口监控等功能组成。

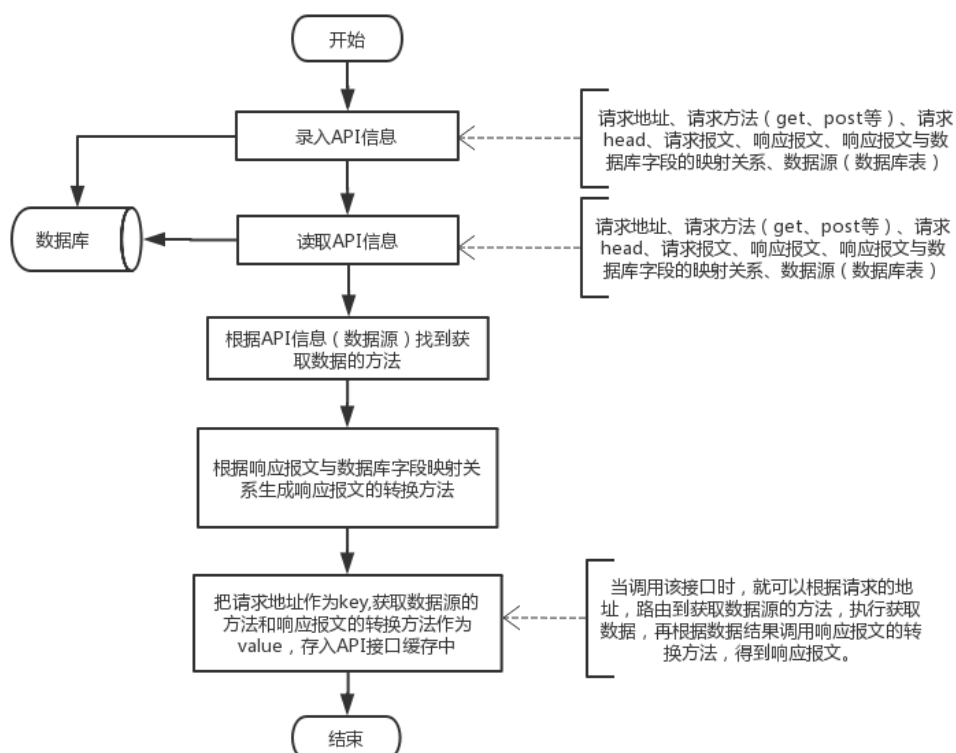
通过大数据服务平台对接服务，将各业务系统能力对接，通过大数据服务平台接口服务进行统一服务交互，同时通过数据接口自动生成服务生成所需的标准化接口，并对接口进行统一管理和监控，将接口提供给其他业务系统调用，以实现各业务系统之间的数据对接。

3.2 产品功能

3.2.1 接口注册

数据服务平台接口注册服务功能，可以智能的实现重建信息系统功能模块、数据库、原生数据接口等多种类型数据源的 API，达到各种数据资源跨系统、跨层级、跨平台的数据实时流转和数据汇聚整合，以实现业务之间协同互操作。通过接口自主注册,将各个业务系统的数据源资源整合形成标准规范化的接口 API，可自定义 API 的参数信息，根据不同的业务调用去定义所需要的 API，以实现以统一接口 API 形式提供各类数据服务。

接口注册以及调用接口的业务逻辑流程图如下所示：



(1) 页面录入注册 API 所需要的信息，包含请求地址、请求方法 (get、post

等)、请求 head、请求报文、响应报文、响应报文与数据库字段的映射关系、数据源（数据库表）。

(2) 数据服务平台读取录入的 API 信息。

(3) 根据 API 信息中的数据源，定位读取数据源的方法。

(4) 根据 API 信息中的响应报文与数据库字段的映射关系，利用反射原理生成把数据源转换成响应报文的方法。

(5) 把 API 信息中的请求地址作为 Key，把读取数据源的方法和响应报文的方法作为 Value，存入 API 接口的缓存中。

(6) 当有请求过来时，首先根据请求的地址，从 API 接口缓存中匹配到对应的获取数据源和转换响应报文的方法。

(7) 先执行获取数据源的方法，得到数据。

(8) 然后根据 API 信息中响应报文与数据库字段映射关系，执行转换响应报文的方法，得到响应报文，返回给客户端。

以上就是接口注册生成整个过程以及接口调用的请求响应过程。

3.2.1.1 主要功能

1、API 参数配置

功能描述：主要是对 API 接口生成在页面实现可配置，可对具体的数据源、请求参数等信息进行所需定义。

2、接口规范定义

功能描述：可根据不同用户的接口规范来制定一套标准化的接口规范定义，

形成数据标准化。

3、错误码配置

功能描述：可根据用户业务配置相应错误码和对应说明信息，当接口调用返回错误码时进行参考作用。

3.2.1.2 产品特性

1、配置灵活，可自用，可外放

- 支持图形化界面进行 API 参数配置，资源、渠道、数据源等灵活配置，具备灵活的数据源管理能力，可支持自对接及外放对接。

2、多种不同类型的数据源生成接口

- 支持从不同类型的源系统生成接口，可以对 B/S、C/S 架构系统、数据库、原生系统数据接口等生成数据接口 API。

3、用户接口规范标准化定义

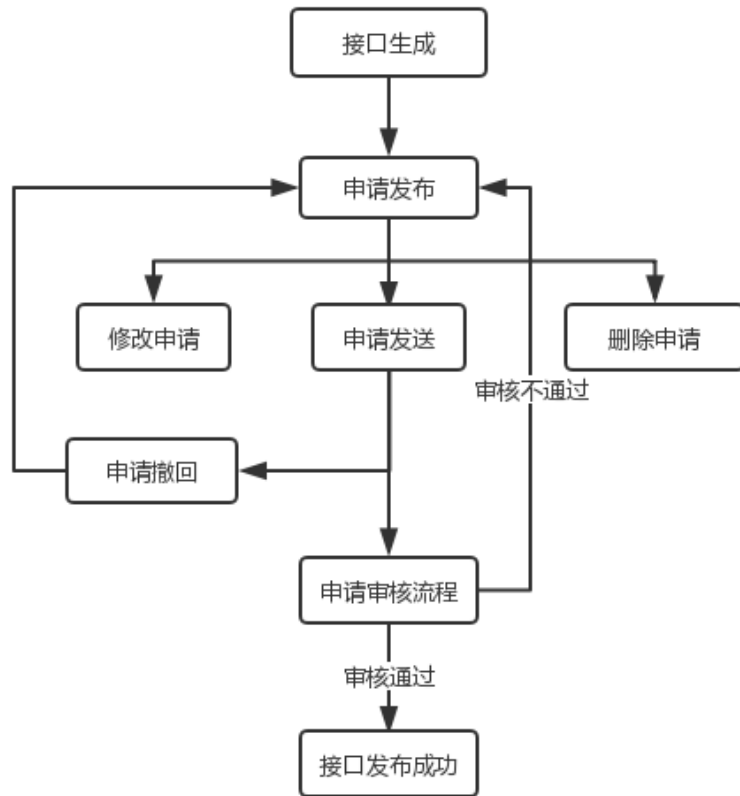
- 支持生成的数据接口可根据用户的接口规范进行标准化定义。
- 提供字段映射功能，可灵活配置数据源和目标之间的字段映射关系。
- 提供错误数据的检测盒处理，将发生错误的数据记录到相应的错误表中，过滤掉不符合规范的数据并捕获这些数据。
- 提供丰富的数据校验手段，按照用户定义的逻辑规则校验。
- 高可靠性，高并发，能快速响应，低延迟
- 数据服务平台状态正常情况下，支持数据接口的本地查询响应时间不应超过 2s，能快速返回响应结果。

- 高并发，低延时，支持亿级 API 访问量。
- 多种标准协议格式，稳定安全
- 采用 HTTP、TCP/UDP 等多种标准协议格式，保证数据稳定安全进行传输，防止恶意信息泄露。

3.2.2 接口发布

通过数据服务平台接口注册服务自主生成数据接口后,可以通过数据服务平台中接口发布功能对需要发布的接口服务提出发布申请,包括新增申请、修改申请、删除申请、查询申请、申请发送和撤回功能。通过接口实现跨不同业务系统数据协同,形成数据资源互通,有统一的数据标准 and 对接规范。

接口发布功能业务流程图如下所示:



3.2.2.1 主要功能

1、新增申请功能

功能描述：支持对自主生成的接口进行发布申请操作，申请后需进行审核操作，审核通过其他业务系统方可对接使用该接口。

2、修改申请功能

功能描述：支持对发布申请进行申请信息修改操作。

3、删除申请功能

功能描述：支持对发布申请进行删除操作。

4、查询申请功能

功能描述：支持列表显示已申请的信息，并可以从不同维度进行信息查询。

5、申请发送和撤回功能

功能描述：新增申请后，可支持申请发送和撤回操作，申请发送则会进入审核环节，也可以在未审核之前进行撤回操作。

3.2.2.2 产品特性

1、接口申请发布全流程管理

- 接口发布全流程管理，从申请发布到接口审核通过，严格把控发布接口整个过程，避免出现错误。

2、多种功能操作集合

- 支持对发布申请的多种操作，如修改、删除、发送、撤回等，人性化管理发布申请信息。

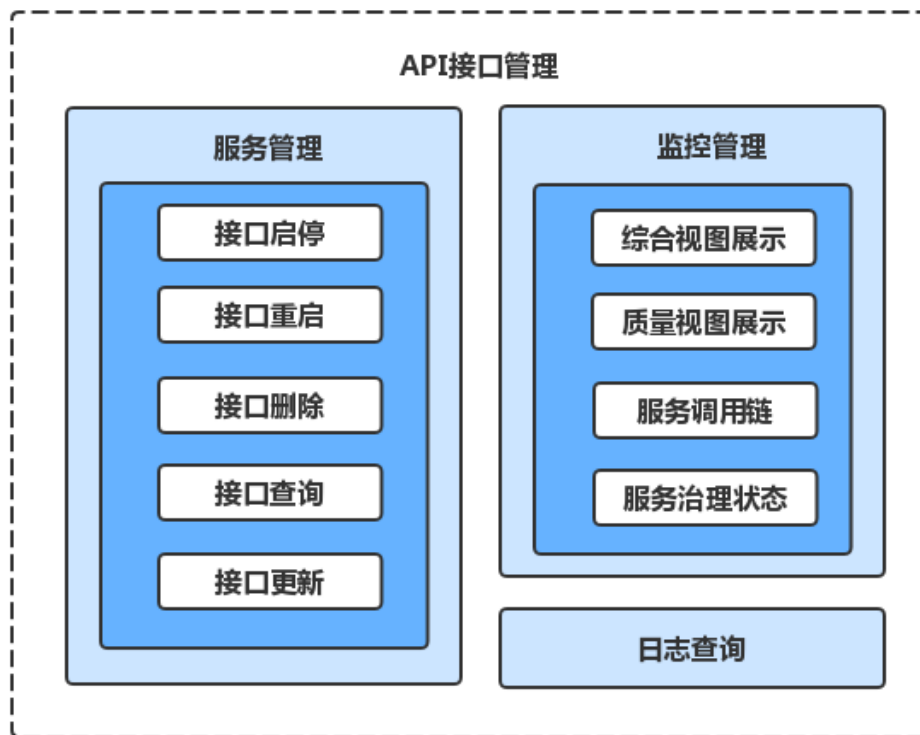
3、多维度查询信息

- 页面支持从不同维度进行接口发布信息查询，个性化配置搜索查询条件，以便能及时查找到所需信息。

3.2.3 接口管理

接口注册生成和申请发布通过后，需对数据服务平台的接口进行统一管理，并在此基础上提供接口平台运行监控管理功能，达到对数据服务平台接口对接的整个生命周期管理。

模块具体的功能结构图如下所示：



3.2.3.1 主要功能

1、接口启停、重启功能

功能描述：支持对接口进行启动和停止操作。启动接口后，该接口便可对外提供服务使用，停止该接口后，该接口就不能继续对外提供使用，且在接口有效使用期内，可通过重启操作继续使用该接口的功能。

2、接口删除功能

功能描述：支持对已有的接口信息进行删除操作。

3、接口查询、查看功能

功能描述：支持通过多维度条件输入查询接口信息，并可以查看 API 接口的详细信息，包括接口各种参数信息、响应信息等，以及运行情况，以便及时发现

接口服务是否运行异常。

4、接口更新功能

功能描述：当 API 接口服务参数及映射等信息发生更新后，可通过更新操作及时获取到最新的接口信息。

5、日志查询

功能描述：可以查看到使用接口的具体信息，如请求参数、成功、失败数、请求频率、异常上报日志等来实时观测该接口是否正常使用。

6、综合视图

功能描述：展示了 API 接口平台的总请求量，以及调用量最大的接口和应用组信息等，可以直观清楚的看到 API 接口的使用情况。

7、质量视图

功能描述：展示了 API 接口平台接口的成功率占比趋势走向，以及接口延时占比趋势走向，可以清楚的看到接口使用质量情况。

8、服务调用链

功能描述：搭建服务调用链，显示服务之间的调用关系，以便可以清晰的对服务的整个调用链路各节点有完善的监控。

9、服务治理状态

功能描述：通过服务治理状态，显示当前接口服务注册情况、服务状态、熔断等。

3.2.3.2 产品特性

1、接口界面可视化，易维护

- API 接口管理系统化或平台化，可以直接在可视化 API 管理界面上方便的维护接口。

2、接口服务多功能使用，操作方便

- 支持对接口服务启动、停止、重启、删除等进行运行控制，方便对接口服务进行操作，实时响应。

3、全图形化操作

- 丰富的图形化设计和监控信息展示，支持页面显示接口的全生命周期管理、运行时状态监控。

4、实时监控

- 提供强大的 API 接口监控功能，实时了解 API 运行状态，一旦发现 API 或节点异常则产生告警并及时通知运维人员，减少由于服务不可用导致的损失。

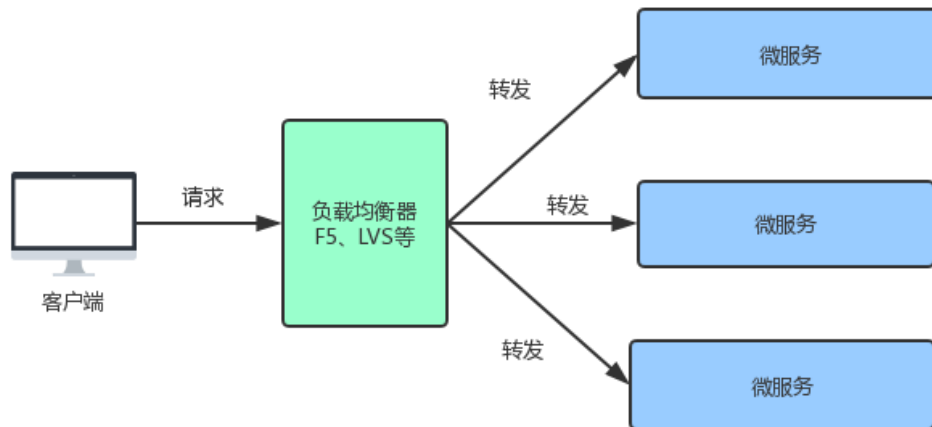
5、降低成本，提高系统灵活性

- 提供 API 服务治理与 API 性能指标监控，降低 IT 运维的成本，提高系统灵活性。
- 安全可靠的将业务系统通过微服务架构进行连接，高效便捷。

3.2.4 负载均衡

负载均衡是微服务架构中必须使用的技术，通过负载均衡来实现系统的高可

用、集群扩容等功能。负载均衡可通过硬件设备及软件来实现，硬件比如：F5、Array 等，软件比如：LVS、Nginx 等。负载均衡架构图如下：

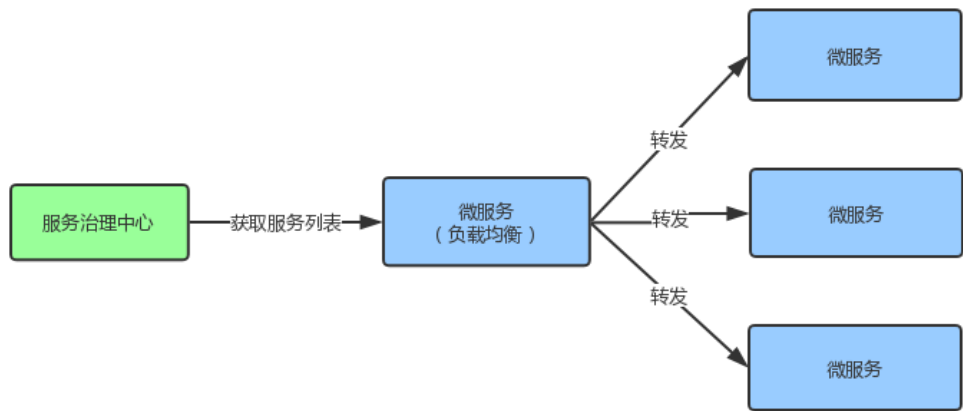


当客户端发送请求时，会先到达负载均衡器（也相当于一个服务），负载均衡器根据负载均衡算法（轮训、随机、加权轮训）从可用的服务端列表中取出一台服务端的地址，并将请求转发到相应的微服务。通过硬件或软件实现负载均衡均会维护一个服务端清单，利用心跳检测等手段进行清单维护，保证清单中都是可以正常访问的服务节点。所以负载均衡可以为微服务集群分担请求，降低系统的压力。

而对于客户端负载均衡来说，与服务端负载均衡的主要区分点在于服务清单的存放位置。在客户端负载均衡中，客户端自己会存储一份服务端清单，它是通过从注册中心进行抓取得到的，同时也需要对此进行维护。

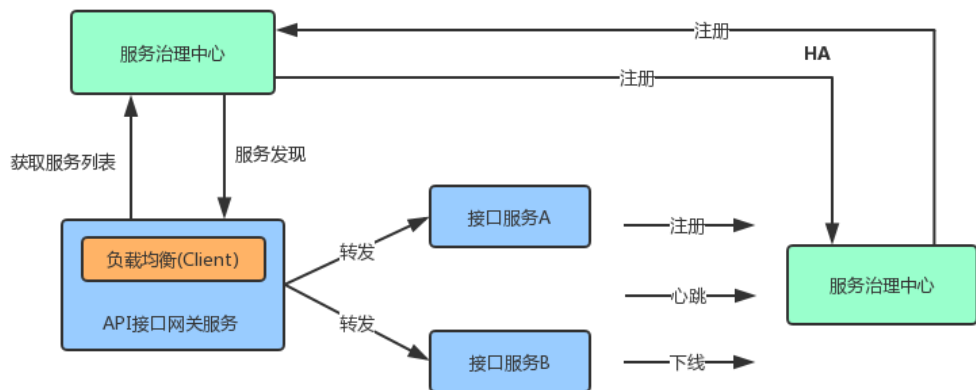
微服务负载均衡是基于客户端的负载均衡工具，它可以将面向服务的 REST 模板请求自动转换成客户端负载均衡的服务调用。负载均衡会从服务治理中心获取服务列表，然后根据负载均衡算法直接请求到具体的微服务，中间省去了负载

均衡服务。负载均衡的流程图如下所示：



- (1) 在微服务中使用负载均衡，它会先从服务治理中心获取服务列表。
- (2) 然后根据负载均衡的算法进行负载均衡，将请求路由转发到其它微服务进行处理。

大数据服务平台 API 接口网关主要使用了负载均衡为客户端提供负载均衡功能,根据算法将请求转发给相应的接口服务。具体的负载均衡流程图如下所示：



3.2.4.1 主要功能

1、获取服务

功能描述:从服务治理中心获取服务信息,根据策略规则去路由,并根据 Ping 判断服务的可用性。

2、服务一致性

功能描述:负载均衡器具有提供服务一致性的功能,通过读取客户端所发出请求内的信息,进行重写报头程序,然后将请求发送至合适的服务器上,该服务器会维护着该客户端信息。

3、选择服务器

功能描述:当服务集群中的某个节点无法成功处理请求时,那么该请求就会被发往其他节点上去,当请求成功发至另一节点后,原节点上的请求信息就会自动消失了。

4、统计计量

功能描述:由于所有的客户端请求都会先经过负载均衡器,所以负载均衡器具有统计计量的功能,利用该功能负载均衡器就可以准确统计出各阶段各种流量流动以及各项目进行的次数等,从而网络就可以适当调整系统性能。

3.2.4.2 产品特性

1、高并发,高性能,高可用

- 解决并发压力。
- 提高应用处理性能,增加吞吐量,加强网络处理能力。

- 提供故障转移，实现高可用，当集群里的 1 台或者多台服务器 down 掉的时候，剩余的没有 down 掉的服务器可以保证服务的继续使用。
- 使用了更多的机器保证了机器的良性使用，不会由于某一高峰时刻导致系统 CPU 急剧上升。

2、扩展性强，可伸缩性

- 通过添加或减少服务器数量，提供网站伸缩性（扩展性）。

3、安全防护机制

- 负载均衡设备上做一些过滤，黑白名单等处理。

4、优化访问请求分配，提升性能

- 负载均衡器优化了访问请求在服务器组之间的分配。
- 消除了服务器之间的负载不平衡，从而提高了系统的反应速度与总体性能。

5、实时监控运行状态，灵活性强，扩展能力高

- 负载均衡器可以对服务器的运行状况进行监控，及时发现运行异常的服务器，并将访问请求转移到其它可以正常工作的服务器上，从而提高服务器组的可靠性。
- 采用负载均衡器后，可以根据业务量的发展情况灵活增加服务器，系统的扩展能力得到提高，同时简化了管理。

3.2.5 路由管理

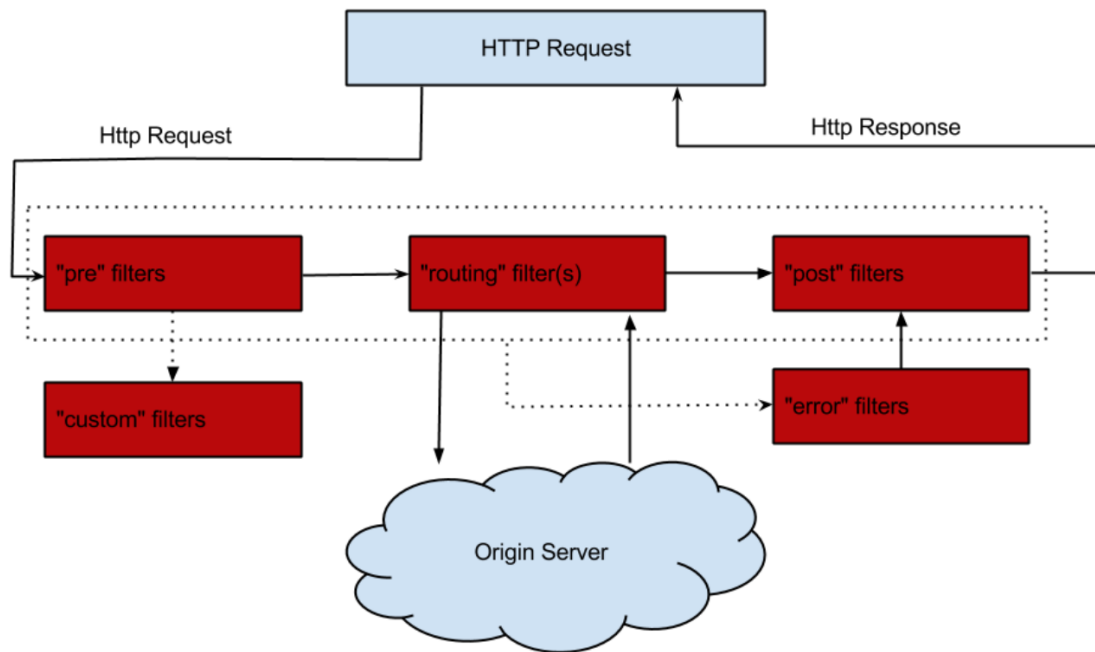
API 服务网关作为微服务系统的网关组件，用于构建边界服务，致力于动态

路由、过滤、监控、弹性伸缩和安全。负载均衡以及服务治理结合可以实现智能路由和负载均衡的功能；网关将所有服务的 API 接口统一聚合，统一对外暴露。外界调用 API 接口时，不需要知道微服务系统中各服务相互调用的复杂性，保护了内部微服务单元的 API 接口；网关可以做用户身份认证和权限认证，防止非法请求操作 API 接口；网关可以实现监控功能，实时日志输出，对请求进行记录；网关可以实现流量监控，在高流量的情况下，对服务降级；API 接口从内部服务分离出来，方便做测试。

API 服务网关和服务治理进行整合，将 API 服务网关自身注册为服务治理下的应用，同时从注册中心中获得其他微服务的消息，也即以后的访问微服务都是通过 API 服务网关跳转之后获得。

API 服务网关通过 Servlet 来实现,通过自定义的 Servlet 来对请求进行控制。核心是一系列过滤器，可以在 Http 请求的发起和响应返回期间执行一系列过滤器。API 服务网关采取了动态读取、编译和运行这些过滤器。过滤器之间不能直接通信，而是通过 RequestContext 对象来共享数据，每个请求都会创建一个 RequestContext 对象。

API 服务网关包含了对请求的路由和过滤两个最主要的功能。其中，路由功能负责将外部请求转发到具体的微服务实例上，是实现外部访问统一入口的基础。而过滤功能则是负责对请求的处理过程进行干预，是实现请求校验、服务聚合等功能的基础。Zuul 生命周期如下图所示：



当一个客户端 Request 请求进入 Zuul 网关服务时，网关先进入“pre filter”，进行一系列的验证、操作或者判断。然后交给“routing filter”进行路由转发，转发到具体的服务实例进行逻辑处理、返回数据。当具体的服务处理完成后，最后由“post filter”进行处理，该类型的处理器处理完成之后，将 Request 信息返回客户端。

3.2.5.1 主要功能

API 服务网关主要包含了对请求的路由和过滤两个核心功能，它的核心是一系列过滤器，通过加载动态过滤机制，从而实现以下各项功能：

1、身份认证与安全

功能描述：识别面向各类资源的验证要求，并拒绝那些与要求不符的请求。

2、过滤机制

功能描述：负责对请求的处理过程进行干预，是实现请求校验、服务聚合等

功能的基础。

3、动态路由

功能描述：以动态方式根据需要将请求路由至不同后端集群处，负责将外部请求转发到具体的微服务实例上，是实现外部访问统一入口的基础。

4、负载分配

功能描述：为每一种负载类型分配对应容量，并弃用超出限定值的请求。

3.2.5.2 产品特性

1、审核与监控

- 在边缘位置追踪有意义的数据和统计结果，从而带来精确的生产视图。

2、静态响应处理

- 在边缘位置直接建立部分响应，从而避免转发到内部集群。

3、多区域弹性

- 跨越 AWS Region 进行请求路由，实现 ELB 使用多样化，让系统边缘更贴近使用者。

4、压力测试

- 逐渐增加指向集群的流量，以了解性能。

3.2.6 故障隔离

微服务架构通过定义明确的服务边界，能有效地隔离故障。和其他分布式系统一样，微服务在网络、硬件和应用层上都会存在更多的问题。由于服务之间

是互相依赖，因此任何组件都可能出错导致用户不能访问。为尽可能减少部分中断带来的影响，我们需要构建容错能力强的服务，以从容应对发生的某些中断。

微服务架构将应用逻辑拆分成服务，服务之间通过网络交互。由于是通过网络调用，而不是在进程中调用，因此这给需要在多个物理和逻辑组件间进行协作的系统带来了潜在的问题和复杂性。分布式系统变得越来越复杂，也导致网络特定故障发生的可能性增大。

在微服务架构中，服务是互相依赖的。这就是为什么需要故障隔离并且尽可能降低它们的负面影响。

3.2.6.1 主要功能

1、服务降级

功能描述：当组件出现故障时，能隔离这些故障并且能做到优雅地服务降级。

2、超时

功能描述：在服务之间调用时，设置一个主动超时时间，超过了这个时间阈值后，如果“被依赖的服务”还没有返回数据的话，“调用者”就主动放弃，防止因“被依赖的服务”的故障所影响。

3、限流

功能描述：限制最大流量，系统能提供的最大并发有限，请求又太多，服务不过来的时候就会进行请求排队限流。

4、延迟处理

功能描述：设置一个流量缓冲池，所有的请求先进入这个缓冲池等待处理，

真正的服务处理方按顺序从这个缓冲池中取出请求依次处理,这种方式可以减轻后端服务的压力。

5、熔断

功能描述: 当流量过大或者错误率过大的时候,自动断开链路,不对外提供服务。当流量恢复正常,或者后端服务稳定后,就会自动链接上,正常提供服务。

3.2.6.2 产品特性

1、优雅的服务降级

- 当组件出现故障时,能隔离这些故障并且能做到优雅地服务降级。可将不是那么重要的功能模块进行降级处理,停止服务,这样可以释放出更多的资源供给核心功能的去用。
- 可以对用户分层处理,优先处理重要用户的请求。

2、自我修复能力

- 自我修复能帮助恢复应用。在大多数情况下,是通过外部系统监控实例的状态,当服务出现故障一段时间后则会重启服务。在大多数情况下,自我修复的功能是相当有用的。

3、隔离模式

- 这种模式可以对不同类型的请求使用线程池来进行资源隔离,每种类型的请求互不影响,如果一种类型的请求线程资源耗尽,则对后续的该类型请求直接返回,不再调用后续资源。这种模式使用场景非常多。

4、故障转移缓存

- 因网络问题和系统中的变更，服务通常会出现故障，而故障中断大多是暂时的，会进行自我修复。故障转移缓存能使服务即使在出现故障的时候也能工作，能帮助我们的应用提供必需的数据。
- 保障数据持续的传输。

3.3 功能展示

注册 API 生成接口，需要进行数据资源的数据源信息接口对接，对接所需平台服务的 API 接口，填写基本信息后，点击下一步，进行请求示例的填写，先根据需要进行选择不同的请求方式（GET/POST），然后填写入参和出参示例，最后需要进行错误码填写，以便当请求响应出错时可快速知道请求出错原因。如图所示：

注册API

×

基本信息

请求示例

错误码

* API来自平台/服务:

请选择

找不到你需要的服务?点击[创建服务](#)

* API名称:

请输入API名称

* API网关地址:

platform/service/

请使用驼峰命名

* 使用API服务路由:

☐ 是

☒ 否

API接口地址:

http://

请输入合法URL

* API测试类型:

☒ 伪接口

☐ 测试接口

* API开发者:

* API描述:

请输入API描述

* 是否IP鉴权:

☐ 是

☒ 否

取消

下一步

基本信息

请求示例

错误码

* 请求方式:

POST

✓

* POST请求内容类型:

application/json

✓

入参示例:

提示: 若是解析失败, 请到[json.cn](#)确认JSON对象是否填写正确

入参说明会根据入参示例进行解析

参数名	类型	是否必须	中文说明
暂无数据			

* 出参示例:

```
{
  "code": "0",
  "msg": "success",
  "data": []
}
```

✓

上一步

下一步

基本信息

请求示例

错误码

错误码:

+ 添加

错误码	说明	操作
1	1	删除

上一步

提交

API 接口信息填写完成后，提交生成所需的接口。

提供 API 接口管理信息界面化展示，通过此界面可对 API 接口进行多维度查询、条件重置、导入、启动、停止、重启、更新等操作。如图所示：

40



查看 API 接口详细信息。如图所示：

- 基本信息：



- API 返回码：

> 当前API返回码:

错误码	说明
1	111111111111

- 请求参数：

> 请求参数:

名称	类型	必填	中文说明
currentPage	number	是	
pageCount	number	是	
searchText	string	是	
advanceSearchText/authForUser	布尔值	是	

● 响应参数:

> 响应参数:

名称	类型	中文说明
code	数字	API平台统一出参字段（API平台自动解析转换）（返回码）
msg	字符串	API平台统一出参字段（API平台自动解析转换）（返回消息）
data	对象	API平台统一出参字段（API平台自动解析转换）（返回数据）
currentPage	数字	
totalCount	数字	
moduleList/level1Name	字符串	
moduleList/level2Name	字符串	
moduleList/level3Name	字符串	
moduleList/description	字符串	

● 请求示例:

> 请求示例:

线上服务地址	
开发测试地址	
入参示例	keyword=xxx
请求方式	GET
返回格式	JSON
出参示例	<pre>{ "code": 0, "msg": "msg", "data": [{ "userId": 1, "userType": 1, "userName": "weedustzhao", "nickName": "张三" }, { "userId": 2, "userType": 1, "userName": "zachhuang", "nickName": "张三" }] }</pre>

● 错误码参照信息:

错误码	说明
5000	系统错误（平台代码执行出现异常）
5001	平台依赖接口响应异常
4003	没有权限
4004	API地址错误或请求方式非法
4007	apikey非法
4008	缺少参数
4009	入参json解析错误
4010	请求源非法
4011	{0}错误

日志查询，可以查看到操作使用接口的具体信息，包括请求响应信息、耗时等等信息。如下图所示：

日志查询

平台:

服务:

接口名:

接口UR:

最小耗:

最大耗:

时间范: ~

业务返:

平台返:

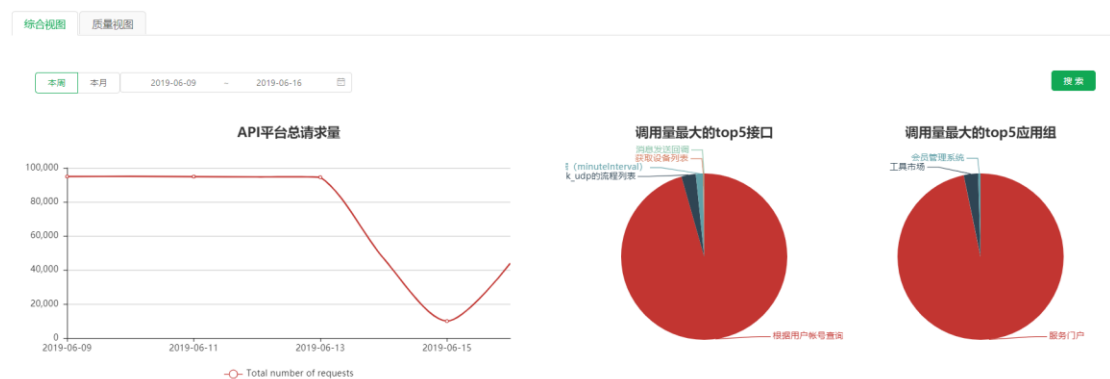
时间	平台	服务	接口	url	业务返回码	平台返回码	耗时(ms)	接口负责人	应用组名	应用负责人	操作
 暂无数据											

▼ 接口信息			
平台: 统一登录	服务: 用户信息查询	接口: 根据用户帐号查询用户信息	url: http://wang.oa.com/unified/userInfo?byAccount&account=zyBackendAdmin
▼ 服务调用方			
应用组名: 服务门户	应用组负责人: robinhuang	来源IP: 172.25.0.14	
▼ 服务提供方			
接口负责人: uckyly			
▼ 调用日志			
业务返回码: 0	平台返回码: 0	耗时(ms): 1812	时间: 2019-06-16 10:06:26
▼ 请求参数		▼ 响应参数	
header	DefaultHttpHeaders[X-Forwarded-For: 172.25.0.14, clientip: 172.25.0.14, User-Agent: GuzzleHttp/6.2.1 curl/7.47.1, PHP/7.1.1, Uid: 0, AppId: 0, content-length: 0, GUID: 0506880413233599, Request-Id: HH05d05a421db1f, Api-Chain: 0-0506880413233599, GW-Token: , connection: close, host: api.lo.gin.zhiyun.qcloud.com]		
body			
▼ 异常上报日志			
序号	异常信息	异常日志/异常堆栈信息	
暂无数据			

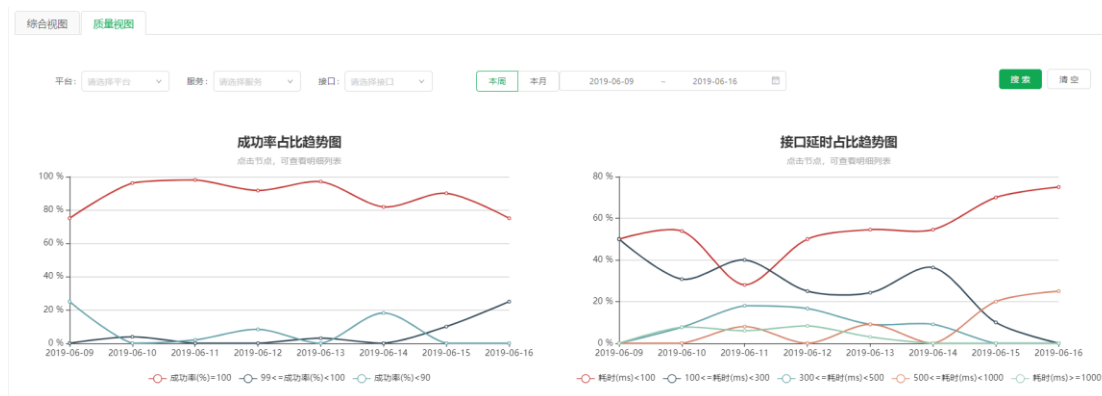
监控管理主要是对 API 接口平台的监控信息作一个综合视图和质量视图的

展示，如下图所示：

● 综合视图



● 质量视图



4 数据集成平台

4.1 总体介绍

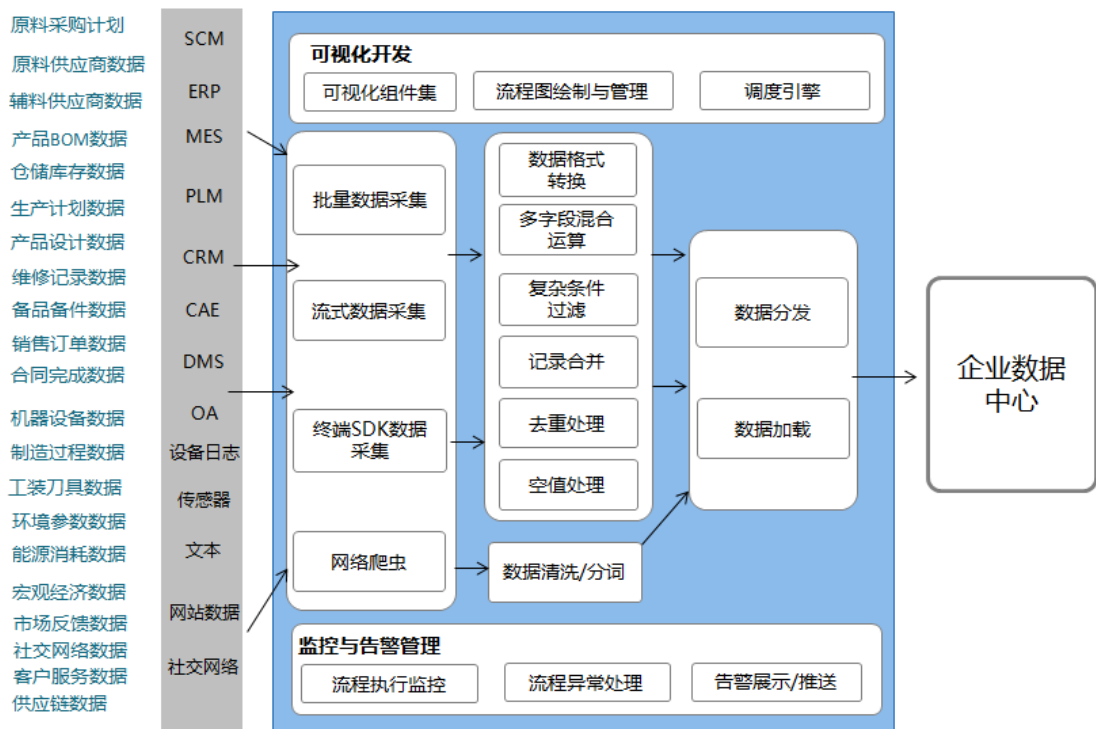
针对企业多源异构数据的集成与处理，汇通提供数据集成平台。

它支持实时、批量、日志等多种数据采集方式，拥有多种系统和设备的开发工具套件，能够快速对接各类源数据；对采集到的多种数据源进行经过抽取、转换、加载，创建多种的任务调度，以满足不同的数据需求，并且能够根据企业的需求快速扩展。



汇通国信数据集成平台架构

数据流图如下所示：



数据处理流图

产品特点：

1. **多源异构数据整合能力**：提供百种数据采集接口，支持各种主流的关系型数据库、第三方云平台、文本文件、HDFS 文件等；可快速对接和集成数据，能满足企业在复杂业务场景下的各类数据整合要求。

2. **灵活的流式数据处理引擎**：支持 20 多种的流式数据源，自定义触发任务条件，在实时传输过程中进行数据处理。

3. **全图形化操作**：丰富的图形化界面设计和监控数据采集过程执行状态。

4. **灵活的数据清洗规则配置**：支持记录级的多字段的混合运算、过滤、去重、清洗、复杂条件过滤等，实现数据的标准化。

5. **统一的作业调度及监控功能**：提供多种任务逻辑的组合处理及配置。

4.2 产品功能

4.2.1 数据采集转换与加载

产品具备从不同数据源 (RDBMS、Hadoop、MPP 等) 进行数据采集与处理，包括批量采集处理、流式采集处理等。支持格式转换、数据核查功能，包括记录级的多字段的混合运算、过滤、去重、清洗、复杂条件过滤等，实现数据的标准化。

4.2.1.1 主要功能

功能	功能细项	功能描述
数据库导入	数据库导入	数据库到 Hive、HDFS, 包含: MySql、Oracle、DB2、SqlServer 等常见数据库

	Hive 到 Hive	支持常见数据库互导以及导入到 HDFS 和 Hive
本地文件导入	本地文件	本地文件到 Hive、HDFS
	Excel、CSV	Excel、CSV 到 Hive、HDFS
	客户端工具导入	客户端导本地路径文件源、客户端导数据库数据源
	dump 文件	数据上传到 Hive、HDFS、DB
大数据类数据库导入	大数据	Redis、HBase、Impala、MongoDB
Http 流式上传	Http 流式上传	提供 http 流式上传方式, 开放上传接口, 上传到 kafka
消息队列服务	Kafka	
其他类型导入	FTP	提供 FTP 上传到 HDFS
资源库	创建数据库连接	创建资源库连接信息并保存, 在之后的上传过程中, 可以选择已经保存过的数据库, 自动进行连接选择, 不用再输入信息, 方便操作
	管理数据库信息	对保存过的资源库连接进行列表展示, 并可以对资源库信息进行管理操作
迁移任务	迁移任务管理	展示所创建的任务信息及任务运行情况, 并可对任务进行管理

	执行历史记录	查看所有数据迁移任务的执行历史和日志 (比如某条任务是每小时执行一次, 就会产生多条执行历史)
--	--------	--

4.2.1.2 产品特性

1. 提供几十种数据采集接口, 包括: 各种主流的关系型数据库如 mysql、oracle 等、第三方云平台、文本文件、HDFS 文件等。
2. 支持文件批量采集, 数据库迁移同一个任务最大支持 1000 张表同时导入。
3. 具备全量和增量数据采集, 允许灵活定义多种抽取策略, 允许对抽取的字段进行动态选择。
4. 具备灵活的数据源管理能力, 允许以统一方式进行配置并获取访问。
5. 客户端自动同步本地资源, 对本地文件数据实现轻松管理, 告别频繁上传、文件管理不再愁。
6. 提供图形化的界面来实现灵活的数据清洗规则配置, 支持记录级的多字段的混合运算、过滤、去重、清洗、复杂条件过滤等, 实现数据的标准化。
7. 提供字段映射功能, 可灵活配置迁移的源和目标之间的字段映射关系。
8. 提供错误数据的检测和处理, 将发生错误的记录到响应的错误表中, 过滤不符合规范的数据并且捕获这些数据。
9. 提供丰富的数据校验手段, 按照用户定义的逻辑规则校验等。

4.2.2 流程开发和任务调度

使用任务调度系统(Task Scheduler)提供的功能,用户可以通过拖拽点击等操作及可完成 ETL 流程的设计和参数的配置,无需繁琐的编码过程;支持多种任务管理设置;支持重跑流程、停用流程等。

4.2.2.1 主要功能

功能	功能子项	功能描述
流程管理	流程管理	流程的新建、复制、删除、修改、启用、停止、查询, 定义任务调度策略
实例管理	实例查询	快速检索查询当前平台的所有流程实例、流程执行实例依赖关系图形化、实例执行流程图查看
	实例管理	重跑、补跑、任务重试、终止
	调度历史	实例相关流程调度历史时长图形化与列表两种方式展示

4.2.2.2 产品特性

1. 通过图形化拖拽方式形成工作流程, 无需命令行入口, 上手简单。
2. 系统抽象多种大数据计算组件, 全面支持和管理各种大数据业务场景。系统抽象出丰富的组件, 包括: 保存、线条、开始、结束、分支、合并、依赖、导数、Hdfs、M/R、Hive、Sparksubmit、SparkSql、Mysql、Oracle、Shell、Kettle、

DB2、postgresql、sqlserver、http、java、sparkpy、sparkR、SparkMpp、存储过程、通知等工具，并可图形化添加多种高质量组件，支持扩展。

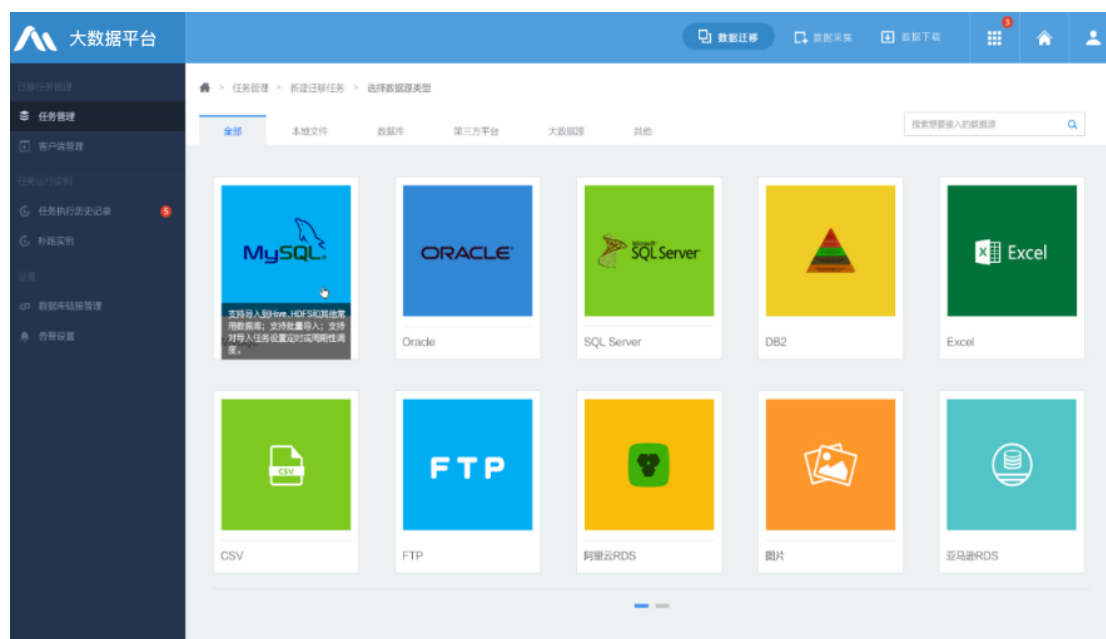
3. 支持多种的任务调度管理

支持多种任务管理设置；支持重跑流程、停用流程；支持流程提交历史调度时间点流程实例重新生成并调用。

4. 丰富的图形化界面监控数据采集处理过程状态。

4.3 功能展示

提供几十种数据源的导入导出，包括：各种主流的关系型数据库如 Mysql、Oracle 等、第三方云平台、文本文件、HDFS 文件等。



具备灵活的数据源管理能力，可以增加、修改资源库信息：



图、灵活的数据源管理能力

支持数据的批量迁移，包含配置源和目标、数据表配置、导入规则设置等步骤，即可执行。



图、数据的批量迁移

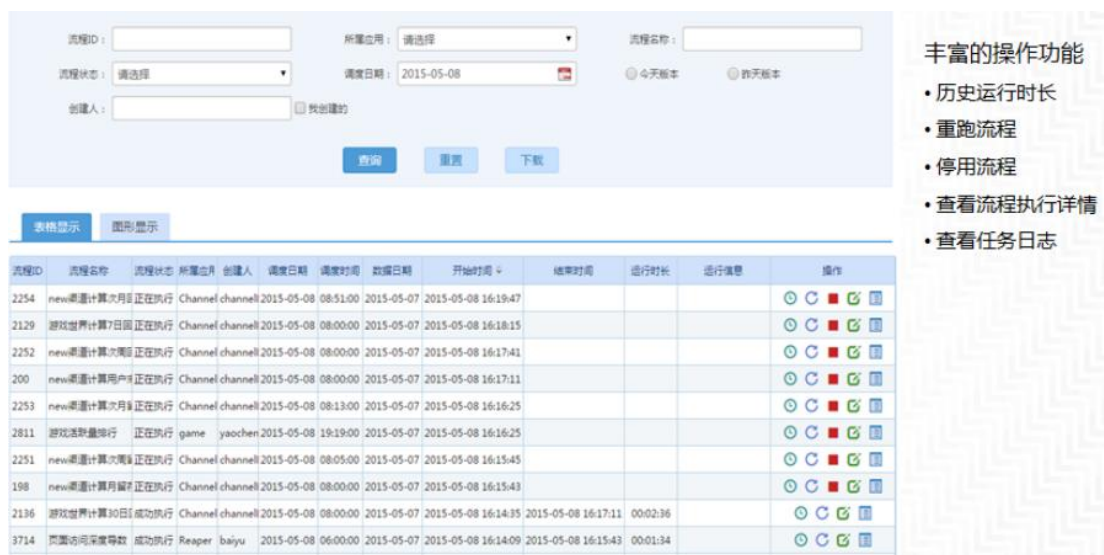
灵活的数据清洗规则配置界面，支持记录级的多字段的混合运算、过滤、去重、清洗、复杂条件过滤等。



支持多种告警规则、告警方式设置：



对 ETL 作业各个步骤的运行情况进行监控并显示监控信息：



图、作业运行情况监控

4.4 应用场景

可应用于多源、异构数据的整合场景，为后续数据中心/数据集市建设奠定数据基础。

例如：某企业，已逐步实施某些企业信息业务系统，较好地支持了研发、生产制造、销售、客户服务等各个环节的运营使用，取得了丰富的成果，积累了大量数据。但是随着业务应用的不断发展，发现存在企业内系统数据孤岛现象较严重，数据各自割裂；数据没有很好地集成，无法实现共享；数据口径的不一致，数据可信度较差，容易造成生产能力的浪费、库存成本的增加、运营效率的降低等现象。



为了实现企业内部分散数据和外部数据的融合，快速发掘隐藏在数据背后的巨大商业价值。企业通过数据集成工具 所提供的百种数据采集接口适配能力，快速与相关主流的关系型数据库、第三方云平台、文本文件、HDFS 文件等进行对接，实现生产数据、销售数据、供应上数据、设备数据、工厂数据、网站数据等采集转换。对于异构数据深度清洗后，自动构建统一的元数据管理，构建大型异构数据统一数据中心。通过深度机器学习，提供数据治理的自动化异构数据整合方案，减少不必要的重复开发工作，缩短企业数据的集成时间。



数据采集清洗转换后，再将数据加载到企业数据中心，为后续数据的进一步处理和分析应用奠定了良好基础。

5 数据计算存储平台

5.1 总体介绍

面对海量的数据，政务服务系统经常面临的困难和挑战是如何快速实现海量数据稳定可靠的存储与管理？如何为多样的应用场景提供不同的分析引擎？如何更方便易用地管理大数据平台？

为了应对这些困难和挑战，汇通国信提供大数据计算平台来帮助客户应对挑战。汇通国信大数据计算平台是整个大数据存储处理和分析的基础平台。它基于 Hadoop/Spark 生态系统，对复杂开源技术进行高度集成和性能优化，自研 10 余种实用工具。在分布式存储系统的基础上，建立了统一资源调度管理系统，深度优化大规模批处理、交互式查询计算、流式计算等多种计算引擎，具备业界超高性能；同时具备良好的横向扩展能力。平台具有使用简便、运行高效、易于扩展、安全可靠等特点。



图、汇通国信大数据计算平台功能架构

主要功能：

功能	内容描述
----	------

大数据平台核心组件	
分布式文件存储	分布式文件存储、多副本备份与同步机制，提供容错机制，可修改副本策略；
	大文件写入、流式数据访问、高吞吐量数据访问；
	支持数据存储分布策略，支持机架感知与负载均衡，支持高可用；
NoSQL 数据库	分布式、列存储、多维结构存储，支持结构化和非结构化大数据量的高速读写操作；
	面向列表（簇）的存储和权限控制，列（簇）独立检索，以及二级索引，支持数据多版本；
	面向列的数据压缩，高压缩比，有效降低磁盘 I/O
数据仓库工具	海量结构数据批量离线分析
	提供基于 HQL 的数据查询机制，支持 UDF，自定义存储格式，扩展数据类型，函数和脚本
批量计算框架	数据划分和计算任务调度；
内存计算引擎框架	分布式内存计算引擎
流数据计算引擎	基于 Storm 与 Spark Streaming 的流式计算引擎；
分布式数据库 MPP	支持基于 Spark 的 MPP 架构数据库，基于 Spark 扩展 CRUD 操作；
多维分析引擎	提供数据分析能力，支持 SQL 查询
分布式消息队列	支持消息队列的负载均衡、分区存储、数据压缩等

服务	
分布式协作服务	配置管理、配置更新通知、节点主备容灾、节点心跳管理等；
统一资源调度	支持资源封装、调度、隔离以及配额管理
	支持 Capacity(静态)、FIFO(先进先出)、Fair(公平、动态)等调度模式；
交互式分析引擎	支持基于 SQL 的查询分析；支持基于 JDBC/ODBC 的数据库连接，支持 BI 可视化工具连接
数据导入导出	支持传统数据库到 Hadoop；支持 Hadoop 到传统数据库；
全文搜索引擎	基于 Lucene 的全文搜索服务器；
日志采集服务	分布式、可靠的日志采集服务
日志分析服务	提供一个分布式多用户能力的全文搜索引擎；支持日志搜集处理框架、快速的日志综合处理能力；支持日志搜索、可视化、分析能力
缓存服务	基于 Key-value 的数据缓存库，支持数据同步；
安全保障	支持 Kerberos 认证和系统集成
任务调度	
实例管理	快速检索查询当前平台的所有流程实例、流程执行实例依赖关系图形化、实例执行流程图查看；
	重跑、补跑、任务重试、终止；

	实例相关流程调度历史时长图形化与列表两种方式展示；
配置管理	数据库等资源的连接配置；
	依赖的 hadoop/hdfs 等相关的配置；
	支持短信/邮箱告警服务，如邮箱配置、SMS 配置，告警短信配置；
内存计算功能增强	
Spark SQL 功能增强	SQL 语 法 兼 容 能 力 增 强 ； 支 持 SQL 99/2003 标 准 ， 兼 容 SqlServer/Oracle 语法，支持 TPC-DS 测试集 99 个 SQL 语句；
	支持 PL/SQL 语法，支持存储过程、
	查询结果可以通过交叉表进行进过展示，默认显示前 100 行
事务处理能力增强	基于 Spark 扩展数据增删改 CRUD 操作
用户管理	
用户管理	用户及账户的添加、修改、删除，用户启停用；
项目管理 - 多租户管理	提供项目管理、人员分配、权限管理等功能
费用管理	当前计算、存储资源消耗费用计算，并可查看详情；
资源管理	当前租户下各项目资源使用情况，对项目进行资源池分配；

个人中心	密码修改、用户注销；
运维监控	
安装部署	安装文件拷贝、环境检测与主机环境配置、组件自动化部署；
集群监控	指标监控、监控热图、历史配置信息、版本信息；
服务管理	添加与删除服务、服务启停、部署与移动；
	参数配置、配置组、历史版本；
	HA 配置，支持全组件的 HA 配置，包括 Manager 管理节点；
主机管理	添加与删除主机节点、主机监控指标、主机及相关组件的告警信息；
告警管理	报警历史记录；
	告警组、告警通知；
版本管理	平台及各个组件版本管理、版本升级；
	平台授权信息注册；
用户管理 - 运维	用户添加、修改、删除；
用户	角色添加、修改、删除；
日志管理	根据检索内容做简单的信息统计，统计不同类型输出信息数量；
	根据组件、关键字信息做信息搜索；

产品整体特性：

1. 深度优化与自研：深度集成与优化近 20 种组件，自研 10 余种实用工具，功能丰富，满足多样化场景要求；
2. 完整支持 SQL 标准：支持最全 SQL 语法、存储过程，全面支持 SQL99/2003 和存储过程方式；全面高标准通过 TPC-DS 标准测试集的全部 99 个测试项；性能优异；
3. 支持分布式事务处理，支持对数据的增删改操作(CRUD),是业界少数拥有此能力的厂家之一，使得用户更好地迁移历史应用，并降低使用人员的专业要求门槛。
4. 基于汇通国信实战经验的混合云架构优化：基于 Docker 容器技术的弹性计算资源分配和多租户管理能力，系统支持将公有云和私有云进行混合和匹配，以获得最佳的效果，满足不同规模客户的需求；
5. 系统具备 99.99%的一流可靠性。

产品价值：

1. 构建企业级智能大脑
2. 打造企业统一信息视图，消除数据孤岛
3. 强大流处理实时分析业务变化

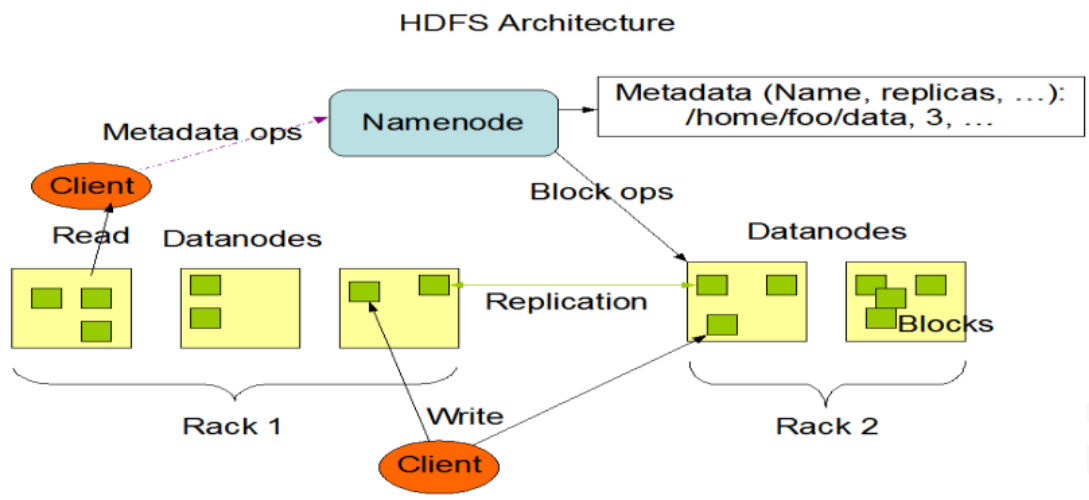
5.2 产品功能

提供分布式存储系统、批处理、流处理、交互式等多样的计算引擎、NOSQL 数据库、增强的内存计算引擎、统一资源调度管理、任务调度、权限管理、运维管理等功能。

5.2.1 分布式存储系统

汇通国信基于改造后的分布式文件系统 HDFS 作为存储海量数据的分布式存储软件架构。HDFS 被设计成适合运行在通用硬件(commodity hardware)上的分布式文件系统。HDFS 是一个高度容错性的系统，适合部署在廉价的机器上。HDFS 能提供高吞吐量的数据访问，非常适合大规模数据集上的应用。HDFS 放宽了一部分 POSIX 约束，来实现流式读取文件系统数据的目的。

HDFS 采用 master/slave 架构。一个 HDFS 集群是由一个 Namenode 和一定数目的 Datanode 组成。Namenode 是一个中心服务器，负责管理文件系统的名字空间(namespace)以及客户端对文件的访问。集群中的 Datanode 一般是一个节点一个，负责管理它所在节点上的存储。HDFS 暴露了文件系统的名字空间，用户能够以文件的形式在上面存储数据。从内部看，一个文件其实被分成一个或多个数据块，这些块存储在一组 Datanode 上。Namenode 执行文件系统的名字空间操作，比如打开、关闭、重命名文件或目录。它也负责确定数据块到具体 Datanode 节点的映射。Datanode 负责处理文件系统客户端的读写请求。在 Namenode 的统一调度下进行数据块的创建、删除和复制。



汇通国信分布式文件系统，针对海量非结构化数据存储设计的分布式系统，构筑在普通的 Linux 机器集群上，可为外部提供高可靠和高并发的存储访问。高可扩展、高可用、高性能、面向互联网服务.特性如下：

- 采用扁平化的数据组织结构；
- 使用 HA 架构和平滑扩容；
- 支持多种客户端；
- 支持大小文件存储、流式数据访问、高吞吐量数据访问；
- 可为外部提供高可靠和高并发的存储访问；
- 支持数据存储分布策略，支持机架感知与负载均衡，支持高可用。

功能支持：

- 支持类 POSIX 文件访问接口，即文件系统提供类 POSIX 的 Java/Shell
- 文件访问接口,支持对文件/目录进行创建、追加、删除、读、写操作。
- 支持通过本地系统快照进行快速恢复，也支持通过备份文件进行恢复。
- 支持对文件和目录的读、写、增、删、查、设置操作的记录审计日志。
- 支持配额控制，即文件系统支持对租户的存储空间和文件数量进行配额

- 管理，保证各租户按照存储配额使用。

5.2.2 多样的计算引擎

(1) 基于批量的分布式计算引擎

汇通国信通过对开源 MapReduce 进行改造，基于它写出来的应用程序能够运行在由上千个商用机器组成的大型集群上，并以一种可靠容错的式并行处理上 T 级别的数据集。

一个 MapReduce 作业 (job) 通常会把输入的数据集切分为若干独立的数据块，由 map 任务 (task) 以完全并行的方式处理它们。框架会对 map 的输出先进行排序，然后把结果输入给 reduce 任务。通常作业的输入和输出都会被存储在文件系统中。整个框架负责任务的调度和监控，以及重新执行已经失败的任务。

通常，MapReduce 框架和分布式文件系统是运行在一组相同的节点上的，也就是说，计算节点和存储节点通常在一起。这种配置允许框架在那些已经存好数据的节点上高效地调度任务，这可以使整个集群的网络带宽被非常高效地利用。

(2) 分布式数据仓库引擎

汇通国信基于开源的 Hive 进行改造优化。可以将数据存放在分布式文件系统或分布式数据库中，并使用类 SQL 语言进行海量数据统计、查询和分析操作。

Hive 提供了一种类似 SQL 的查询语言——HiveQL，它支持在 from 子句中嵌套 select, project, join, aggregate, union all 和子查询。HiveQL 支持数据定义 (DDL) 语句来创建特定序列化格式的内部表和分区，以及 Buckt 类型的列。

用户可以从外部源加载数据，将查询结果通过数据操作（DML）语句分别加载和插入到 Hive 表中。HiveQL 目前还不支持更新和删除现有表中的行。

HiveQL 支持多表插入操作，用户可以通过使用一个 HiveQL 语句来执行在一个相同输入数据中的多个查询。Hive 通过共享该输入数据的扫描来优化这些查询。HiveQL 也是有很好的扩展性的。它支持 Java 实现的用户定义的列变换（UDF）和聚合（UDAF）函数。

（3） 基于内存的分布式计算引擎

汇通国信基于开源 Spark 进行深度改造，提供的基于内存的分布式快速计算。拥有 Hadoop MapReduce 所具有的优点，但不同于 MapReduce 的是 Job 中间输出和结果可以保存在内存中，从而不再需要读写 HDFS，因此 Spark 能更好地适用于数据挖掘与机器学习等需要迭代的 map reduce 的算法。与 Hadoop 相比，Spark 的中间数据放到内存中，对于迭代运算效率更高。Spark 更适合于迭代运算比较多的 ML 和 DM 运算。提供的数据集操作类型包括 map、filter、flatMap、sample、groupByKey、reduceByKey、union、sort、partitionBy 等多种 Transformations 操作类型。

同时还提供 Count, collect, reduce, lookup, save 等多种 actions 操作。这些多种多样的数据集操作类型，给给开发上层应用的用户提供了方便。各个处理节点之间的通信模型不再像 Hadoop 那样就是唯一的 Data Shuffle 一种模式。用户可以命名，物化，控制中间结果的存储、分区等。可以说编程模型比 Hadoop 更灵活。

Spark 的适用场景：

- Spark 是基于内存的迭代计算框架，适用于需要多次操作特定数据集的应用场合。需要反复操作的次数越多，所需读取的数据量越大，受益越大，数据量小但是计算密集度较大的场合，受益就相对较小
- 由于 RDD 的特性，Spark 不适用那种异步细粒度更新状态的应用，例如 web 服务的存储或者是增量的 web 爬虫和索引。就是对于那种增量修改的应用模型不适合。

(4) 实时数据处理

汇通国信对开源 Spark streaming 进行了改造，基本的原理是将流数据分成小的时间片断（一般是秒级），以类似 batch 批量处理的方式来处理这小部分数据。这里的批处理引擎是 Spark，也就是把 Spark Streaming 的输入数据按照 batchsize（如 1 秒）分成一段一段的数据（Discretized Stream），每一段数据都转换成 Spark 中的 RDD（Resilient Distributed Dataset），然后将 Spark Streaming 中对 DStream 的 Transformation 操作变为针对 Spark 中对 RDD 的 Transformation 操作，将 RDD 经过操作变成中间结果保存在内存中。整个流式计算根据业务的需求可以对中间的结果进行叠加，或者存储到外部设备。

5.2.3 NOSQL 数据库

对于 NoSQL 数据库，汇通国信支持 Apache HBase 的实时分布式数据库，用来解决关系型数据库在处理海量数据时的理论和实现上的局限性。HBase 存储面向列、可压缩，有效降低磁盘 I/O，提高利用率。同时具有灵活的表结构，

可动态改变和增加（包括行、列和时间戳）Column 以及 Column Family，并支持单行的 ACID 事务处理。平台提供的 HBase 通过使用索引来加快数据的查询速度。

汇通国信基于开源 NoSQL 数据库 HBase 软件进行深度优化，具备对大型数据提供随机、实时的读写访问，用来解决关系型数据库在处理海量数据时的理论和实现上的局限性。

HBase 是一个分布式的、按列存储的、多维表结构的实时分布式数据库。它可以提供大数据量结构化和非结构化数据的高速读写操作，为高速在线数据服务而设计。Hbase 具有以下特性：

海量数据存储能力（PB 级）；

- 支持数据的自动切分和分布，支持动态扩容，无需停机；
- 高可靠，数据冗余备份，数据存放在 HDFS 分布式文件系统之上，不会丢失；
- 支持面向列的数据压缩，8:1 以上的压缩比，有效降低磁盘 I/O，提高利用率；
- 提供高效的数据整合能力，支持灵活的表结构，可动态改变和增加维度（包括行、列和时间戳）；
- 高效，海量数据秒级查询，支持多维度查询，支持全文检索；
- 提供多种 API 调用与二次开发接口；

HBase 支持半结构化数据（JSON/BSON，XML 形式存储）和非结构化数据例如纯文本，图片或者大对象的高效存取。由于越来越多的应用在考虑对半结构

化数据，非结构化数据做查询，检索和分析，对这些数据存储的支持能简化应用程序的开发工作，同时数据平台的优化使得对这类数据的操作性能更高。

5.2.4 统一资源调度管理

汇通国信基于开源 Yarn 进行深度优化和改造，可提升分布式的集群环境下的资源利用率，这些资源包括内存、IO、网络、磁盘等。Yarn 支持多种计算框架，基于一个全局的资源管理器，所有接入的框架要先向该全局资源管理器申请资源。提供资源隔离与动态调整。不同的框架中的不同任务往往需要的资源（内存，CPU，网络 IO 等）不同，它们运行在同一个集群中，会相互干扰，资源统一管控与调度平台提供一种资源隔离机制避免任务之间由资源争用导致效率下降。

Yarn 的调度器维护一群队列的信息。用户可以向一个或者多个队列提交应用。每次 NM 心跳的时候，调度器，根据一定的规则选择一个队列，再在队列上选择一个应用，尝试在这个应用上分配资源。不过，因为一些参数限制了分配失败，就会继续选择下一个应用。在选择了一个应用之后，这个应用对应也会有很多的资源申请的请求。调度器会优先匹配本地资源是申请请求，其次是同机架的，最后的任意机器的。

支持先进先出调度（FifoScheduler）、容量调度（CapacityScheduler）、公平调度（FairScheduler）等多种调度模式，适合不同的应用场景。

先进先出调度（FifoScheduler）：最简单的调度器，按照先进先出的方式处理应用。只有一个队列可提交应用，所有用户提交到这个队列。可以针对这

个队列设置 ACL。没有应用优先级可以配置。

容量调度 (CapacityScheduler) : 可以看作是 FifoScheduler 的多队列版本。每个队列可以限制资源使用量。但是, 队列间的资源分配以使用量作排列依据, 使得容量小的队列有竞争优势。集群整体吞吐较大, 延迟调度机制使得应用可以放弃夸机器或者夸机架的调度机会, 争取本地调度。

公平调度 (FairScheduler) : 多队列, 多用户共享资源。特有的客户端创建队列的特性, 使得权限控制不太完美。根据队列设定的最小共享量或者权重等参数, 按比例共享资源。资源抢占特性, 是指调度器能够依据公平资源共享算法, 计算每个队列应得的资源, 将超额资源的队列的部分容器释放掉的特性。

5.2.5 分布式应用协调

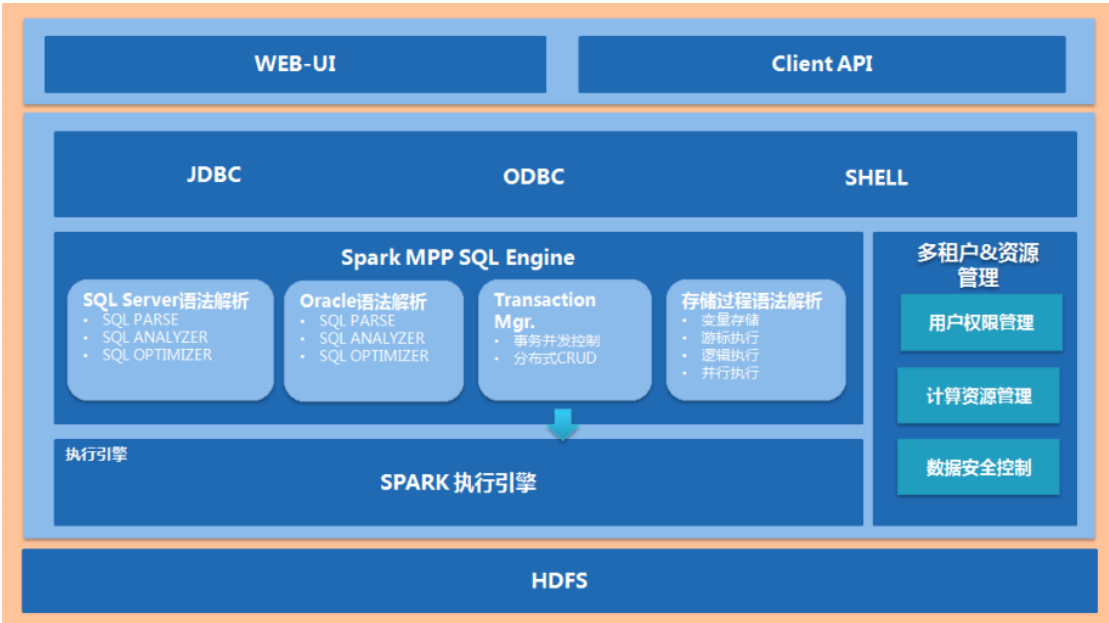
汇通国信基于 Zookeeper 系统进行优化, 应用于分布式应用的协作服务。使得分布式应用可以基于这些接口实现诸如同步、配置维护和分集群或者命名的服务。Zookeeper 很容易编程接入, 它使用了一个和文件树结构相似的数据模型。可以使用 Java 或者 C 来进行编程接入。Zookeeper 针对大型分布式系统的可靠协调系统, 提供的功能包括: 配置维护、名字服务、分布式同步、组服务等。它可以维护系统配置、群组用户和命名等信息。

- 提供数据同步机制, 数据严格按时间更新、查询和发布, 保证数据的一致性。
- 提供高并发能力, 保证集群的高性能, 在以读为主的应用中, 可以提供数千台客户端的并发查询。

5.2.6 内存计算增强 Spark

汇通国信 Spark 引擎是对原生 Spark 引擎进行了大量的改进、优化和扩展，能够完整支持 SQL 标准、支持存储过程；增强分布式事务处理能力，支持数据的增删改(CRUD)处理；全面高标准通过 TPC-DS 标准测试集的全部 99 个测试项。它具有高性能、稳定性好、功能丰富等特点，能够降低使用人员的专业要求门槛，更方便进行原有应用的迁移。

总体功能架构：



包括：Spark 执行引擎优化、Spark MPP SQL 引擎、以及接口支持。

5.2.6.1 主要功能

功能	功能描述
Spark 执行引擎优化	对 Spark 引擎进行了完善和优化，支持 Spark HA，增强服务的稳定性和性能。

SQL Server 语法解析	包括 SQL Server 解析器、SQL Server 分析器、SQL Server 优化器；完整支持 SQL Server 语法要求。
Oracle 语法解析	包括 Oracle 解析器、Oracle 分析器、Oracle 优化器；完整支持 OracleSQL 语法要求。
事务管理	包括事务并发控制、数据分布式增删改(CRUD)操作。
存储过程语法解析	包括支持变量存储、游标执行、逻辑执行、并行执行等功能。
接口支持	包括支持 JDBC、ODBC、Shell 等。

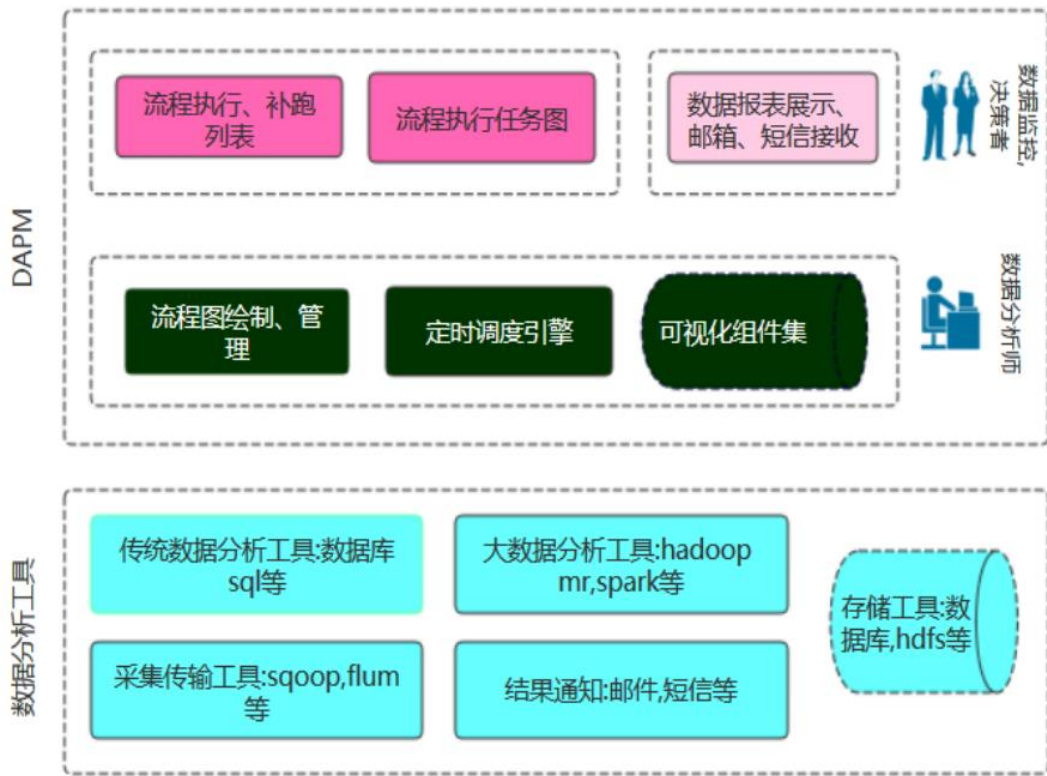
5.2.6.2 产品特性

1. 完整支持 SQL 标准、支持存储过程
2. 增强分布式事务处理能力，支持数据的增删改(CRUD)处理
3. 全面高标准通过 TPC-DS 标准测试集的全部 99 个测试项，性能优良
4. 扩展支持 Spark HA，增强系统稳定性

5.2.7 任务调度 Task Scheduler

统一任务调度系统(Task Scheduler)是一套高效的图形化大数据 workflow 配置与执行管理系统，支持可视化的大数据计算任务构建能力。通过封装底层技术复杂性，提供对 SQL 脚本、MR、Spark Scala、Shell 脚本、MySQL/Oracle、数据导入/导出等多种计算模块的可视化操作，使得开发人员更专注于计算本身而不在关注底层技术细节。

总体功能架构：



包括抽象出的各类数据分析工具组件、数据分析流程管理 DAPM (DataAnalysis Process Management)。DAPM 通过流程图绘制/管理将可视化组件编辑成数据分析流程，通过定时调度引擎实现自动化调度，通过流程执行，补跑列表实现执行监控，通过报表展示，短信，邮箱发送方式通知分析结果。

5.2.7.1 主要功能

主要功能包括：新建流程、流程查询、实例列表、补跑实例、资源配置、运行环境配置、告警配置、流程分布等。

功能	功能描述
新建流程	新建数据分析 workflow 任务组或草稿任务组，并设置定时

	任务告警共用变量等。
流程查询	对已经存在的流程或者流程草稿 进行修改 调度停用或者删除。
实例列表	查询展示所有调度流程实例,各实例任务组执行情况,以及历史时长、调度历史、广义工作流依赖图查询等。
补跑实例	支持流程提交历史调度时间点流程实例重新生成并调用。
资源配置	各项目经理配置对应 mysql、kettle 等数据源信息。
运行环境配置	Hadoop 环境、服务本身等相关配置。
告警配置	短信、邮箱服务配置。
流程分布	各项目经理修改对应管理项目 创建人流程分布情况展现以及批量修改创建人流程所属项目、创建人功能。

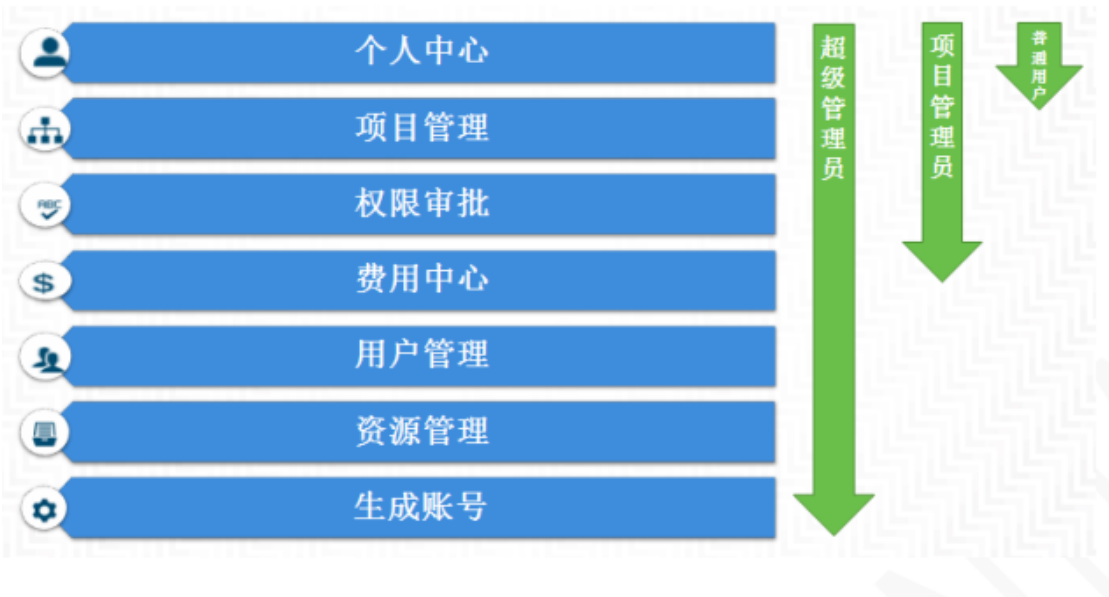
5.2.7.2 产品特性

1. 傻瓜式分布式计算开发；
2. 可视化流程定制；
3. 计算任务集中管理；
4. 抽象多种大数据计算组件；
5. 计算任务监控；
6. 流程故障自动报警

5.2.8 用户管理 User Admin

用户管理(User Admin)是汇通国信大数据平台的用户权限管理系统。用户权限管理采用多租户的思路，将数据能力按需、可控的进行开放，提供多租户下的库表资源管理，权限分配，基于项目提供存储、计算资源的分配、使用监控与计费等服务。

总体功能架构：



5.2.8.1 主要功能

功能	功能描述
监控管理	该功能旨在对 Hadoop/Spark 下各个组件的性能指标进行可视化监控，主要包括指标、热图、配置历史。
服务管理	主要展示了 HDFS、MapReduce2、YARN、Hive、Hbase、Sqoop、ZooKeeper、Metrics、Spark2、Flume、Kafka、ELK、Impala、

	UserAdmin、Task Scheduler、SQL Editor、Sentry 等各个组件服务的安装和运行情况概览、热图及配置更改与记录
主机管理	对系统集群中的物理服务器或虚拟主机,支持集群设备、平台、服务的部署、配置、升级管理能力,支持集群中在线添加、移除或者迁移计算节点、存储节点,能够在节点上自动部署 Agent 代理,实现主机的集中监控与管理,主要包含功能如下: 主机的列表,主机操作,主机服务配置,主机报警
告警管理	该功能记录了所有的报警历史记录,包括警报定义的名字、警报当前状态、警报对应服务、当前状态持续时间、警报的启停控制。
管理功能	显示系统安装的版本信息和服务信息;显示系统服务的用户及用户组;可开启禁止认证;可查看许可证的概览数据和详情数据。
版本	可查看版本信息,包括堆、名称、版本、集群名称;可以注册版本。

5.2.8.2 产品特性

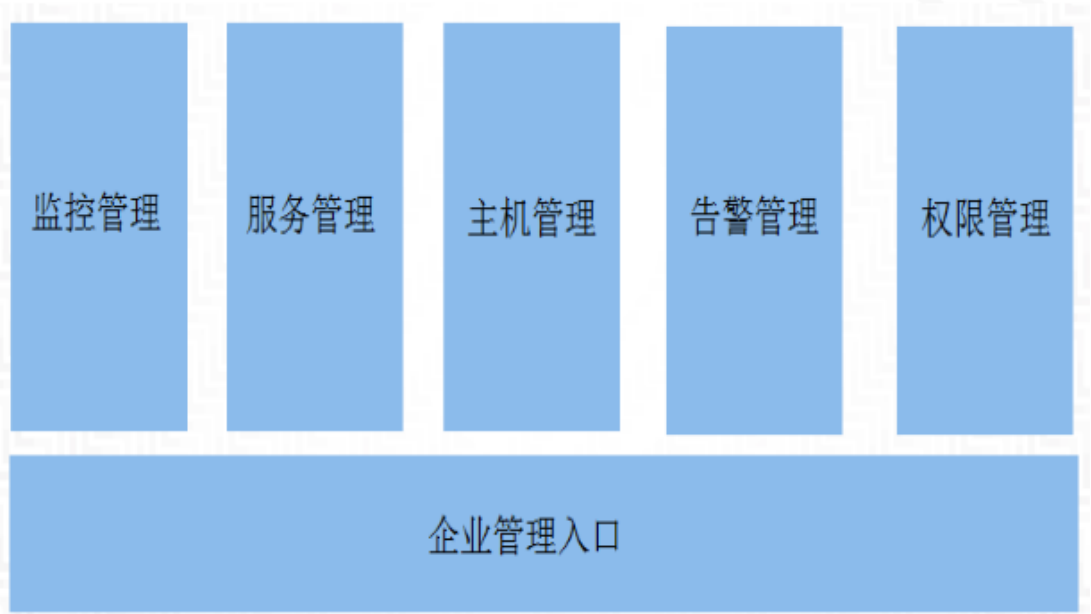
1. 完善的项目权限管理流程。
2. 丰富的用户组和角色权限管理。
3. 支持基于角色的访问控制 RBAC (Role-Based Access Control) 的权限控

制。

- 4. 支持对计算资源的自动化计费。

5.2.9 运维监控管理 Manager

Manager 是汇通国信大数据平台的运维管理系统。它具备便捷的图形化监控运维能力。在提升易用性的同时，可提供软件的自动部署、各节点运行状态的实时监控、各类用户的权限控制、统一的资源配额调度和系统自动告警等多种功能。运维人员可以轻松地部署和集中操作整个的管理集群。



5.2.9.1 主要功能

功能	功能描述
监控管理	该功能旨在对 Hadoop/Spark 下各个组件的性能指标进行可视化 监控，主要包括指标、热图、配置历史。

服务管理	主要展示了 HDFS、MapReduce2、YARN、Hive、Hbase、Sqoop、 ZooKeeper、Metrics、Spark2、Flume、Kafka、ELK、Impala、UserAdmin、Task Scheduler、SQL Editor、Sentry 等各个组件服务 的安装和运行情况概览、热图及配置更改与记录。
主机管理	对系统集群中的物理服务器或虚拟主机，支持集群设备、平台、服务的部署、配置、升级管理能力，支持集群中在线添加、移除或者迁移计算节点、存储节点，能够在节点上自动部署 Agent 代理，实现主机的集中监控与管理，主要包含功能如下：主机的列表，主机操作，主机服务配置，主机报警。

5.2.9.2 产品特性

1. 支持向导式自动化的平台安装部署；
2. 支持集群中节点的自动化配置；
3. 支持 Hadoop/Spark 组件的管理功能；
4. 支持对集群各节点的可视化、图形化、指标化监控；
5. 支持统一的告警服务，提供消息、邮件及界面可视化等多种方式展示；
6. 支持对服务、组件的定期健康检查与日志输出；
7. 支持基于整体平台或独立组件的版本升级，并保障数据的持续可用。

5.3 功能展示

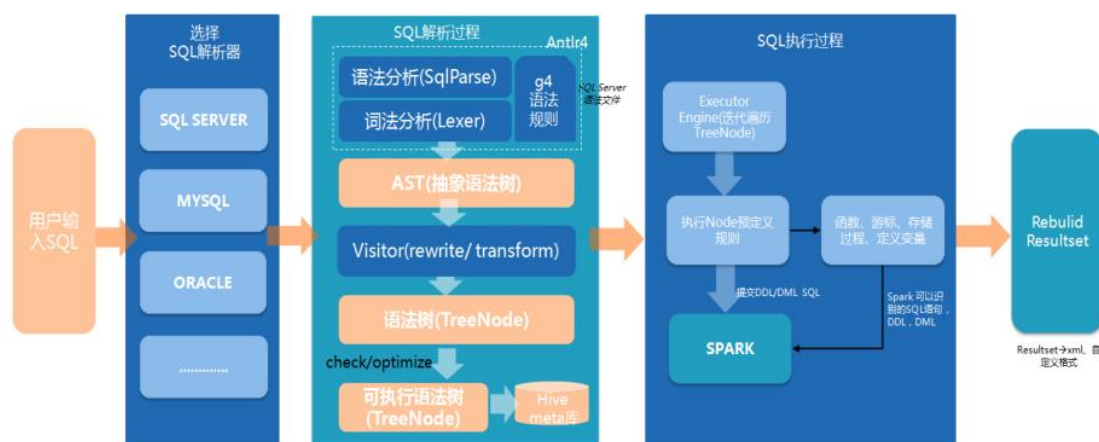
5.3.1 Spark 内存计算引擎

1、完整支持 SQL 标准、支持存储过程

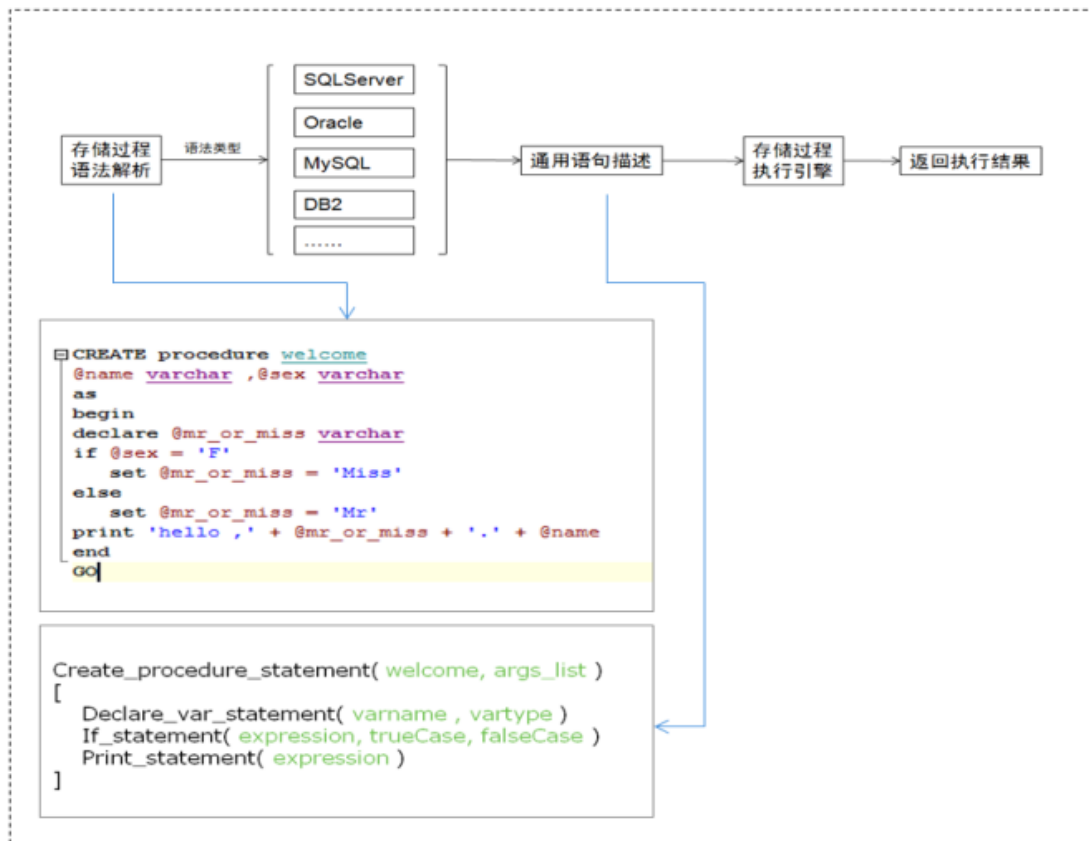
汇通国信 Spark 引擎进行了大量优化，全面支持 SQL99/2003 和存储过程方式（兼容 Oracle、SQL Server 等），支持游标、函数、变量等逻辑访问大数据基础数据库，全面高标准高性能通过 TPC-DS 标准测试集的全部 99 个测试项；保证传统企业数据处理流程的无缝迁移。



SQL 语句执行逻辑示意图：



Spark 存储过程技术支持示意图：



- 语法支持：自适应多种数据库语法类型，如 Oracle, Sqlserver , Mysql 等，包括 T-SQL 及其扩展的特殊语法及相关函数的识别、解析及执行；
- 语句支持：支持丰富的语句语法 if, while , goto , exception, transcation 等，多各类语句的语义解析及执行；
- 执行过程优化：去掉与结果无关语句、无依赖语句并行执行；

2、增强分布式事务处理能力，支持数据的增删改(CRUD)处理

- CRUD 支持：Spark SQL 支持传统数据库的 insert , update , delete 操作；
- 高性能保障：充分利用 Spark 的高性能特性，支持大规模数据集下的数据更新操作；
- 高性能优化：优化的数据读写结构，减少 IO 操作；
- 多版本数据控制；

- ACID 支持：分布式环境下的数据 ACID 支持；

CRUD - Update 测试

```
0: jdbc:hive2://10000/> select * from file_orc;
+-----+-----+
| name   | age  |
+-----+-----+
| athmodeus | 29  |
| leon    | 56  |
| kakath  | 30  |
| eve     | 25  |
+-----+-----+
4 rows selected (0.127 seconds)
0: jdbc:hive2://10000/> update file_orc set name='athmodeus' where age='30';
+-----+
| Result |
+-----+
+-----+
No rows selected (0.519 seconds)
0: jdbc:hive2://10000/> select * from file_orc;
+-----+-----+
| name   | age  |
+-----+-----+
| athmodeus | 29  |
| leon    | 56  |
| athmodeus | 30  |
| eve     | 25  |
+-----+-----+
4 rows selected (0.143 seconds)
```

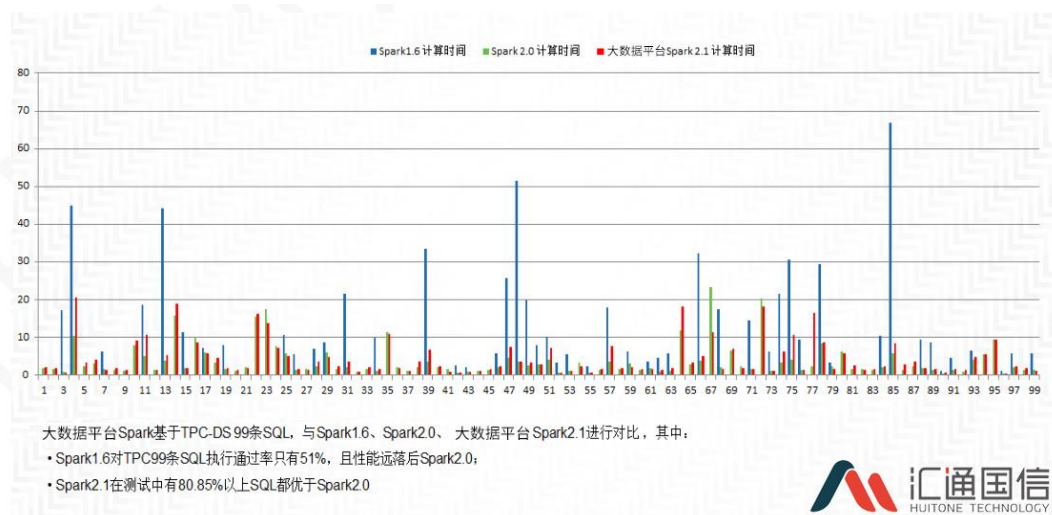
CRUD - Delete 测试

```
0: jdbc:hive2://10000/> select * from file_orc;
+-----+-----+
| name   | age  |
+-----+-----+
| athmodeus | 29  |
| leon    | 56  |
| athmodeus | 30  |
| eve     | 25  |
+-----+-----+
4 rows selected (0.143 seconds)
0: jdbc:hive2://10000/> delete from file_orc where name='athmodeus';
+-----+
| Result |
+-----+
+-----+
No rows selected (0.549 seconds)
0: jdbc:hive2://10000/> select * from file_orc;
+-----+-----+
| name   | age  |
+-----+-----+
| leon    | 56  |
| eve     | 25  |
+-----+-----+
2 rows selected (0.149 seconds)
```

3、性能优异

在 SQL 执行计划优化方面，实现了多种优化策略，性能大幅提升；全面高标准通过 TPC-DS 标准测试集的全部 99 个测试项，性能优异；

下图是 汇通国信 Spark 与开源标准 Apache Spark 的性能对比。



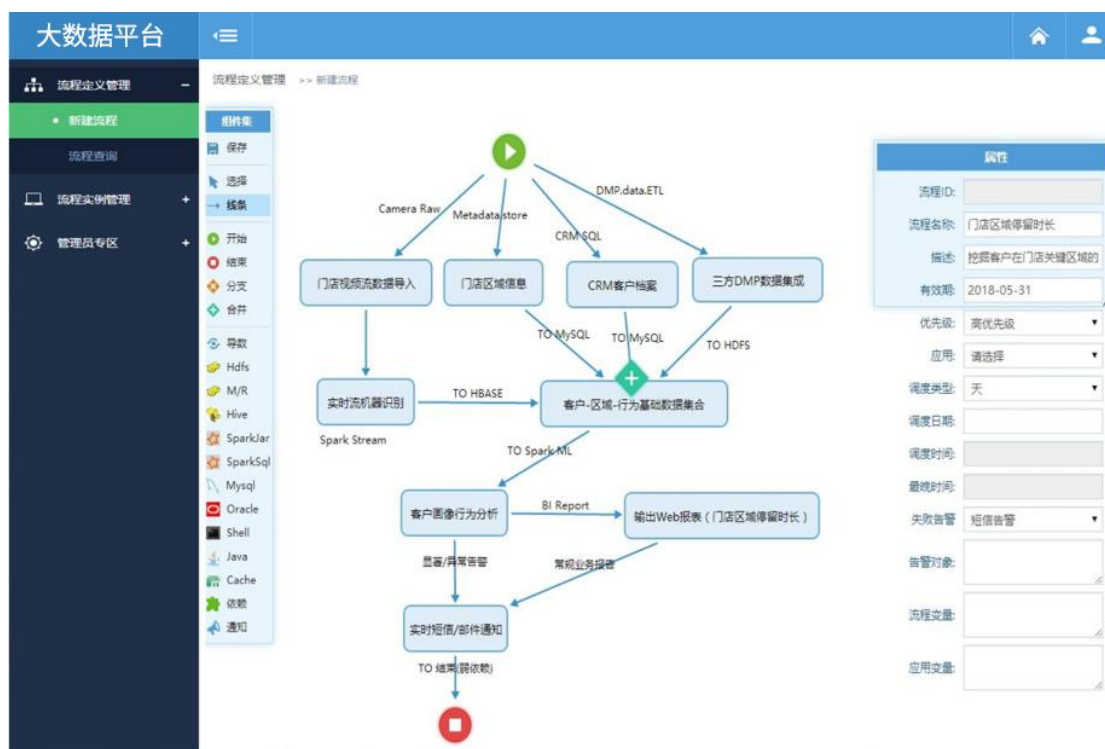
4、扩展支持 Spark HA，增强系统稳定性

通过扩展支持 Spark 的高可靠行，增强了系统的稳定性和可靠性。

5.3.2 任务调度使用

1、极简的分布式计算开发，可视化流程定制，易学易操作。

创建各种大数据任务和查询操作全部图形化完成，通过拖拽方式形成工作流程，无需命令行入口，上手简单。



2、抽象多种大数据计算组件，支持扩展

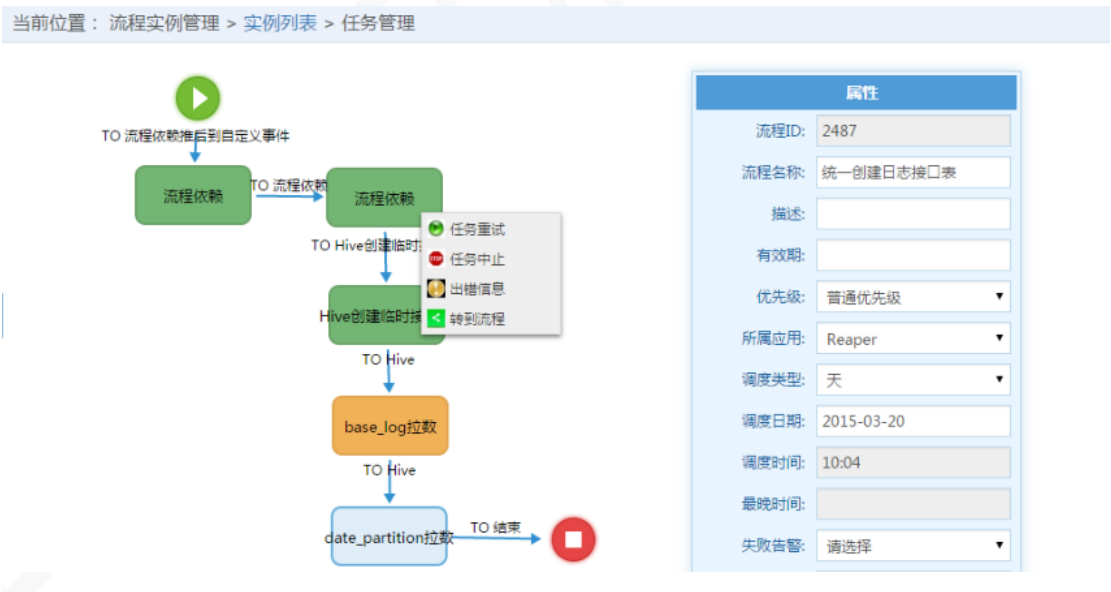
系统抽象出丰富的组件，包括：保存、线条、开始、结束、分支、合并、依赖、导数、Hdfs、M/R、Hive、Sparksubmit、SparkSql、Mysql、Oracle、Shell、Kettle、DB2、postgresql、sqlserver、http、java、sparkpy、sparkR、SparkMpp、存储过程、通知等工具，并可图形化添加多种高质量组件，全面支持和管理各种大数据

业务场景。



3、支持多种的任务调度管理

支持多种任务管理设置；支持重跑流程、停用流程；支持流程提交历史调度时间点流程实例重新生成并调用。支持查询展示所有调度流程实例,各实例任务组执行情况,以及历史时长、调度历史、广义 workflow 依赖图查询等。

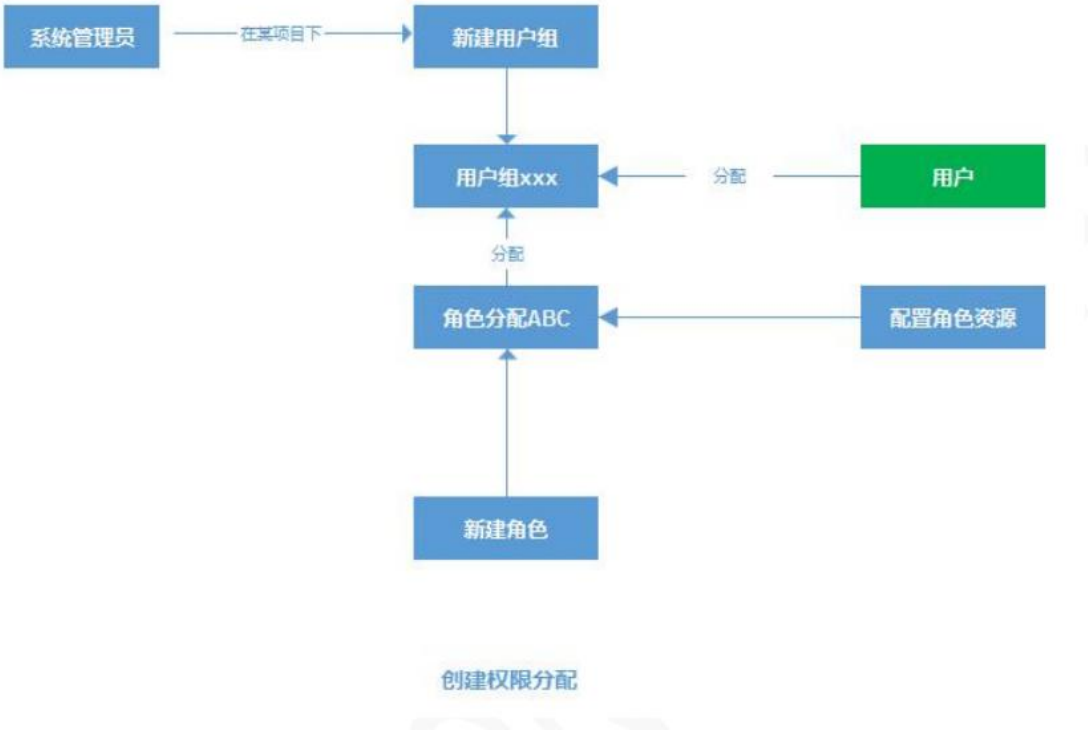


4、支持多种的通知告警方式

支持通过报表展示，短信，邮箱发送方式通知分析结果。

5.3.3 用户管理使用

1、完善的项目权限管理流程

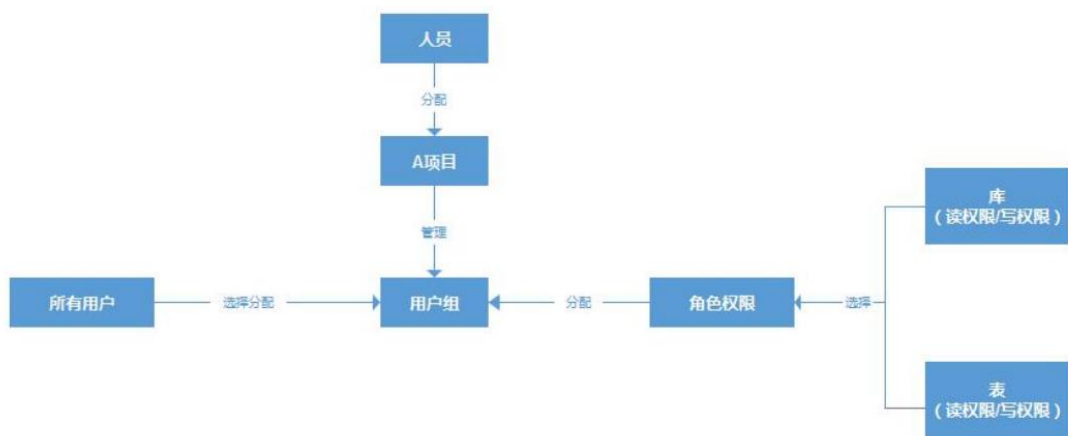


2、丰富的用户组和角色权限管理

支持按用户组和角色进行权限控制，一个项目可以分配多种角色，一个角色可以关联多个用户组（一个用户组也可以关联多个角色），一个用户组可以关联多个用户。

大数据平台					English	权限管理	
个人中心	项目管理						
项目管理							
权限审批							
费用中心							
用户管理							
资源管理							
生成账号							

项目名	项目管理员	创建时间	项目人数	项目权限管理	
lenovo_test		2017-11-20 09:34:34	0	用户组管理	角色管理
demo_test	demo1	2017-11-20 09:10:27	1	用户组管理	角色管理
FsySXOzEm		2017-11-17 19:37:38	0	用户组管理	角色管理
pZzduYbhw		2017-11-17 19:37:38	0	用户组管理	角色管理
cqjhllRI		2017-11-17 19:37:38	0	用户组管理	角色管理
WnDoMpqIs		2017-11-17 19:37:38	0	用户组管理	角色管理
VENaanmxy		2017-11-17 19:37:38	0	用户组管理	角色管理
vAVIzceHX	admin_lrc,demo1_lrc2	2017-11-17 19:37:37	6	用户组管理	角色管理
CNalftqGr		2017-11-17 15:36:08	0	用户组管理	角色管理
tOGMrHWkE		2017-11-17 15:36:08	4	用户组管理	角色管理



3、支持基于 RBAC（基于角色的访问控制 Role-Based Access Control）的权限控制。

支持批量或单一的库表权限授权，支持到表级的读写权限控。

新建角色

角色名称:

权限: 权限列表

库表	读	写
abc	☒	☐
abc2	☒	☒
abc3	☒	☐
datahub01	☒	☒
datahub02	☐	☐
datahub_rf	☐	☐
db1	☐	☐
default	☐	☐
hive_1	☐	☐



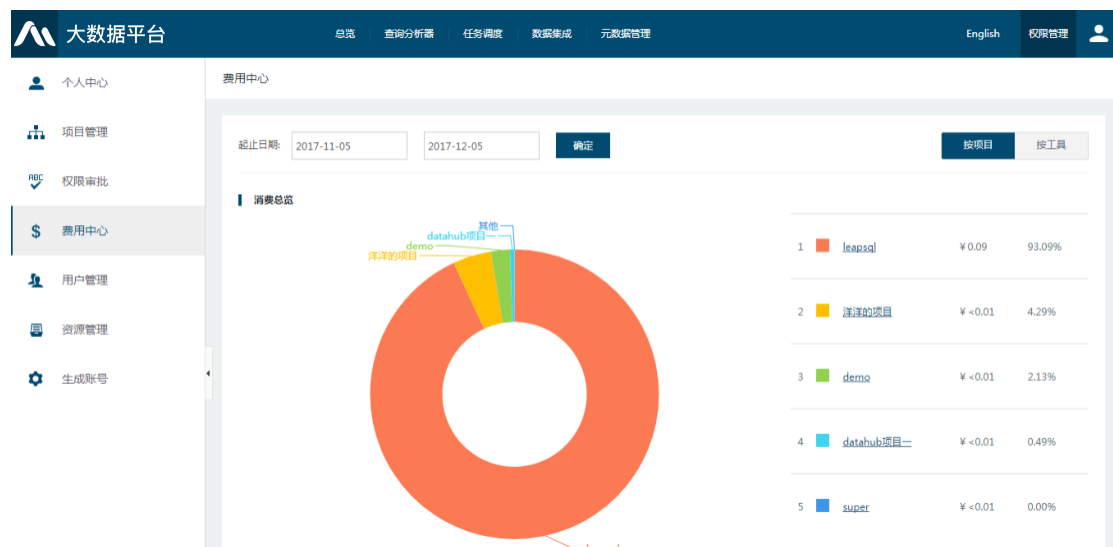
已选中(4)

库表	读	写
abc	☒	☒
abc2	☒	☒
abc3	☒	☒
datahub01	☒	☒

4、支持对计算资源的自动化计费

(1) 提供对计算资源的自动化计费功能，支持基于项目类型和工具类型分别收费。

(2) 支持查看不同项目或者不同工具的消费占比和总体趋势，支持查看时段内所有项目每天的存储量和在不同工具下的花费详情。





5.3.4 运维监控使用

1、支持向导式安装部署：

Manager 支持向导式自动化的平台安装、部署过程，支持以本地源提供安装包；提供快速配置大数据集群主机脚本，批量初始化服务器环境、拷贝安装平台介质；支持向导式平台配置，对主机、服务等快速配置。运维人员可以轻松地部署和集中操作整个的管理集群。应用程序自动执行安装过程中，部署时间从几周减少到几分钟。



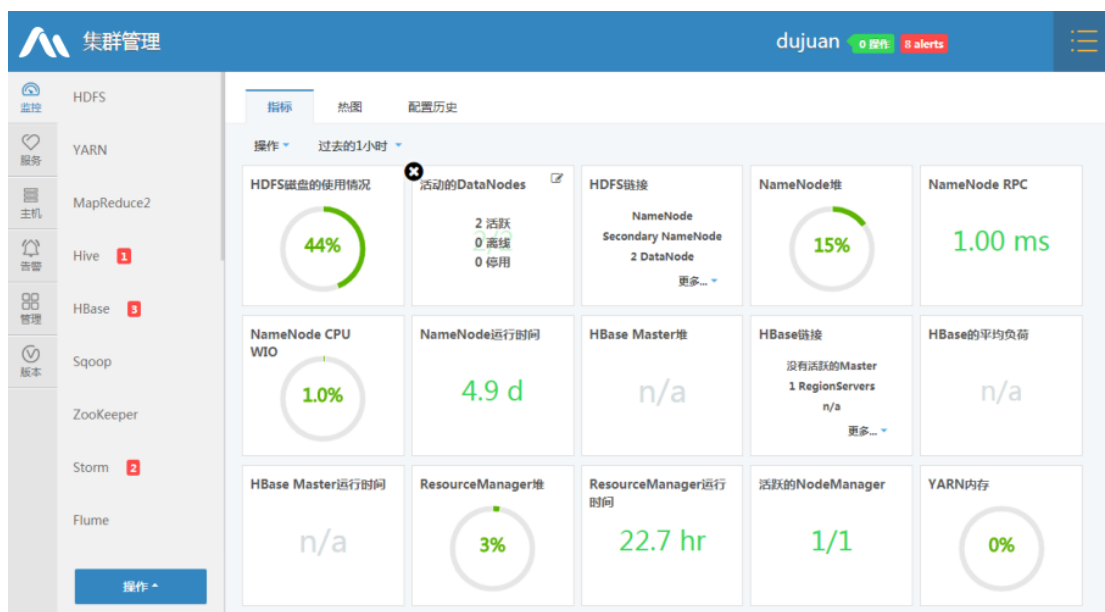
2、支持全图形化平台运维

Manager 可以为您提供群集范围内的实时视图的节点运行和服务；提供了一个单一节点的中央控制台。

系统管理					wap12	0 事件	5 alerts
操作	组: 全部 (71)						
告警定义名称	状态	服务	状态改变时间	状态			
Any	全部	全部	Any	全部			
ZooKeeper Failover Controller Process	CRIT (2)	LEAP-HDFS	2 hours ago	启用			
DataNode Storage	OK (2) CRIT (1)	LEAP-HDFS	12 hours ago	启用			
Percent DataNodes With Available Space	CRIT	LEAP-HDFS	2 days ago	启用			
Host Disk Usage	OK (2) WARN (1)	Ambari	2 days ago	启用			
NodeManager Web UI	OK (3)	LEAP-YARN	about an hour ago	启用			
NodeManager Health	OK (3)	LEAP-YARN	about an hour ago	启用			
Metrics Monitor Status	OK (3)	LEAP-Metrics	2 days ago	启用			
ZooKeeper Server Process	OK (3)	LEAP-ZooKeeper	about an hour ago	启用			

3、支持预警监控

Manager 提供集群主机 CPU 利用率、集群主机内存利用率、集群 HDFS 磁盘利用率、集群服务器运行负载和各组件安装运行管理的监控，以便及时响应系统故障，保障系统正常运行。提供设备、平台的关键指标趋势分析、历史信息对比。数据粒度可以到天、小时、分钟。支持将产生的告警通过邮件或短信等方式通知给运维人员。



系统管理

wap12

0 操作

5 alerts

首页

监控

服务

主机

告警

管理

版本

操作

组: 全部 (71)

告警定义名称	状态	服务	状态改变时间	状态
Any	全部	全部	Any	全部
🔊 ZooKeeper Failover Controller Process	CRIT (2)	LEAP-HDFS	2 hours ago	🟢 启用
⚡ DataNode Storage	OK (2) CRIT (1)	LEAP-HDFS	12 hours ago	🟢 启用
+ Percent DataNodes With Available Space	CRIT	LEAP-HDFS	2 days ago	🟢 启用
📁 Host Disk Usage	OK (2) WARN (1)	Ambari	2 days ago	🟢 启用
🌐 NodeManager Web UI	OK (3)	LEAP-YARN	about an hour ago	🟢 启用
📁 NodeManager Health	OK (3)	LEAP-YARN	about an hour ago	🟢 启用
📊 Metrics Monitor Status	OK (3)	LEAP-Metrics	2 days ago	🟢 启用
🔊 ZooKeeper Server Process	OK (3)	LEAP-ZooKeeper	about an hour ago	🟢 启用

4、Manager 支持灵活的告警管理。

Manager 支持统一的告警服务，提供消息、邮件及界面可视化等多种方式展示；支持灵活的告警规则配置及告警管理。

告警列表

状态: 未关闭 主机IP: 告警对象名称: 开始时间: 2013-05-14 20:50 结束时间: 2013-05-14 20:50

确认 处理 关闭

☐	级别	告警名称	告警对象	状态	发生时间	关闭时间	主机IP	告警内容
☐	6	HDFS文件系统可用性告警	NameNode	成功	2013-05-22 10:12:32		10.20.21.134	HDFS文件系统可用性不可用
☐	6	FTP连接失败	任务/节点	成功	2013-05-22 10:12:32		10.20.21.134	DP_GDS_GPRS_MMS任务节点FTP连接失败
☐	6	读HDFS文件失败	任务/节点	失败	2013-05-22 10:12:32		10.20.21.134	DP_GDS_GPRS_MMS任务读HDFS文件失败
☐	5	数据库连接失败	ScheduleManage	失败	2013-05-22 10:12:32		10.20.21.134	ScheduleManage数据库连接失败

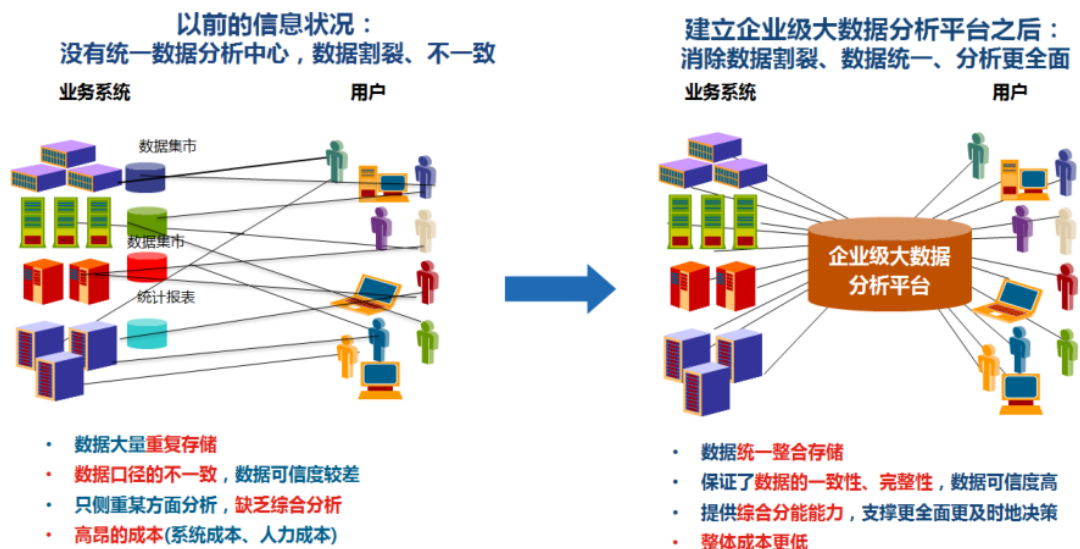
5.4 应用场景

可应用于构建企业级数据中心/数据湖、实时数据处理、明细清单实时查询等多样场景。

5.4.1 场景 1：企业级数据中心/数据湖系统

企业存在 ERP、CRM、SCM 等 IT 域数据，也存在 MES、SCADA 、DCS 、PLC 等系统/设备的 OT 域数据；各类数据的来源方式多样化，如:关

系型数据库数据、数据仓库数据、实时数据、文件数据、图片数据等，各类数据进行数据 ETL 过程后，需要将这些海量的数据进行存储管理，需要建立了统一资源调度管理能力，能够通过批处理、交互式查询计算、流式计算等多种计算引擎来进行相应的数据处理和数据加工，为后续的深度分析奠定基础。



通过企业级数据中心的建设，消除数据割裂、实现企业级数据整合、形成统一的、完整的数据，向上可提供综合分析能力。

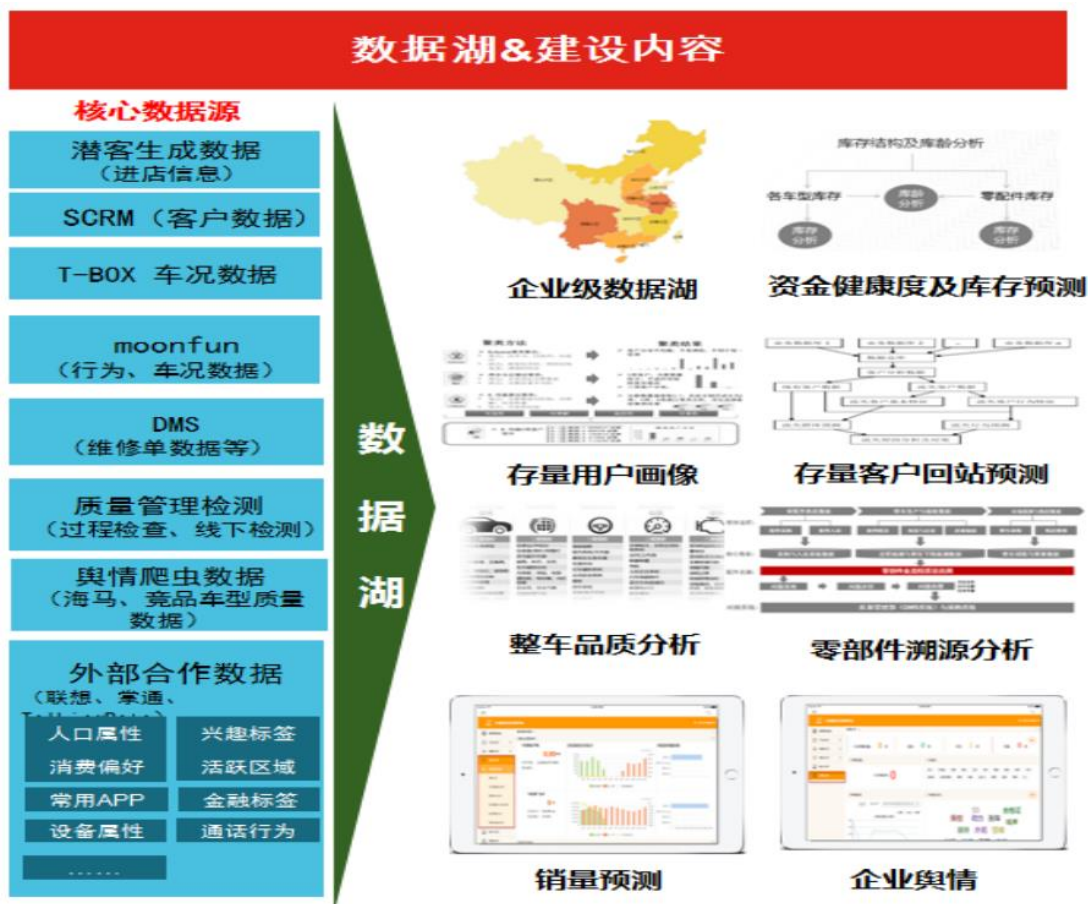
“(数据湖)是一种采集、管理和分析数据的全新模式”

“A new way of capturing, organizing and analyzing data.”

——《Big Data Requires a Big, New Architecture》Forbes



数据湖的建设让低成本的全量数据存储成为可能；让低成本、高性能、全量数据分析成为可能；促进企业内部数据开放和共享。



实现的价值：

- 高效互联

全量采集，开放共享，按需获取，不留孤岛或瓶颈。

- 全量存储

不再需要抛弃有价值的数据，全量治理和全量分析变成可能。

- 知识沉淀

分析算法和模型成为组织知识资产，实现统一管控和持续迭代。

- 敏捷创新

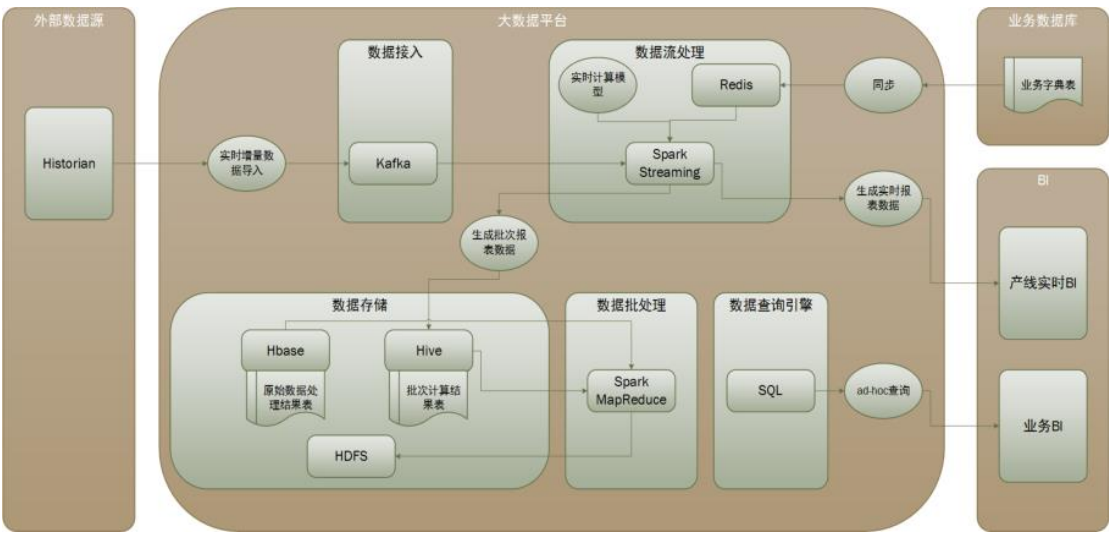
新增应用变得简单高效，创新应用迎来爆发期。

5.4.2 场景 2：实时数据处理场景

适用场景：通过构建高性能的实时数据采集、实时处理，实时分析模型，现实时高效、毫秒级数据触发能力，实现秒级单位时间窗口的数据统计分析能力。

例如：某卷烟生产企业因传统 IT 系统无法实时获取数据和实时分析，导致制丝和卷包的生产工艺提升遭遇严重瓶颈，产品质量下降形成大浪费。迫切需要在产线物联网升级的基础上建立车间海量时序数据的实施处理和分析能力，实现智能化的实时监控以及控制优化。

系统基于数据插件+kafka 消息队列，从时序库毫秒级实时采集到大数据平台；车间时序消息吞吐量>20 万条/秒，关键数据完整率达到 99.99%；通过实时处理框架秒级完成指标计算，实时流处理速度提升 12 倍；质量实时监控界面秒级更新，基于大数据平台 SQL 的 OLAP 分析能力提升 8 倍。



汇通国信大数据平台与 MES 等系统集成，帮助其利用系统历史沉淀数据构建各类训练集，并对整合后的实时数据进行拟合、分析，构建工艺流量稳态识别模型，并根据流量波动情况，挖掘波动根因，并协助修改，最终提升产品质量。



5.4.3 场景 3：明细清单实时查询

传统的明细清单查询系统存在建设成本高、查询性能不足、系统扩展性差等问题。

- 建设成本高。传统清单查询系统部署在小型+Oracle+高端盘阵，成本高，扩容成本更高。
- 查询性能不足。随着清单量的快速增加，系统部署过于集中，清单查询、预处理等应用部署在同一台主机上，应用之间相互影响较大，降低了查询性能；
- 系统扩展性差。传统数据库的横向扩展性差，且无法做到在线扩展，只能通过系统升级项目应对不断增长的数据量和访问量

在大数据时代，利用大数据相关组件的特性构建分布式清单查询系统，能很好地解决上述问题。

分布式清单查询系统包括清单导入、清单预处理、清单库、查询服务、系统管控等功能。

6 智能分析平台

6.1 总体介绍

目前，人工智能技术已广泛应用于图像识别、自然语言处理、生产制造、新材料、工程设计与仿真分析、市场营销等多个领域，并带来了不同以往的成果和体验，为经济社会转型升级提供重要支撑。希望通过人工智能相关技术手段，深挖数据价值，构建以数据为依据的前瞻性决策，引领企业发展。

汇通国信数据智能分析 Intelligence 平台提供深度学习分布式框架、机器学习组件、知识图谱组件、预测库、优化库、知识库等，具备特征工程、数据建模以及机器算法学习库的功能，可以辅助用户发掘隐藏在数据背后的巨大商业价值，加快从数据到业务的价值实现。



图、汇通国信数据智能分析平台 Intelligence 功能架构

其中，DeepNEX 是一款多租户的深度学习私有云平台，它支持多种深度学习框架的分布式深度学习，降低现今主流开源深度学习框架的使用门槛；HyperMiner 是一款自动化机器学习平台，它支持多种机器学习模型，实现自动建模、自动调参的分布式机器学习；HyperGraph 是企业级大数据图分析平台，可在秒级别对百亿级知识图谱进行分析和挖掘，为企业客户提供一个海量多维度数据实时响应、可视化交互、多源数据整合的一站式平台。

主要功能：

功能	功能描述
深度学习分布式框架	深度学习分布式框架面向深度学习用户，提供易管理、支持多种部署模式、支持超参数自动搜索优化的深度学习平台。通过深度学习分布式框架，在硬件平台实现进行资源的申请(包括运行环境、处理器核数、内存、GPU 数量)等功能。 深度学习分布式框架同时提供方便、快捷的 Python 编程环境，Shell 终端来时刻运行训练和调参的环境。深度学习的编程人员可以在申请资源后，秒级打开窗口进入编程状态，深度学习分布式框架将深度学习的系统配置、部署、资源分配完全自动化，可以极大提高深度学习的工作效率。
机器学习组件	机器学习组件高速内存计算框架对常用的机器学习算法的实现库，同时包括相关的测试和数据生成器。机器学习库支持常见的机器学习问题，如：二元分类，回归，聚类以及协同过滤，同时也包括一个底层的梯度下降优化基础算法。

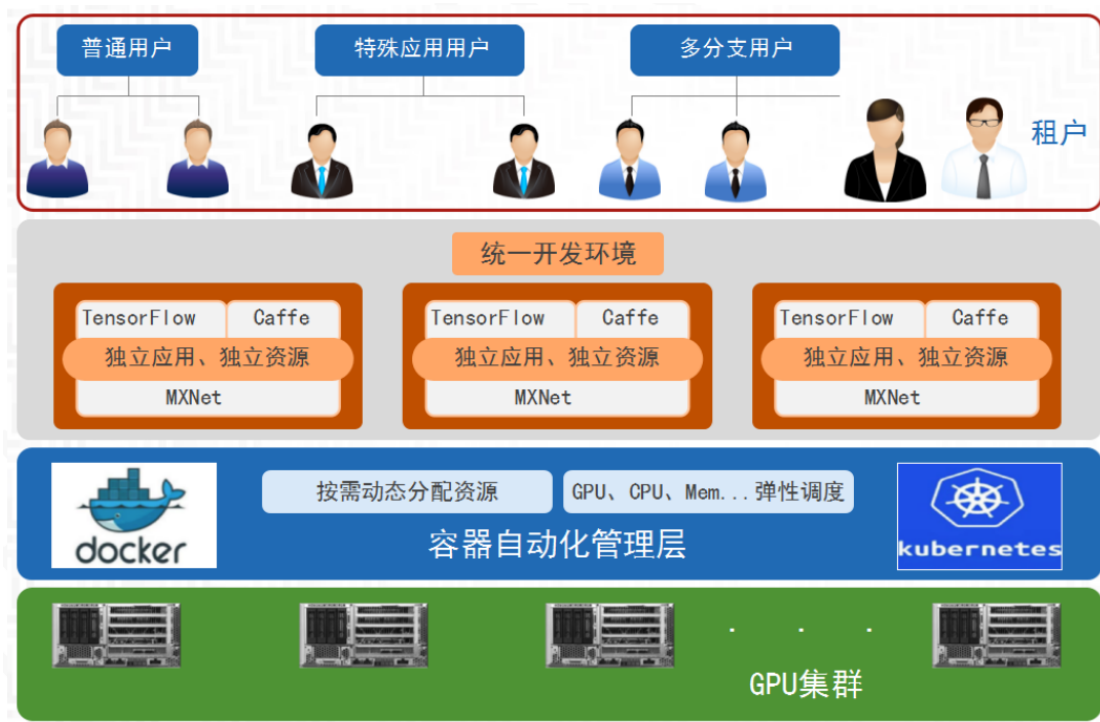
	高速内存计算框架 机器学习库实现了大部分的主流机器学习算法，如二元分类、线性回归、聚类 、协同过滤、隐性反馈 vs 显性反馈 、梯度下降基础算法等等，也提供 R 统计分析工具、Python 工具等。
知识图谱组件	汇通国信知识图谱组件是一个可以在秒级别实现对数以百亿计规模的图数据进行操作的大数据关联关系综合分析工具箱，旨在给企业用户提供一个可以不断拓展数容量和维度、打通不同来源的异构数据并且自由探索和挖掘关联数据的工具。
预测库	基于训练数据完成 0-1 事件预测、时间序列预测（如：需求预测、故障预测、事件预测、价格预测等）、关联规则挖掘等。
优化库	针对多种场景提供最优化算法，如：调度算法、排队算法、动态规划等，以便实现多个场景下的最优调度、最优资源分配、最优定价、最优库存等。
知识库	提供对自然语言处理的文本理解和知识图谱。知识图谱针对大规模时空数据和复杂关系型数据提供中枢节点发现、社区发现、关系衡量和用户属性传播等分析功能，实现从关系数据多源融合、图并行计算、可视分析到业务智能的端到端解决能力，深度解决用户现有及潜在业务痛点。
分析应用模板	分析应用模板专门针对大数据领域的开发了多维分析框

	架，预置丰富的分析模型，支持汽车、钢铁、能源、烟草、零售等多个行业的快速领域分析和系统建模需要，极大地降低信息计算人员的使用门槛。
可视化工具	数据可视化工具支持支持多方数据源接入，提供多种数据报表分析展现，包括：智能报表展示、仪表盘展现、自助分析报表、多维分析、Office 插件支持、移动可视化保险等，操作方式简单、高效并且深度结合业务场景需要，真正为用户提供的一站式、可视化、低门槛、高价值的数据分析服务。

6.2 产品功能

6.2.1 深度学习 DeepNEX

DeepNEX 是一款多租户的深度学习私有云平台。DeepNEX 旨在降低现今主流开源深度学习框架的使用门槛。通过帮助客户解决深度学习开发和训练过程中计算资源短缺、集群环境配置复杂、资源管理缺失等问题，让个人和企业更加方便、高效地进行深度学习相关的研发。



图、深度学习 DeepNEX 功能架构

6.2.1.1 主要功能

功能	功能描述
深度学习云平台	- 多租户功能深度学习云平台 - 自定义环境中的处理器、内存、GPU、并行环境中的虚拟容器配置 - 分布式分布式深度学习平台 - 支持多种深度学习框架：Tensorflow，MxNet，Caffe、CNTK、 Torch、Pytorch、Caffe2 - 支持计费
GPU 硬件支持	支持 NVDIVA Gforce 和 Tesla 系统 GPU

无缝扩展	如果用户量、使用深度学习任务的负载持续增大,可以随时增加硬件设备,该系统可以支持无缝扩展。
平台扩展性要求	如果用户量、使用深度学习任务的负载持续增大,可以随时增加硬件设备,、该系统可以支持计算节点的无限水平扩展,支持扩展 GPU 节点集群 ≥ 128 节点
资源调度功能	可根据资源的优先级做计算资源调度: - 对用户透明的计算节点分配 - 按需分配 - 排队机制
作业编排调度	利用容器管理引擎进行编排调度
运行环境	封装支持 Tensorflow, MxNet, Caffe、CNTK、Caldi、Torch、Pytorch、Caffe2 等主流框架,集成 Python 开发环境,简单易用可根据业务要求添加特定的 Python 第三方库
分布式训练	平台可支持分布式训练,可以自动地部署训练任务到集群中。
示例程序	提供应用层的 demo 和案例
扩展性	要求做到计算节点的水平扩展,支持扩展 GPU 节点集群 ≥ 128 节点
环境隔离	容器化隔离用户运行环境
用户管理	- 支持基于用户权限的管理,支持多种权限的用户应用,支持

	集成 LDAP、AD 等管理认证系统 - 管理员用有权创建和删除用户，限制用户创建应用的数目以及每次能分配的资源配额 - 每一个用户（包括管理员）都有权限自行创建应用，以便建立和使用独立的深度学习开发环境
用户目录	- 支持 NFS, Luster, GPFS 等多种文件系统格式； - 使用 NFS 文件系统，支持每个用户建立独立的 NFS 目录 - 支持 HDFS 文件系统
GPU 运行模式	可根据深度学习环境的资源需求采用多种方式共享集群 GPU 卡： - 单卡模式：一个应用独享一块 GPU 卡 - 单机多卡模式：一个应用独享一台服务器及其所有 GPU 卡 - 多机多卡模式：一个应用独享多台服务器及其所有 GPU 卡 - 随机多卡模式：一个应用使用不同节点的多个空闲 GPU 卡 - 网络优化：控制一个应用分配的多 GPU 卡在同一个交换机下
计费功能	每个用户都有独立的 NFS 目录，在创建应用时，用户的

	NFS 目录会自动挂载到应用中。用户可以在自己的多个应用之间方便的共享文件。
计费功能要求	可通过多种方式或多种方式并行计费： - GPU 设备使用量、时间（独用、共享） - CPU 资源使用量、时间 - MEM 资源使用量、时间 - 存储资源使用量（IO）
开发接口要求	所有深度学习软件接口直接向用户开放，用户可以非常方便的要求定制或进行二次开发。
数据存储要求	支持多节点线性扩展，最大可扩展容量 100PB
	支持多个任务同时访问一个机器的多个磁盘的不同文件

6.2.1.2 产品特性

- 1、集成 Tensorflow、MXNet、Caffe 等多种主流深度学习框架，用户可以同时使用，不需要做额外的安装配置。
- 2、支持多人同时使用，每个用户有独立的存储空间和计算资源。
- 3、提供统一的 Web 分析环境，开发调试方便。
- 4、提供深度学习领域最新的 Demo 和应用案例，方便客户的教学与实践。

6.2.2 机器学习 HyperMiner

HyperMiner 是一款自动化机器学习平台。它通过将专业的数据特征处理、

模型选择、模型优化等机器学习流程自动化，配合以拖拽式的图形化操作界面，打造了一个无门槛的机器学习平台。无需机器学习相关背景和专业团队，任何用户和企业都能够高效利用 HyperMiner 内置提供的行业解决方案进行有关工业、金融、零售，医疗、教育等行业数据的建模分析和结果评测，甚至二次开发来满足特定需求。

技术架构：



图、机器学习 HyperMiner 技术架构

6.2.2.1 主要功能

功能	功能描述
机器学习基础库	机器学习基础库，包括 Scikit-Learn,Tensorflow,MLib 等
特征工程	进行数据清洗/预处理等,进行离散特征变换、连续特征变换、时空特征变换等工作。
数据管理	提供基本的数据管理，数据文件的上传、下载、删除和预览

解决方案/模板库	支持从通用解决方案模板创建项目；同时，针对不同行业特点，定制化的自动特征工程
用户管理	支持基于用户权限的管理，支持多种权限的用户应用；
控制中心	支持自动特征工程、自动调参

6.2.2.2 产品特性

1. 自动化的机器学习平台，可以实现自动特征工程、自动选择模型、自动优化模型，用户不需要太多机器学习相关的背景知识即可非常简单地在本平台上进行机器学习相关的应用开发。

2. 可自主设计机器学习算法流程，灵活易用。

3. 对算法精度进行深度优化，支持多种机器算法分布式，专家级的分析结果。

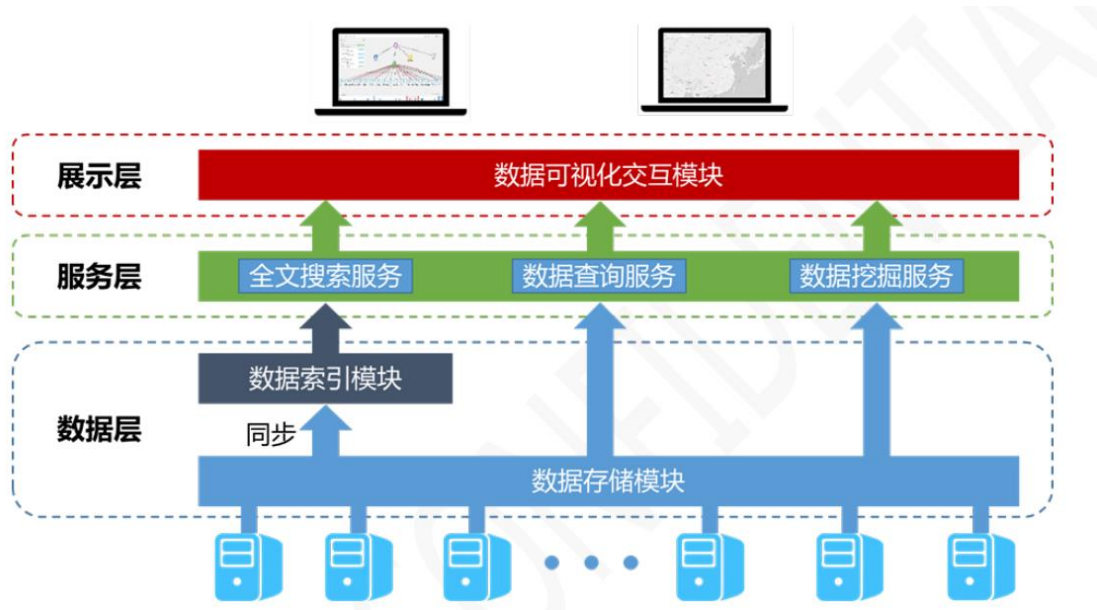
4. 定制化行业方案优化，更适合行业场景。

6.2.3 图分析

HyperGraph 是企业级大数据图分析平台。凭借自主研发的核心技术，HyperGraph 可在秒级别对百亿级知识图谱进行分析和挖掘，为企业客户提供一个海量多维度数据实时响应、可视化交互、多源数据整合的一站式平台。HyperGraph 可以将众多行业日益增长的数据建模成知识图谱，包括金融行的交易数据、制造业贸易数据、公安刑侦数据、医疗数据等。先进的可视化方案可以让用户通过搜索、拖拽、查询的方式进行高级数据挖掘分析，从而满足各行业

的需求。

技术框架：



图、图分析 HyperGraph 技术架构

汇通国信图数据分析平台具有：

- **更佳的扩展性：**系统支持从单台机器到多台机器集群的横向扩展与部署，仅用 5 台普通 PC 就可以对数以百亿计规模的图数据进行秒级延迟的计算；
- **更快的交互性：**系统充分地利用了现代 CPU/GPU 的加速功能，通过高效的内存应用对前台复杂的查询操作请求以及可视化渲染提供快速响应；
- **更易的操作性：**系统提供了包括数据导入、查询交互以及数据挖掘在内的一站式服务，用户可以通过简单的鼠标点击拖曳就可以轻松实现业务需求；
- **更好的接口性：**系统提供了不同颗粒度的图数据存储、交互以及分析的 API 集合，用户可以方便地根据自己的需求进行二次开发，添加业务需要

的功能模块。

6.2.3.1 主要功能

功能	功能细项	功能描述
查询及探索	实体全文搜索	按用户提供的关键字进行精确/模糊搜索，并支持条件筛选缩小候选元素空间
	关系/邻居搜索	通过已有的实体进行邻居搜索，并支持对邻居实体进行条件过滤以精确定位感兴趣的元素
	隐藏关系搜索	通过路径搜索，找到非直接相连的实体之间的最短联系
	全局报表功能	交互式的报表展示功能，让用户掌握全量数据的分布与趋势
	时空数据交互探索	与时空数据进行交互
	可视化布局	提供基于力导向的、层次的以及同心圆的布局
图数据挖掘	分类	对图数据上的节点分类
	聚类	图节点聚类，分成不同社区
	链接预测	对于未发生连接两个节点预测将来发生链接的概率
	特征工程	对图节点进行各类网络特征的提取

	推荐	基于用户及产品之间的二分图进行产品推荐
	复杂模式查询	自定义复杂模式查询，例如交易环，担保圈等等

6.2.3.2 产品特征

1. 自主研发的图分析技术，具有低延时响应、高扩展性。
2. 大规模数据查询实时响应，可以在秒级别实现对数以百亿计规模的图数据进行操作。
3. 友好图数据可视化交互，高效的内存应用对前台复杂的查询操作请求以及可视化渲染提供快速响应。
4. 提供多样的时空图谱数据分析，适合多样场景。
5. 灵活的二次开发 APIs，系统提供了不同颗粒度的图数据存储、交互以及分析的 API 集合，用户可以方便地根据自己的需求进行二次开发，添加业务需要的功能模块。

6.2.4 自助查询分析器 SQL Editor

SQL 查询分析器（SQL Editor）是基于汇通国信大数据平台构建的在线查询系统，借助 SQL Editor 系统，用户无需掌握复杂的大数据开发技术，只要熟悉 SQL 语法，就可以快速对海量数据进行类似关系型数据库的查询操作，并以可视化的方式获得直观的查询结果。

6.2.4.1 主要功能

功能	功能描述
数据查询管理	支持对 Hive、MySQL、SQL Server、Oracle、DB2、phoenix、PostgreSQL 等数据源的数据进行查询、管理等。
文件管理	支持上传文件至 HDFS, 对 HDFS 上的文件实施管理、下载等。
Spark_R	通过 R 语言实施数据查询、管理等。
S_Python	通过 Python 语言实施数据查询、管理等

6.2.4.2 产品特性

1.丰富的数据查询功能:

- (1) 支持对 Hive、MySQL、SQL Server、Oracle、DB2、phoenix、PostgreSQL 数据源的增加、删除、检索等。
- (2) 支持查看表结构、存储信息等, 支持对表执行 CURD 操作。
- (3) 提供 SQL 编辑框, 支持多窗口查询; 支持 SQL 语法及关键字自动补全, SQL 格式化, 撤销, 恢复, 保存等。
- (4) 支持 Spark、Mapreduce、Impala 和 Spark SQL Sever 四种执行方式。
- (5) 支持查询结果可视化, 支持交叉表、柱状图、折线图、饼图的呈现方式。
- (6) 查询结果支持 CSV 格式导出。

(7) 提供对存储过程的创建、编辑、执行操作。

(8) 支持对 SQL 执行历史和保存进行二次编辑，查看结果。

2. 完善的文件管理功能：

(1) 支持对项目下的用户组目录和用户私有目录进行管理。

(2) 支持文件和文件夹的移动、复制、删除、重命名，支持文件的下载和上传，支持新建文件夹。

3. 全面支持 Spark R 功能：

(1) 提供 R 语言编辑框，支持多窗口查询；支持剪切，复制、粘贴、撤销、恢复，保存等。

(2) 支持查询结果可视化，支持交叉表、柱状图、折线图、饼图的呈现方式。

(3) 支持对保存的查询语言再次编辑和执行，支持对文件和文件夹进行复制、移动、重命名、删除；支持新建文件夹。

4. 全面支持 Python 功能：

(1) 提供 Python 语言编辑框，支持多窗口查询；支持剪切，复制、粘贴、撤销、恢复，保存等。

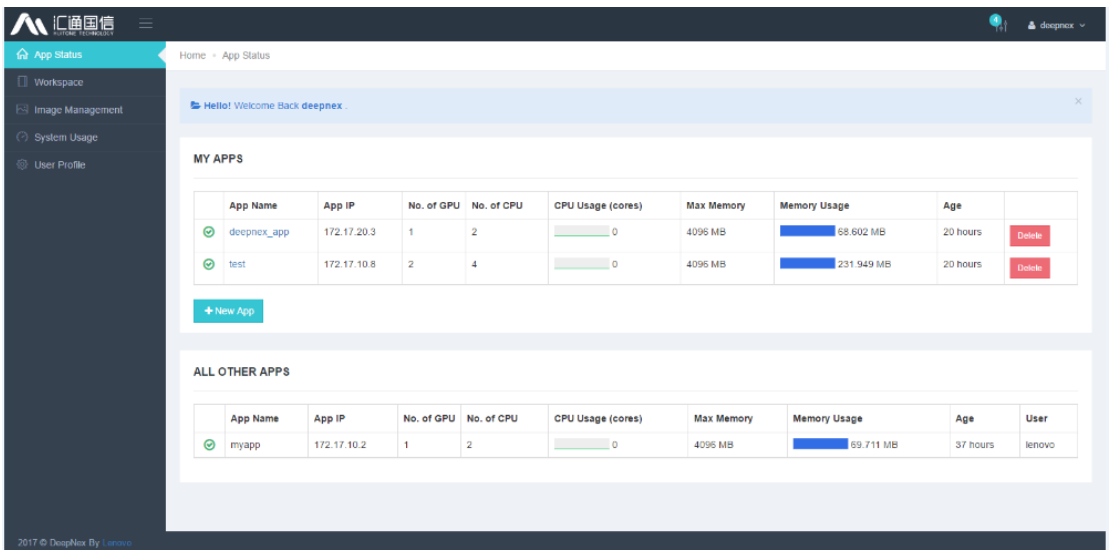
(2) 支持查询结果可视化，支持交叉表、柱状图、折线图、饼图的呈现方式。

(3) 支持对保存的查询语言再次编辑和执行，支持对文件和文件夹进行复制、移动、重命名、删除；支持新建文件夹。

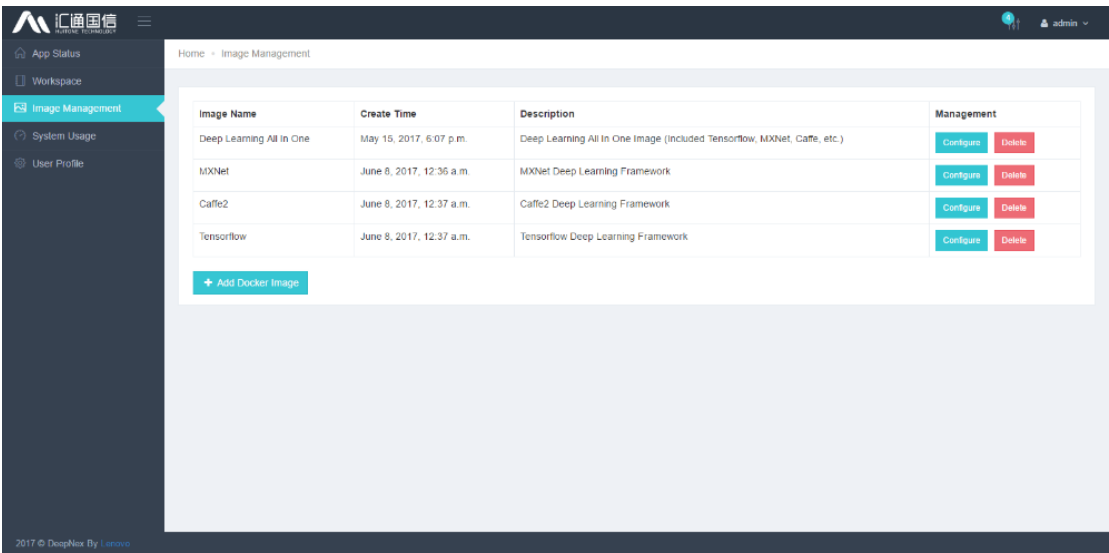
6.3 功能展示

6.3.1 深度学习 DeepNEX

每一个用户（包括管理员）都有权限自行创建应用，以便建立和使用独立的深度学习开发环境。



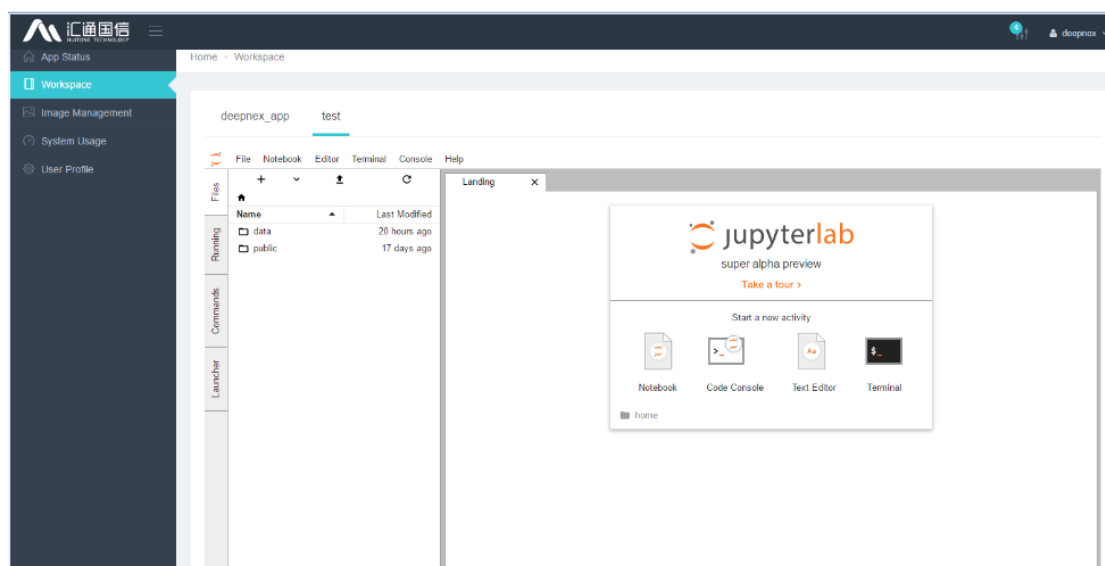
容器镜像管理：



应用程序 Web IDE：

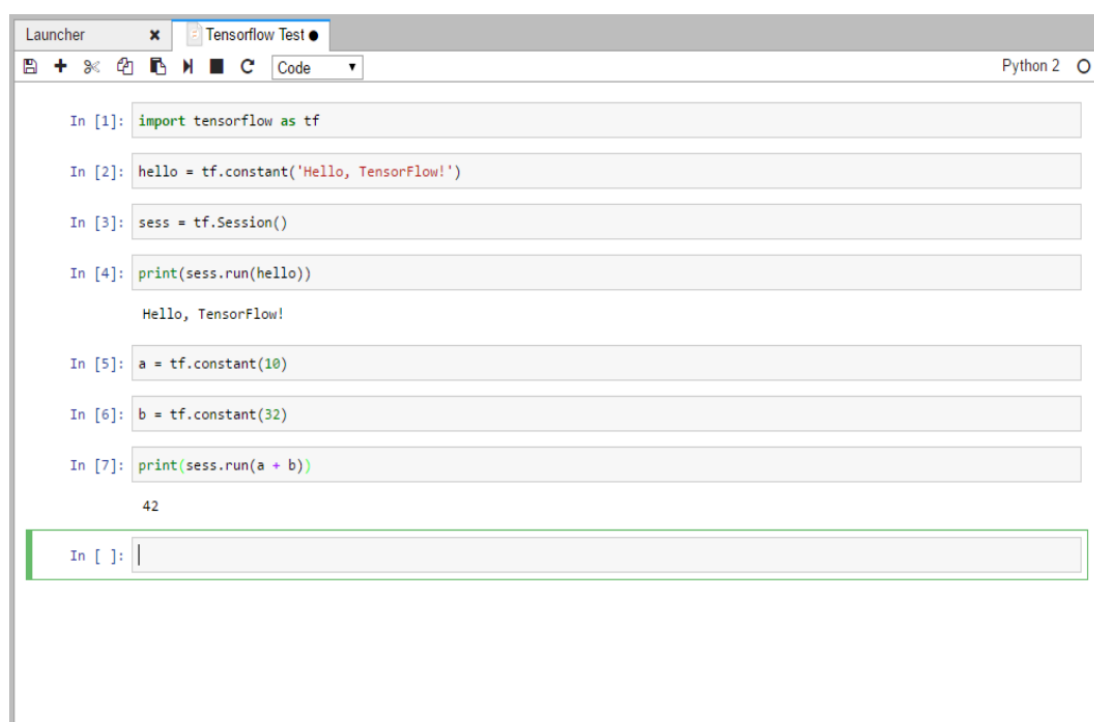
从应用程序列表页面点击应用名或者导航跳转到应用程序的操作界面。在操

作页面可以创建并运行 notebook，打开 terminal 窗口，上传下载文件等。



测试和使用：

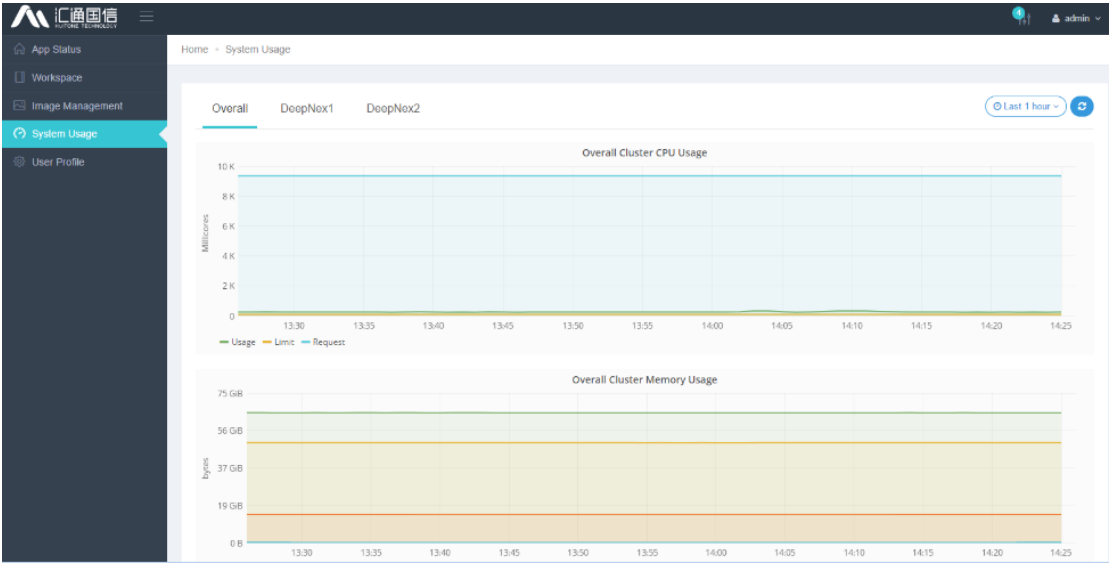
在创建应用后可以打开一个 Jupyter Notebook，并用下面代码简单测试 Tensorflow 是否能正常运行。



用户可以在这个环境中编写程序，运行，调试，察看结果。可以调用已经集成的深度学习模块，可以使用 Python 的各个主流包。这个环境中的环境变量、

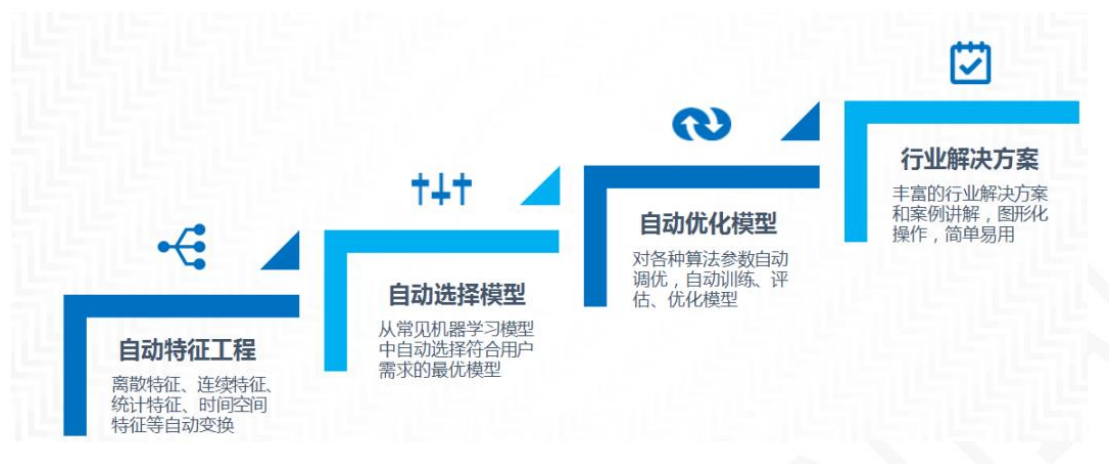
文件等和其他用户完全独立，互不影响。

系统资源监控：



6.3.2 机器学习 HyperMiner

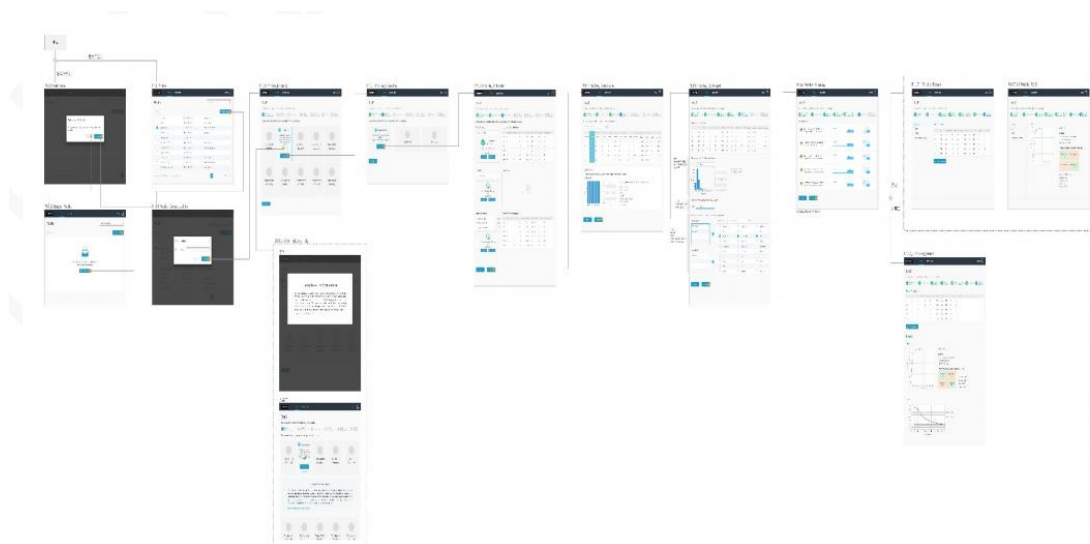
HyperMiner 系统支持机器学习的自动特征工程、自动选择模型、自动优化模型、具有丰富的行业解决方案。



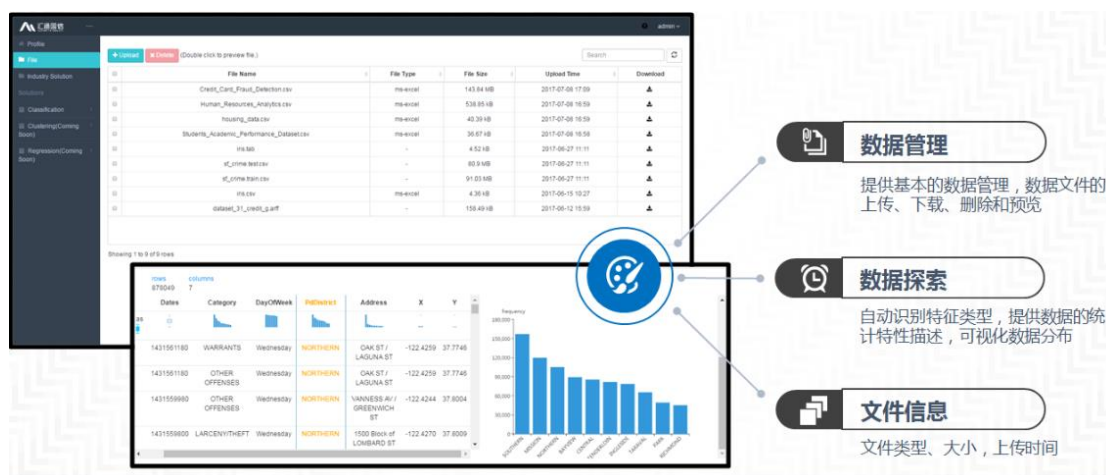
自动特征工程与自动化建模：



良好的用户引导流程:



数据管理:



项目管理:

项目管理
提供基本的项目管理，项目的新增、删除和修改

项目开发
项目开发面板可以从行业解决方案模板创建，也可以自定义创建

简单易用
用户只需要指定训练数据和测试数据，即可自动建立模型

行业解决方案库：

自动特征工程
针对不同行业特点，定制化的自动特征工程

自动模型优化
针对不同行业特点，自动选择模型，自动模型调优

通用解决方案
除了特定行业解决方案，也支持从通用解决方案模板创建项目

行业解决方案模板：

特征探查
特征数据预览，特征类型查看，统计特性展示，数据分布可视化

设置简单
可以采用模板默认参数，也支持自定义参数，向模型引入专家知识

可视化模型评估
模型评估结果可视化，提供多种指标衡量模型效果

6.3.3 图分析 HyperGraph

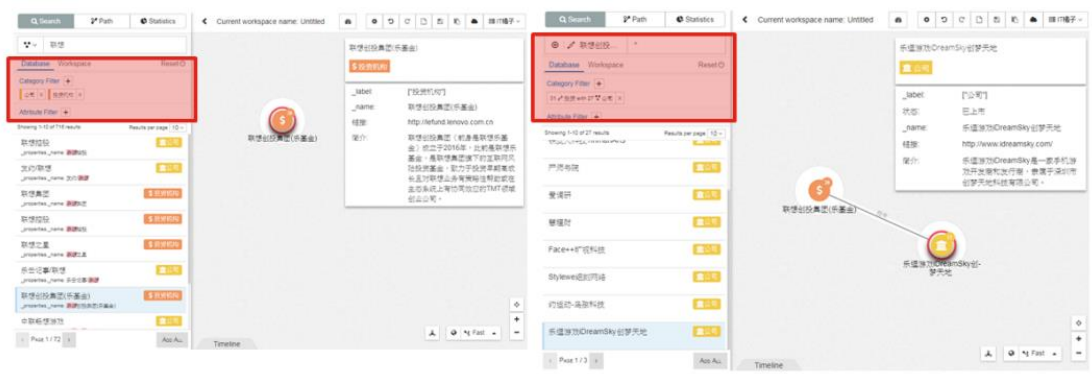
交互式查询及探索分析：



多元化搜索功能

实体/关系搜索

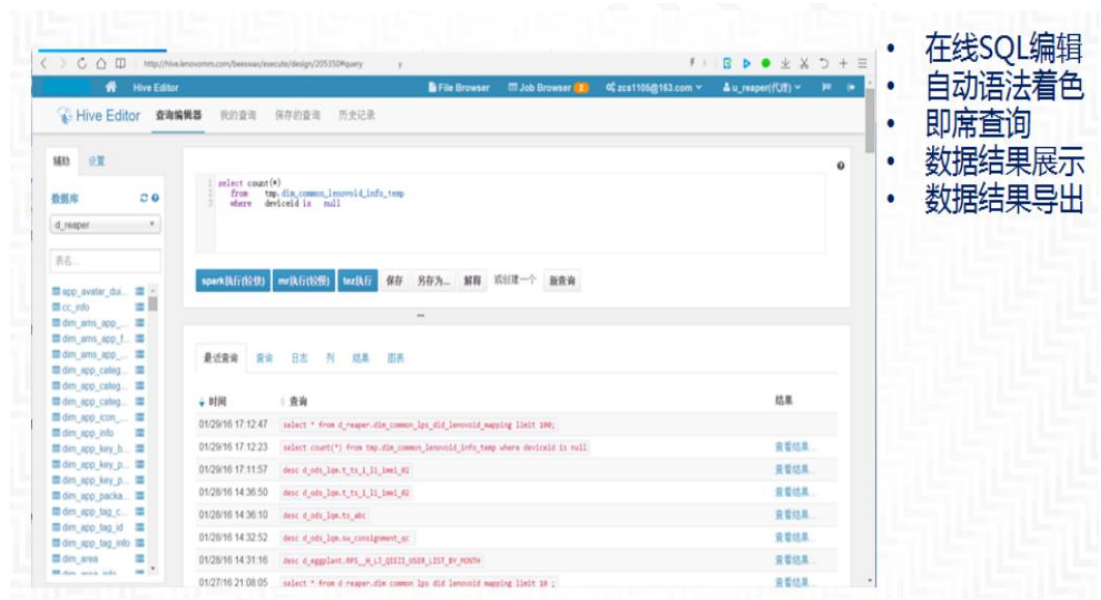
- 按用户提供的关键字进行精确/模糊搜索，并支持条件筛选缩小候选元素空间。
- 通过已有的实体进行邻居搜索，并支持对邻居实体进行条件过滤以精确定位感兴趣的元素。



6.3.4 自助查询分析器 SQL Editor

主要适用场景：

- 有初步 SQL 语法能力的业务分析人员进行自助式灵活查询。



数据查询由数据源管理和 SQL 编辑显示两部分组成。用户可管理的数据源包括所在项目的 hive 数据、自行添加的数据源 (MySQL、SQL Server、Oracle、DB2、phoenix、PostgreSQL)。

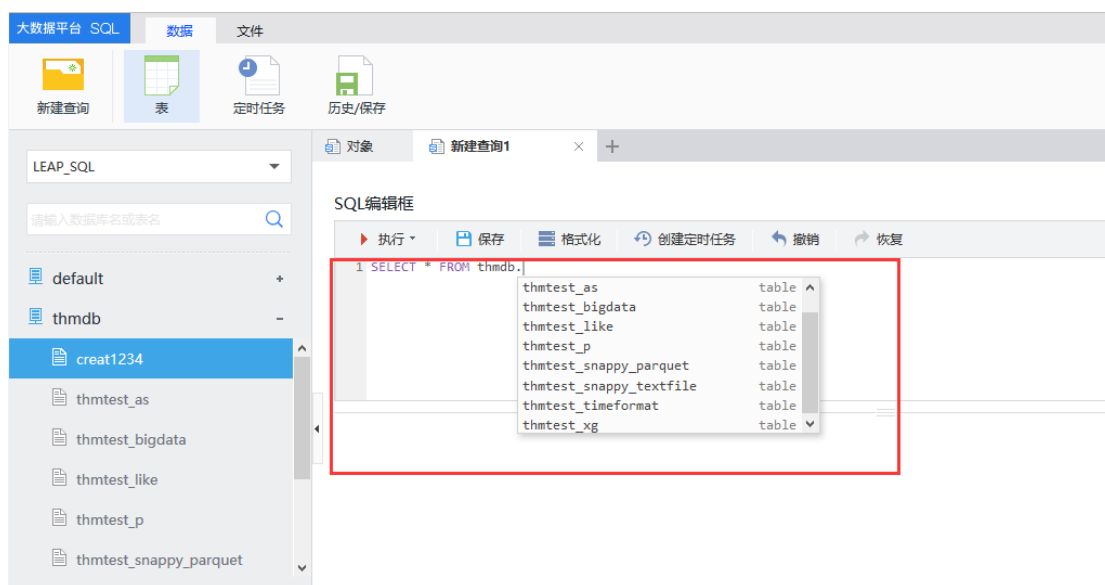


图、数据源

显示数据库、表。



支持在 SQL 编辑框内输入各类型的 SQL 语句，SQL 编辑框提供 SQL 关键字自动补全功能，在进行输入时，会出现相应的关键字提醒，例如在 from 等关键字后输入关键词将会提示相应的库、表；在库后面输入小数点时会提示当前库下的表；出现 SQL 关键字提示时，可通过鼠标点击或按 Enter 键选中。



● 数据源管理

- 1) 获取数据库元数据信息，可以展开并快速检索表信息；
- 2) 添加或删除数据源，支持 Hive、MySQL，Spark 等，并测试数据源连

通性；

- 3) 选择数据表，可以展示单表表结构字段信息，可以快速检索表并修改表结构、删除表。

- SQL 编辑

- 1) 提供成熟的 SQL 编辑器功能，支持语法补全、自动着色、关键字补全、数据库表提示、SQL 格式化，保存，撤销，恢复等操作；
- 2) 支持 SQL 2003 标准，兼容 SqlServer/Oracle 语法，支持存储过程、函数，支持 TPC-DS 测试集 99 个 SQL 语句；
- 3) 提供 MPP 架构数据库，支持海量数据高性能存储及 CRUD 操作；
- 4) 可选择 MapReduce、Impala 和 Spark 等不同执行方式；
- 5) 支持基于 Spark 的 CRUD 操作；
- 6) 查询结果通过交叉表进行展示，支持柱图、线图、饼图，并允许用户自行切换与配置，并且支持以 CSV 格式导出数据；
- 7) 支持多窗口化查询，可同时新建多个窗口，每个窗口进行不同的 SQL 语句查询，方便数据对比和查看；
- 8) 支持输出 SQL 执行日志；
- 9) 执行的 SQL 可以保存，并通过已保存的查询或历史查询快速调出。

定时任务

SQL 查询提供定时任务功能，具体功能如下：

- 1) 支持创建按照小时、日、周、月等时间维度的调度任务，自动化的完成用户的定制化需求；

- 2) 支持查看任务列表，对定制任务执行修改，删除，禁止，启动，重跑，查看日志等操作；
- 3) 支持查看整个集群任务执行历史数据，可查看当前计算任务执行状态，如排队、正在执行、错误、完成等。
- 4) 支持将定时任务记录进行保存，方便二次调度使用。

创建定时任务

序号	创建人	任务名称	下次执行时间	调度周期	状态	操作	执行历史
5	thvntest	impala_多源数据执行	2017-03-09 14:30:00	按月调度	启用		
6	thvntest	spark_load_local	2017-03-09 14:10:00	按小时调度	启用		
4	thvntest	insert_directoty_hdfs	2017-02-16 14:05:00	按周调度	启用		
3	thvntest	insert_directoty_nohdfs	2017-02-10 14:05:00	按天调度	启用		
2	thvntest	LOAD_LOCAL	2017-02-11 10:50:00	按天调度	启用		
1	thvntest	tab_to_tab	2017-02-10 11:45:00	按小时调度	启用		

定时任务列表

2. 完善的文件管理功能：

(1) 支持对项目下的用户组目录和用户私有目录进行管理；

(2) 支持文件和文件夹的移动、复制、删除、重命名，支持文件的下载和上传，支持新建文件夹；



“Spark_R” 的功能支持

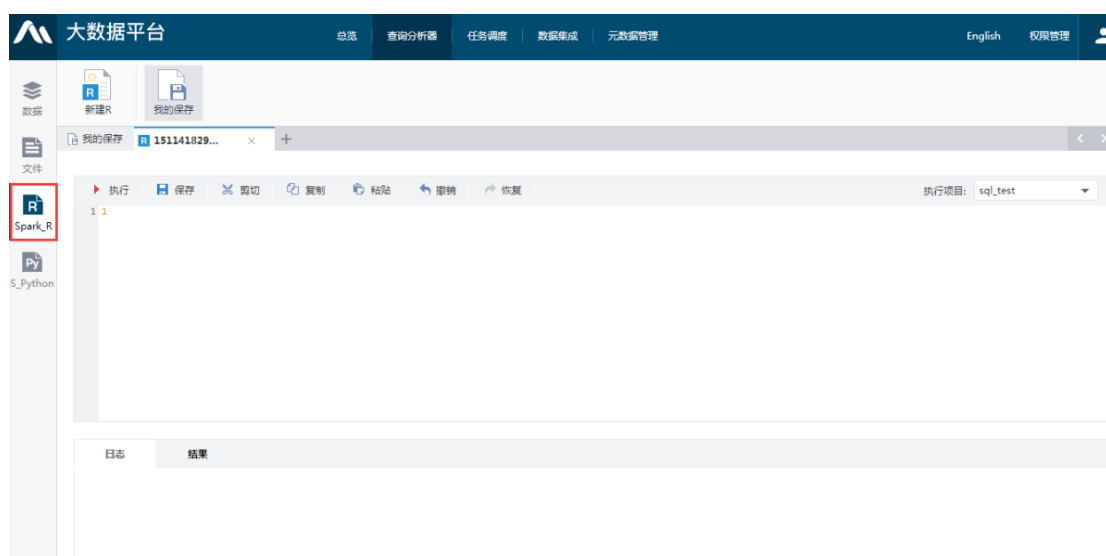
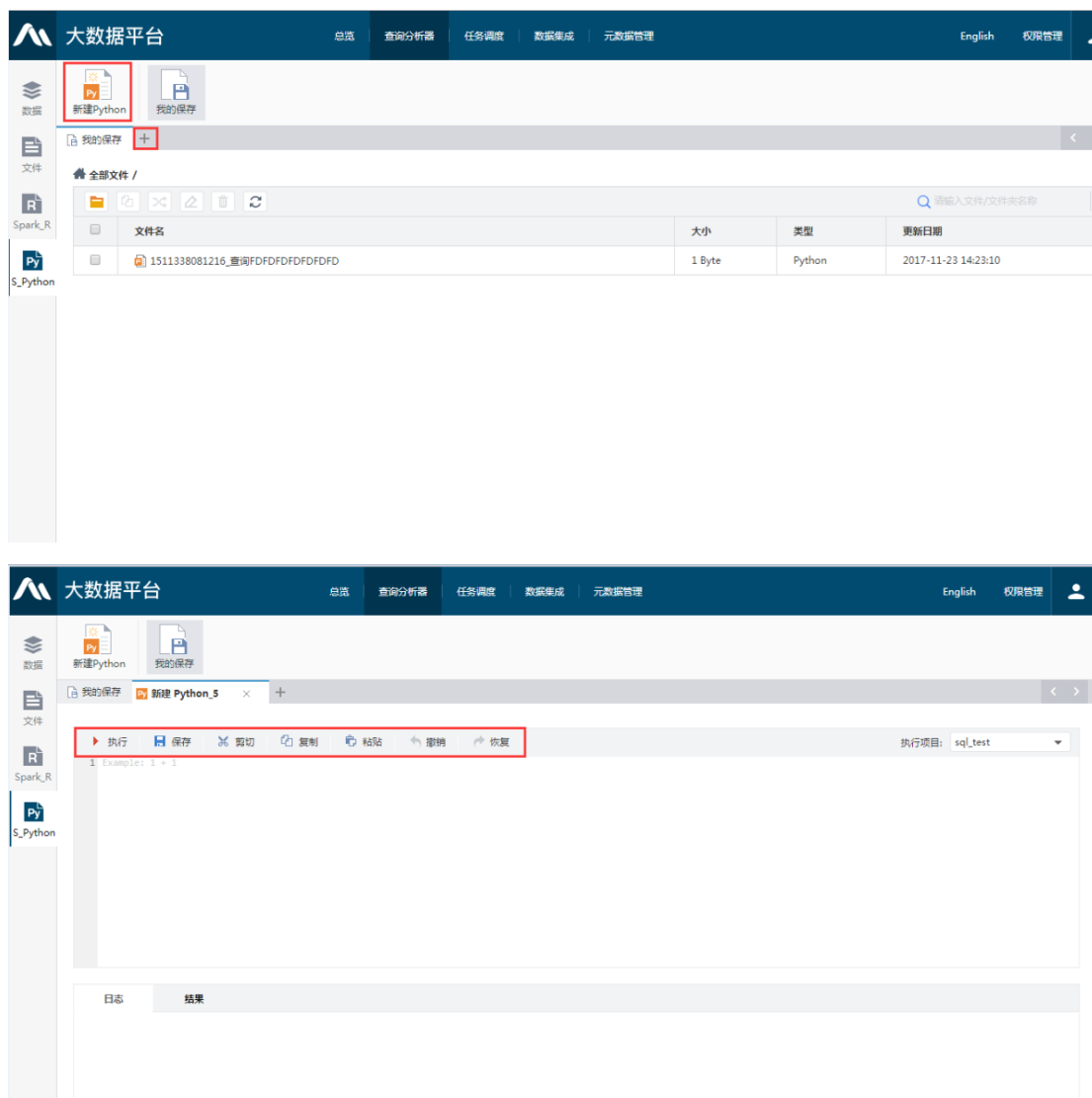


图 “Spark_R”模块

Python 功能支持：

支持新建 Python 语言编辑窗口。在编辑窗口内提供了执行、保存、剪切、复制、粘贴、撤销、恢复等工具供使用。



6.4 应用场景

1、深度学习 DeepNEX

主要适用机构：

- 高校/科研单位：深度学习实验室 / 机器学习实验室 / 人工智能实验室；
- 企业：视觉/语音/自然语言处理；

主要适用场景：

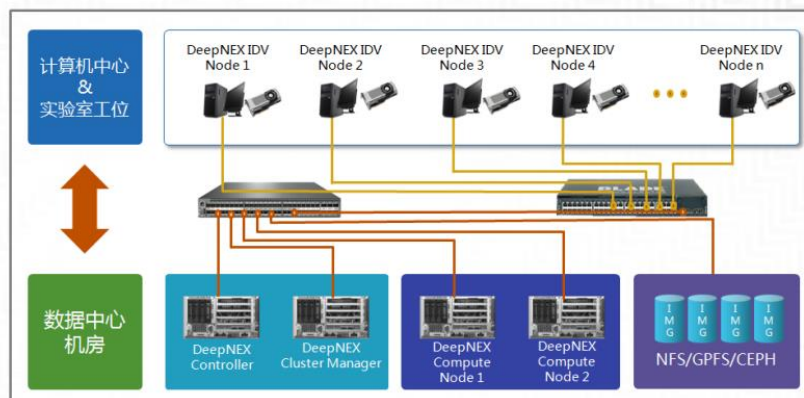
- 机器学习、深度学习、人工智能、图像识别、语音识别、视频搜索、问

答系统等方面。

某高校深度学习科教平台：

客户需求背景：

- 科研经费有限，无法配备足够的GPU工作站集群
- 希望低成本建设一套可共享、多租户的可调度GPU的平台
- 对深度学习有研究渴望，希望找合作共建实验室。



客户收益：

- 全球首创深度学习共享一体机，有限预算下实现更高资源利用率，支持深度学习实验室数十人共享使用。
- 将部分历史闲置的图站改造成深度学习的工位，也可远程供调用，极大地节省了成本。
- 封装了多种深度学习框架和丰富的demo应用案例，全面支持学生的课程学习与实验实训。
- 与客户战略合作，成立联合深度学习实验室。共享资源，联想大数据科学家担任客座讲师。

2、图分析 HyperGraph

可以将众多行业日益增长的数据建模成知识图谱，包括金融行的交易数据、制造业贸易数据、公安刑侦数据、医疗数据等。

应用场景举例：

● 金融分析部门

通过可视化的方式发掘企业上下游关系、企业股东关联关系；

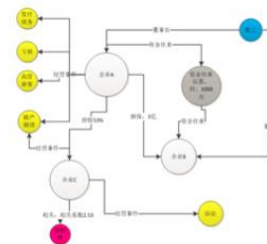
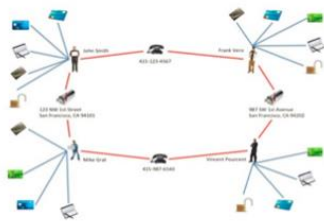
利用企业内部的数据来识别和根除潜在的洗钱、欺诈等犯罪模式；

● 企业数据管理部门

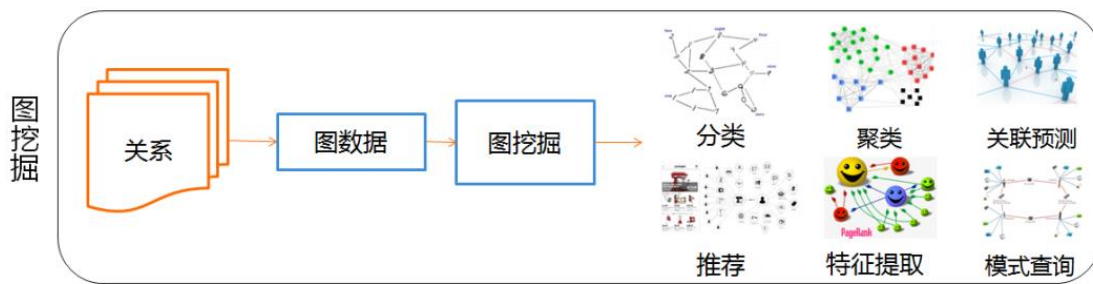
更好地管理企业内部的数据，及时发现隐藏的数据风险；

● 调查刑侦部门

通过有效整合跨系统跨平台的多维信息，构造人物时间的社会化网络进行侦查，分析和延判；

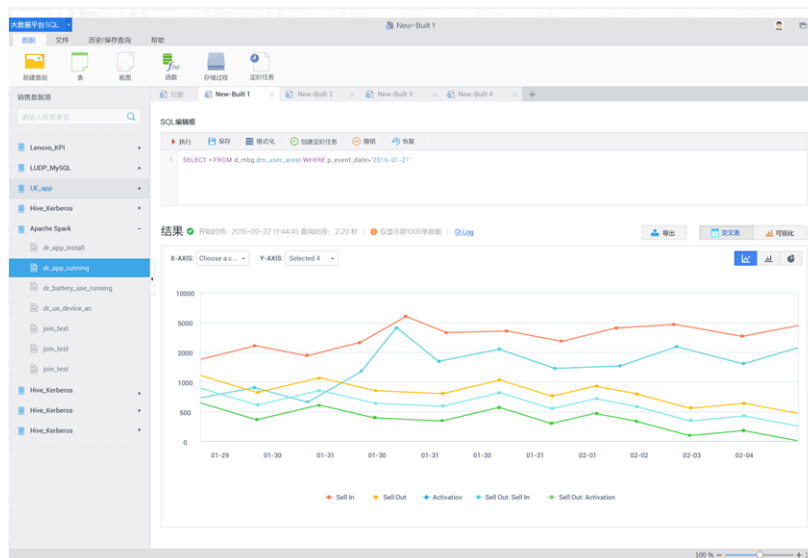


大规模图数据挖掘及预测



自助查询分析器 SQL Editor 主要适用场景：

- 具有初步 SQL 语法能力的业务分析人员进行自助式灵活查询。



查询分析：

提供基于web的即席查询工具，可以灵活地自助查询

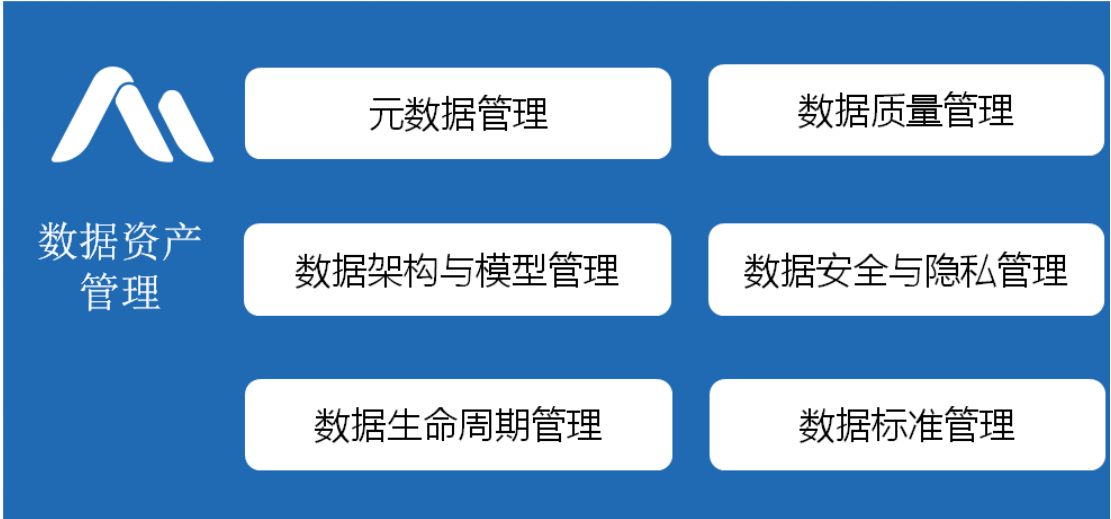
- ✓ 在线SQL编辑，SQL自动补全功能；
- ✓ 自动语法着色；
- ✓ 即席查询与可视化分析；
- ✓ 数据结果展示与导出；
- ✓ 大数据文件和数据管理；

7 数据治理平台

7.1 总体介绍

数据资产管理平台 Governance 将数据对象作为一种全新的资产形态，围

绕数据资产本身建立一个可靠可信的管理机制，提供数据标准管理、数据生命周期管理、元数据管理、数据质量管理、数据安全性与隐私管理等功能，为数据管理人员、运维人员、业务人员和应用开发者提供全方位服务与支撑。



图、汇通国信数据资产管理平台 Governance 功能架构

主要功能：

功能	功能描述
元数据管理	元数据管理是需要将各系统的信息、设计工具信息、生产平台信息，进行收集并统一管理。提供一个视图，以帮助使用人员了解系统的数据分布、数据关系、业务规则、指标口径等。元数据包括：业务类元数据、技术类元数据。 基于完整的元数据管理能力，元数据管理提供血缘分析功能。可以以某个元数据为终止节点，以图形化方式展示前端与其有关系的所有元数据，反应数据的来源与加工过程。从而可判断数据来源，定位数据质量问题。 系统具备影响分析能力。可以以某

	个元数据为起始节点,以图形化方式展示后端与其有关的所有元数据关系,反应数据的流向与加工过程。从而可以判断数据流向,定位数据转换中的错误。
数据标准管理	企业级数据标准主要解决系统之间数据集成的标准,解决跨部门数据标准管理 数据分析时数据一致理解。标准化需要解决的关键问题:标准制定,标准的执行和控制,标准化问题管理。
数据质量管理	数据质量管理针对数据传输加工过程提供波动检查,平衡性检查,阈值检查,有效性,一致性核查,并根据检查结果提供可选的告警形势,比如邮件推送等。提供数据质量稽核规则统一配置,并数据质量管理 支持对质量规则的定义和任意节点的附加。 根据时间发生的顺序,可分为事前质量检查、事中运行监控、事后归纳总结,并能对各过程中所产生事件及信息形成告警信息,通过短信、彩信、邮件的形式进行发送。
数据安全和隐私保护	提供根据设置的安全策略进行访问控制,从而限制用户对数据库数据安全和隐私保 具体文件、目录和数据库表的访问;提供对数据中个人身份信息护和敏感信息,通过脱敏规则进行数据的变形,实现敏感隐私数据的可靠保护,即数据脱敏技术。
数据生命周期管理	对数据从上线、使用到下线整个生命周期的使用情

	况进行跟踪、数据生命周期管理 记录和操作。功能包括：数据生命周期判定原则定义、生命周期查询/更改、数据生命周期监控、过期数据销毁等。
--	--

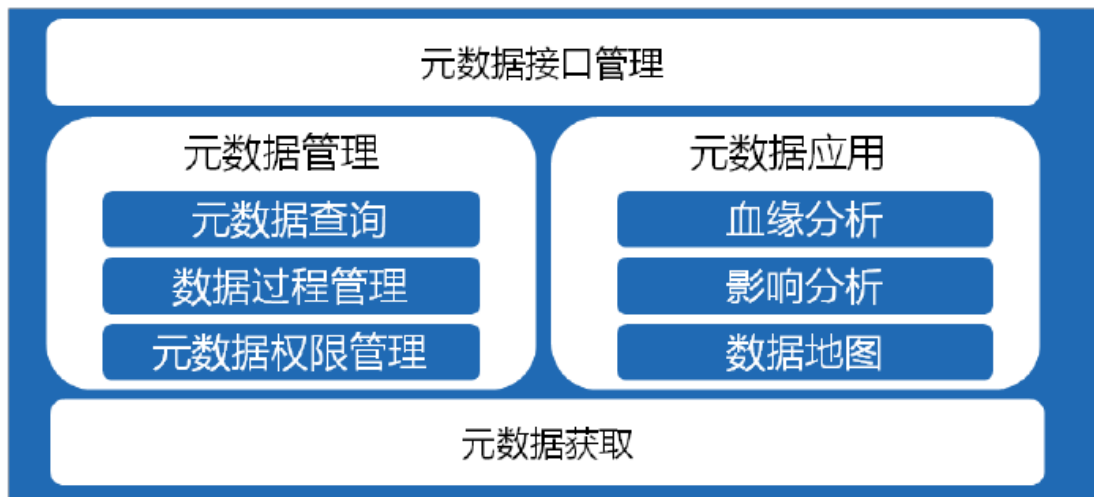
产品整体特性：

1. 支持 Hive，HDFS 的元数据自动获取及全量更新和增量更新的方式。
2. **性能高效**：10 分钟可同步系统单数据源支持 3 万+张表元数据的初始化。
获取；支持单次 10 万行文本数据的导入和导出。
3. **支持数据质量智能评估**、自定义检查规则及自定义告警；提供文件级、记录级、数据关联层面多种数据质量核查机制。
4. 支持全链路的数据处理流程血缘追踪及影响分析。
5. 支持多层级多角色的数据权限模型管理元数据，以及敏感数据过滤及主动探测隐藏。
6. 支持数据操作安全审计及日志分析，主动拦截非法操作。

7.2 产品功能

7.2.1 元数据管理

汇通国信 Metadata 元数据管理作为汇通国信企业级数据治理平台的重要组件，将各系统的信息、设计工具信息、生产平台信息，进行收集并统一管理。提供一个视图，以帮助使用人员了解系统的数据分布、数据关系、业务规则、指标口径等。



汇通国信元数据管理组件可帮助完成从元数据获取到元数据管理与元数据应用的全生命周期管理,并通过元数据接口管理为客户提供多种数据接口方式,便于客户使用元数据。

7.2.1.1 主要功能

功能	功能细项	功能描述
数据集视图	物理视图	系统读取数据展示层级,按照‘数据源-数据库-数据表的层级展示’;物理视图可按照每一层级查看下级列表数据大小等基本信息;
	业务视图	用户根据具体的业务逻辑创建的各个层级的数据主题集合;
	业务视图管理	提供用户创建,删除,修改各个层级的主题及对应的主题下新增及移除元数据;

元数据管理	属性信息	展示和编辑数据表的基本信息(中文名称、数据负责人等)、来源信息、存储信息
	数据结构	展示数据表字段信息和分区信息
	样本数据	展示数据表所存放的样本数据
	变更历史	展示数据表所发生的变更记录
	权限信息	展示用户对该表的可操作权限
	血缘关系	展示数据表的向上、向下关系
	注释信息	展示及编辑数据表的应用举例、适用场景、描述和数据质量问题等信息
元数据搜索查询	搜索查询	可按照数据表的表名、中文名称、数据描述、字段名、字段描述为关键字搜索

7.2.1.2 产品特性

- 1、支持 Hive, HDFS 的元数据自动获取及全量更新和增量更新的方式;
- 2、性能高效: 10 分钟可同步系统单数据源支持 3 万+张表元数据的初始化获取; 支持单次 10 万行文本数据的导入和导出。
- 3、支持全链路的数据处理流程血缘追踪及影响分析;

7.2.2 数据质量管理

企业数据质量通常存在如下问题:

- 1) 稽核规则没有生效，数据质量风险难以及时监控发现；
- 2) 现有实际生效的稽核规则散落在各层的加工代码中，未成体系；

数据质量管理的总体目标：常态化、体系化、标准化、自动化的全面数据质量管理，以达到数据质量控制的全面性、可控性、可度量性、可迅速定位和有效解决。通过流程制度建设、质量评估体系建设、质量检测 IT 平台建设实现 TDQM(Total Data Quality Manage)，即企业级全面数据质量管理。

数据质量管理针对该形势提供波动检查，平衡性检查，阈值检查，有效性，一致性核查，并根据检查结果提供可选的告警形势，比如邮件推送等。

提供数据质量稽核规则统一配置，并支持对质量规则的定义和任意节点的附加。

根据时间发生的顺序，可分为事前质量检查、事中运行监控、事后归纳总结，并能对各过程中所产生事件及信息形成告警信息，通过短信、彩信、邮件的形式进行发送。

7.2.3 数据生命周期管理

对数据从上线、使用到下线整个生命周期的使用情况进行跟踪、记录和操作。功能包括：数据生命周期判定原则定义、生命周期查询/更改、数据生命周期监控、过期数据销毁等。

功能	功能细项	功能描述
生命周期	设置管理	Hive 表支持对生命周期的设定，到期数据直

		接删除；支持文件时间和分区字段判断到期时间
搜索查询	搜索查询	可按照数据表的表名、中文名称、数据描述、字段名、字段描述为关键字搜索

7.2.4 数据的安全与隐私保护

数据资产管理平台可纳入原有的安全管理系统，运维人员对各主机数据库的访问均通过原有安全管理系统进行。运维人员需要访问主机数据库的数据时，需要首先登录原有安全管理系统门户，原有安全管理系统门户对用户进行身份鉴别和权限判别，确定用户的真实身份并判断是否有数据库的访问权限，如果没有权限则无法访问数据库的数据。身份鉴别和权限判别通过后，用户从原安全管理门户登录到要访问的数据库中，根据设置的安全策略进行访问控制，从而限制用户对数据库具体文件、目录和数据库表的访问。如限制用户只能对指定的数据库表执行查询操作，不能进行增删改操作。

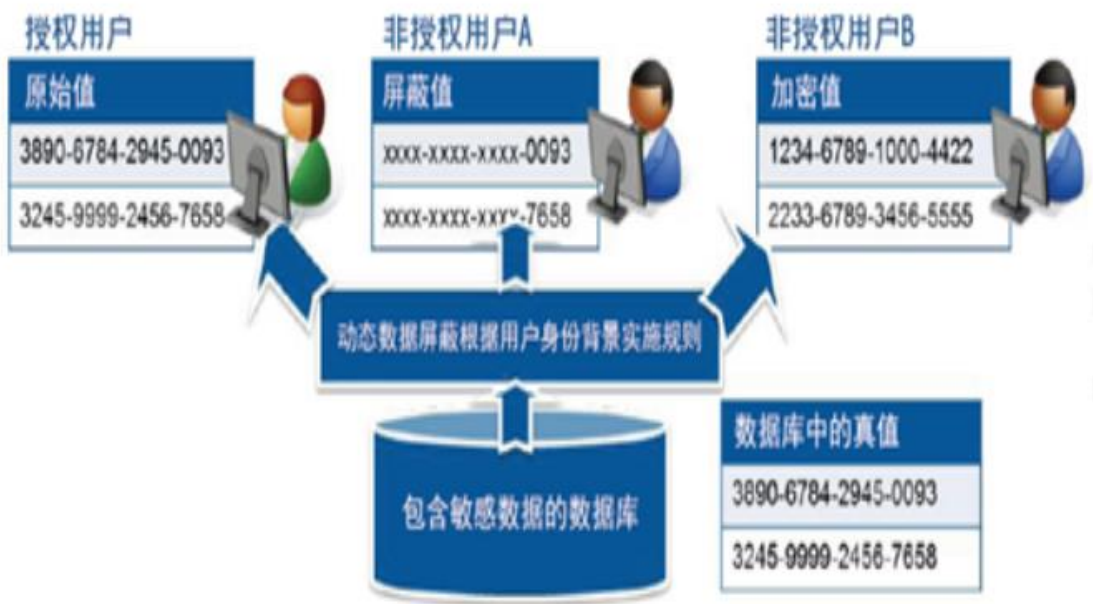
系统支持对目标数据的加密、脱敏能力。数据加密支持 AES128、MD5、DES 等主流加密算法，数据与密钥分离，并能支持用户自定义的加密算法。

在数据的应用迅速发展的同时，世界各国也日益增强了对个人识别信息 (PII) 的保护，不断出台数据监管和个人隐私保护的相关法律。为了保证数据平台的数据安全性，需要对不同层级数据访问者，如数据分析人员、数据管理人员、数据

测试人员等，在保证数据逻辑完整性的前提下采用差异化的数据呈现方式。这需要对数据中个人身份信息和敏感信息，通过脱敏规则进行数据的变形，实现敏感隐私数据的可靠保护，即数据脱敏技术。

模糊化的方式包括如下几种形式：

- 数据替换 - 以虚构数据代替真值。
- 截断、加密、隐藏或使之无效 - 以“无效”或 *****代替真值。
- 随机化 - 以随机数据代替真值。
- 偏移 - 通过随机移位改变数字数据。
- 字符子链屏蔽 - 为特定数据创建定制屏蔽。
- 限制返回行数 - 仅提供可用回应的一小分子集
- 基于其他参考信息进行屏蔽 - 根据预定义规则仅改变部分回应内容（例如屏蔽 VIP 客户姓名，但显示其他客户）



图、数据模糊化

上图展示了数据脱敏的运作方式，图中有三名用户系统数据进行访问。左侧

为授权用户，该用户获得授权，可查看全部详细信息。中间为非授权用户，该雇员仅被授权查看以屏蔽形式显示的数据，以便执行管理任务。右侧也是非授权用户，如开发人员、DBA 或生产支持人员，他们出于 IT 方面的用途而需要以相应格式显示的信息，因此数据安全模块将向他们提供加密过的数值。

7.3 功能展示

元数据展示的集合，提供“物理视图”和“业务视图”两种目录分类逻辑展示元数据。



查看数据库列表：
展示该数据源下所包含的数据库，展示的信息包括数据库名、表数量、存储大小、表新增数量、新增存储大小。

大数据平台

元数据 业务视图 物理视图 表管理 导入/导出

数据表 > 物理视图 > hive > Hive

LEAP-Hive (数据库总数: 8 | 数据库总大小: 100.54 MB | 总新增大小: 166B)

库名	表数量	存储大小	表新增数量	新增存储大小	操作
default	3	6.92 KB	0	0 B	查看更多 删除
h1	8	676B	0	0 B	查看更多 删除
h2	3	18B	0	0 B	查看更多 删除
hive_table	6	305B	1	166B	查看更多 删除
hytest	5	0 B	0	0 B	查看更多 删除
leapidtest	2	0 B	0	0 B	查看更多 删除
rpsql	8	100.53 MB	0	0 B	查看更多 删除
super	3	948B	0	0 B	查看更多 删除

查看数据表列表：

展示该数据库下所包含的数据表，显示的信息包括数据表名、大小、生命周期、描述、创建时间、最后数据时间、新增大小。

大数据平台

元数据 业务视图 物理视图 表管理

数据表 > 物理视图 > hive > Hive > rpsql

rpsql (所属项目: leasql | 数据库大小: 10.54 KB | 表总数: 17 | 新增大小: 10.49 KB)

表名	大小	最后数据时间	创建时间	生命周期	描述	新增大小	操作
rp1	0 B	2017-07-12 1...	2017-07-10 1...		描述30	0 B	查看更多
rp2	76B	2017-07-13 1...	2017-07-10 1...		描述29	76B	查看更多
student110	1.31 KB	2017-07-14 1...	2017-07-14 1...		描述28	1.31 KB	查看更多
student120	1.31 KB	2017-07-14 1...	2017-07-14 1...		描述27	1.31 KB	查看更多
t001	3.43 KB	2017-07-11 1...	2017-07-11 1...		描述26	3.43 KB	查看更多
t002	3.44 KB	2017-07-14 1...	2017-07-14 1...		描述25	3.44 KB	查看更多
t1	56B	2017-07-14 1...	2017-07-14 1...		描述41	56B	查看更多
t10	44B	2017-07-14 1...	2017-07-14 1...		描述40	44B	查看更多
t11	688B	2017-07-18 1...	2017-07-14 1...		描述39	640B	查看更多
t2	20B	2017-07-14 1...	2017-07-14 1...		描述38	20B	查看更多

查看属性信息

“属性信息”选项卡中展示数据表的‘基本信息’、‘来源信息’和‘更多元数据信息’等。

数据集 > 业务视图 > shuo_de ded123 > 123 > 321 > game_by_db2

申请权限★收藏

属性信息

数据结构

样本数据

变更历史

权限信息

血缘关系

注释信息

基本信息

名称	属性
表名	game_by_db2
别名	用户表
数据描述	用于存储APP用户数据
业务意义	可用于用户行为分析
业务负责人	yangyang
是否敏感数据	是
所属主题路径	shuo_de ded123/123/321
数据来源	APP数据采集系统
权限状态	读 写
数据更新频率	按天
数据状态	正常

“数据结构”展示该表的字段信息和分区信息，支持用户对表的字段描述信息进行编辑。

数据集 > 物理视图 > hive > Hive > default > book

申请权限★收藏

属性信息

数据结构

样本数据

变更历史

权限信息

血缘关系

注释信息

字段信息

字段	数据类型	字段描述
name	string	双击输入
age	string	双击输入
birthday	string	双击输入
gender	string	双击输入
weighth	string	双击输入

分区信息

名称	属性
分区字段	weighth,gender

“变更历史”：显示数据表所发生的变更情况详情。变更类型包含三种：新增字段、删除字段、字段类型变更。

数据集 > 物理视图 > hive > Hive > hyy > thmtest_p2

申请权限★收藏

属性信息

数据结构

样本数据

变更历史

权限信息

血缘关系

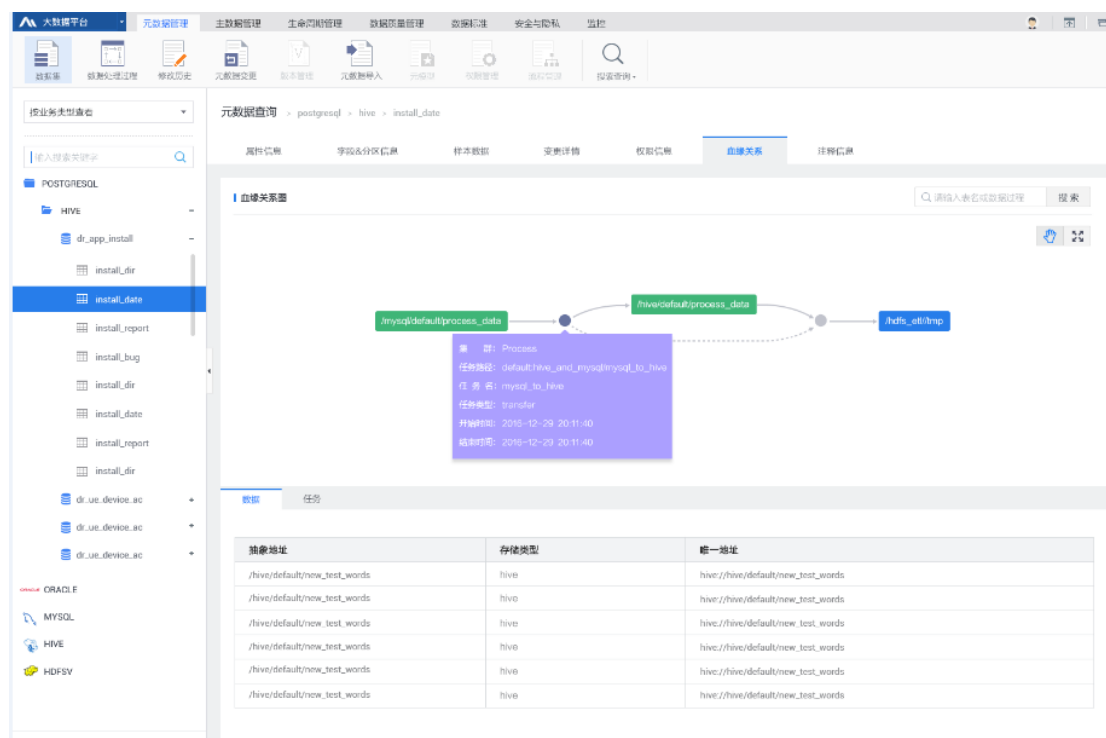
注释信息

字段变更

序号	变更时间	变更	变更详情
1	2017-07-21 14:17:38	新增字段	新增 gmrt_create(int)
2	2017-07-21 14:17:38	新增字段	新增 system_time(string)
3	2017-07-21 14:17:38	字段类型变更	app_key(string) 变为 app_key(int)
4	2017-07-21 14:17:38	删除字段	删除 server_time(string)

血缘分析

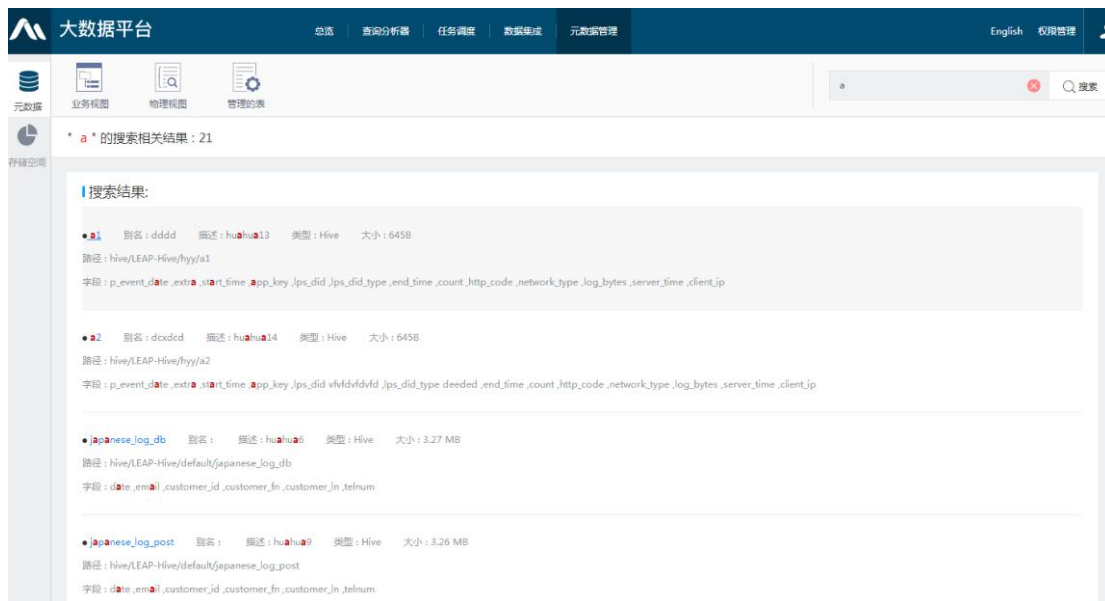
基于完整的元数据管理能力，提供血缘分析功能。可以以某个元数据为终止节点，以图形化方式展示前端与其有关系的所有元数据，反应数据的来源与加工过程。从而可判断数据来源，定位数据质量问题。



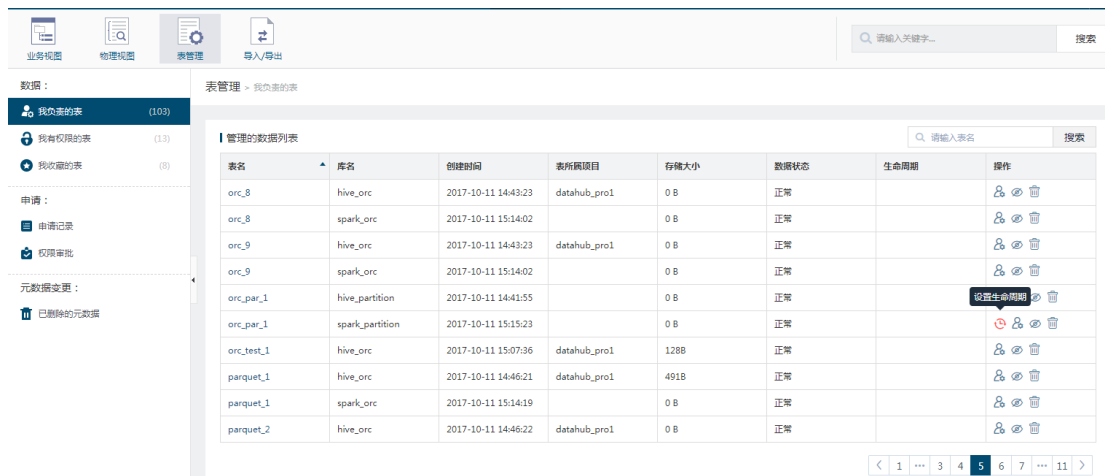
- 数据表的信息：存储类型、库名、表名、当前表的上游表数量、当前表的下游表数量。
- 处理过程的信息：处理工具类型、任务名称、任务类型、所属应用、创建人、流程 ID、开始时间、结束时间。

全局搜索

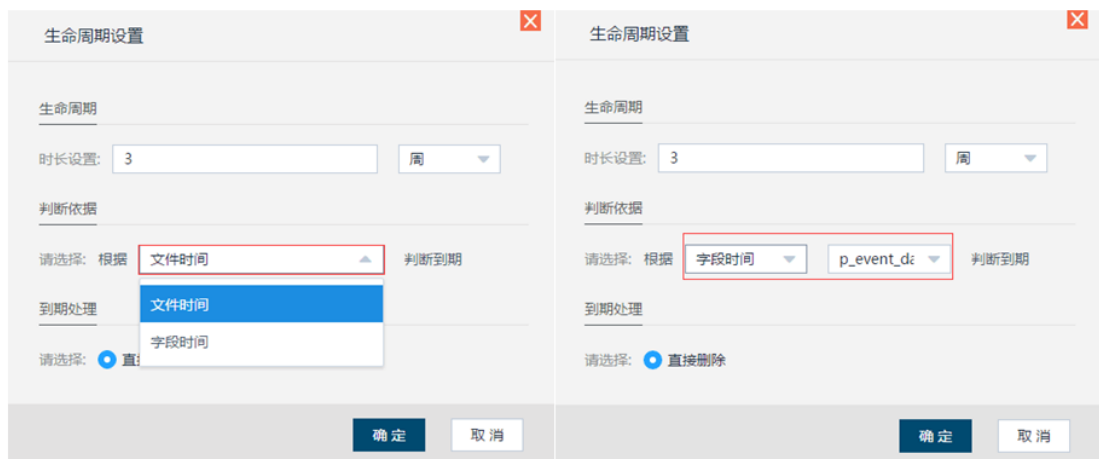
- 可以根据表名、别名、数据描述、字段名称、字段描述查询系统内的数据；
- 支持多关键字搜索
- 默认显示最近 10 条搜索记录，通过点击搜索记录可查看搜索结果；



数据生命周期设置入口：



- 数据保留时间，时长设置：支持按天，周，月，年和永久设置
- 到期判断依据：默认按照文件时间判断到期。
- 到期处理：根据设置的时间，到期的数据会直接删除。



7.4 应用场景

可应用于企业数据梳理、数据治理等场景。

企业在多年的业务发展和经营中，积累了大量的各类数据。但在数据管理和内部决策分析时，面临着相关问题：

1、数据使用有问题：

- 数据量庞大，系统繁多，数据结构复杂，不清楚各类数据的细节管理信息；
- 跨平台的数据整合过程中，遇到数据内容与格式不统一等数据质量问题；
- 数据仓库/大数据分析系统存在一些数据质量问题，不能满足业务发展要求；

2、内部管理决策有需求

- 需要有效地整合利用已经积累的数据信息，为建立企业级统一的业务数据视图，奠定坚实的基础；
- 基于业务部门对数据问题提出的整治需求，满足业务部门深层次分析的要求。

客户希望采取相应的措施和工具来提升数据质量，保证数据能被正确的认识、理解和使用，避免冗余；为共享数据建立决策规则和流程；提升数据资产的价值；提供解决数据问题的机制；促进 IT 和非 IT 人员共同参与决策。通过构建数据治理系统，实现数据标准管理、元数据管理、数据质量管理、数据生命周期管理。



数据标准管理：数据标准管理组织协调者组织数据提供者和执行者参与数据标准属性的收集和整理工作，并按照协商一致的原则形成数据标准初稿，按照数据标准体系分类的要求进行数据梳理。

元数据管理：整理和描述业务术语、信息分类、指标定义（指标口径）、业务规则等信息；针对大数据平台的物理模型、逻辑模型、数据处理过程、接口单元、指标和维度编码等主要实体进行精确定义。

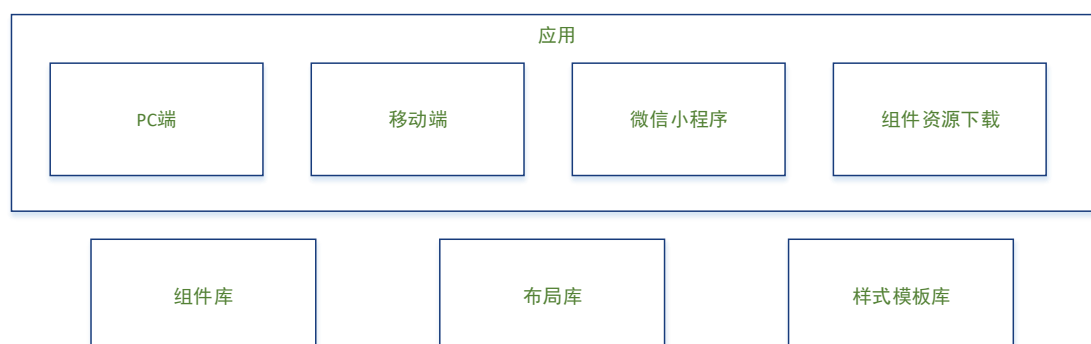
数据质量管理：通过获取大数据平台重点环节的数据质量监控信息，结合元数据库中的有关检查规则，对数据质量情况进行判断，并及时向数据质量监控人员报告。

通过相关的管理流程和系统功能应用，逐步使企业数据中心的数据定义清晰、使用明确，质量可靠，可查可管理，提升了数据资产的价值。

8 前端开发框架

8.1 总体介绍

前端开发框架是一个前端资源 PaaS 平台，为应用开发者提供丰富的前端资源，包括常用的 UI 组件、JS 插件及基础模版，开发者可以通过前端开发框架快速构建页面。它还提供多种样式风格模板供开发者选择，开发者也可以在线拖拽组件组装页面，让前端布局可视化。



8.2 主要功能

8.2.1 PC 组件库

PC 端组件库包括布局组件、导航组件、页面组件、表格组件、按钮组件、表单组件、JS 插件、图表组件、常用模块等。开发人员在开发中可直接将组件拖拉到布局中，或直接 copy 组件代码到对应位置即可在页面指定位置构建指定组件。

PC端组件库						
布局 Layout	导航 Navigation	组件 Components	JS插件 javascript插件	图表 Charts	更多	常用模块 Commonly
两栏布局	横向导航	进度条	弹层 new	饼状图		401提示
三栏布局	纵向导航 new	分页	树形表	区域图 new		403提示
网格布局		流程	拖拉组件 new	柱状图		500提示一
通用布局 new		选项卡	折叠框	折线图 new		500提示二
	表单 Form 更多	时间轴	文件上传	漏斗图		登录异常一
表格 Table 更多	水平布局表单	数据展示	富文本编辑器 new			登录异常二
基础表格	垂直布局表单	加载动画	自定义滚动条			功能开发中
带操作按钮表格	内联布局表单	消息状态	状态开关			服务故障
锁定头部表格	表单验证 new	列表	文本复制			
带分页表格(DataTables)	数字输入框	图片	横向拓扑图			
带排序搜索表格(Kendo UI)	下拉选框	图标	纵向拓扑图			
可变列宽表格(colsize)	autoComplete	面包屑	S型拓扑图			
带导出表格(Tableexport)	时间日期	提示说明	日期处理			
表格编辑(tabedit) new	复选框		DataFlow			
	标签输入框	面板 Panel	DataFlow2			
按钮 Button 更多	日期范围选择器	基础面板	json格式化			
基础按钮	数据选择器 new	面板组合一	页面局部刷新			
按钮组		面板组合二	用户指引			
			Tree Multiselect new			
			markdown解析 new			

8.2.2 移动端组件库

移动端组件库包括按钮、基础组件、弹层组件、列表组件、JS 插件、页面组件、图表组件、提示组件等。开发人员在开发中可直接将组件拖拉到布局中，或直接 copy 组件代码到对应位置即可在页面指定位置构建指定组件。

移动端组件库			
按钮 Button	基础类 Base	弹层 Dialog	JS插件 Plugin
基础按钮	文本	消息通知	滚动刷新
按钮组	面板	加载动画	消息轮播
	表单	对话框	手势划屏
			滚动固定头部表格
列表 List	组件类 Component	反馈提示 Tip	图表 Chart
基础列表	选项卡	空数据提示 new	饼状图
带信息列表	头部菜单	结果提示 new	条形图
带链接列表	导航栏	异常提示 new	线性图
带图文列表	侧边栏		雷达图
带操作列表	搜索		漏斗图
带图标列表 new	流程 new		
	下拉菜单 new		

8.2.3 微信小程序组件库

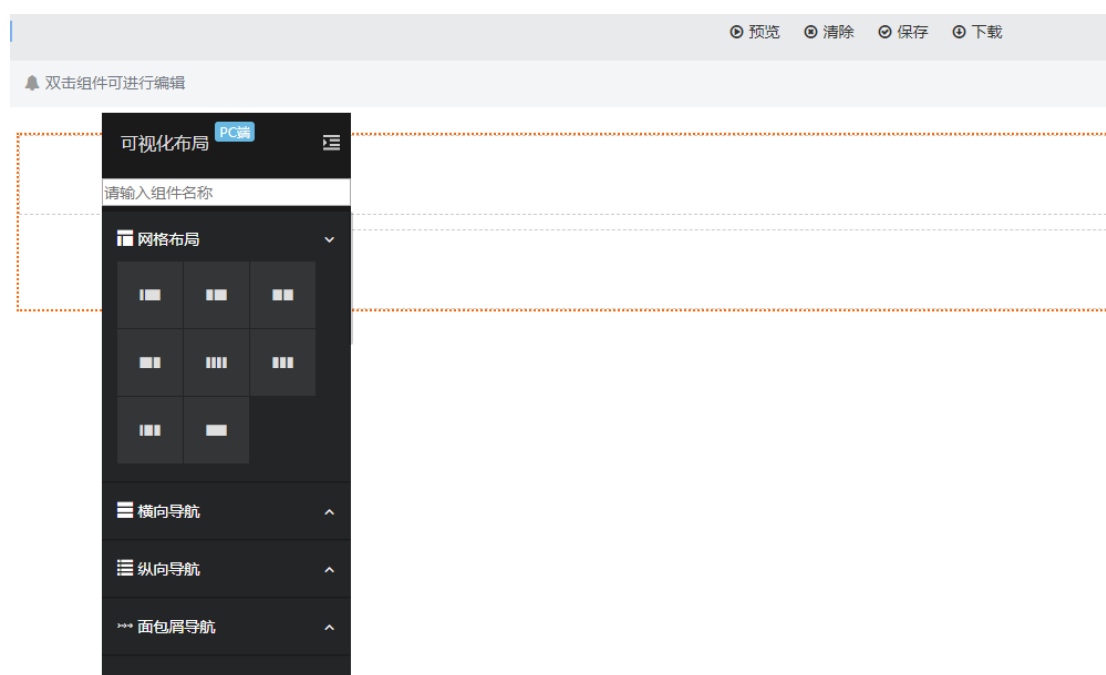
微信小程序组件库包括头部菜单、面板、文本、按钮、搜索、选项卡、底部导航等组件。

8.2.4 PC 端可视化布局

模拟 PC 端页面，提供可视化拖拽式布局，将组件库中的组件拖拉到指定位置进行展示。

可视化布局提供预览、清除、保存、下载等功能。

对布局中的组件，使用鼠标双击对组件展示内容进行编辑。



8.2.5 移动端可视布局

模拟移动端页面，提供可视化拖拽式布局，将组件库中的组件拖拉到指定位置进行展示。

可视化布局提供预览、清除、保存、下载等功能。

对布局中的组件，通过点击组件对组件展示内容进行编辑。



8.2.6 微信小程序可视化布局

微信小程序布局跟移动端相似，只是组件及展示风格按照微信小程序定制。

模拟移动端页面，提供可视化拖拽式布局，将组件库中的组件拖拉到指定位置进行展示。

可视化布局提供预览、清除、保存、下载等功能。

对布局中的组件，通过点击组件对组件展示内容进行编辑。



8.2.7 样式风格

前端开发框架将常用的样式风格内嵌到框架中，在进行前端开发时，只需进行样式风格选择，就能统一前端样式风格开发。



8.2.8 组件搜索

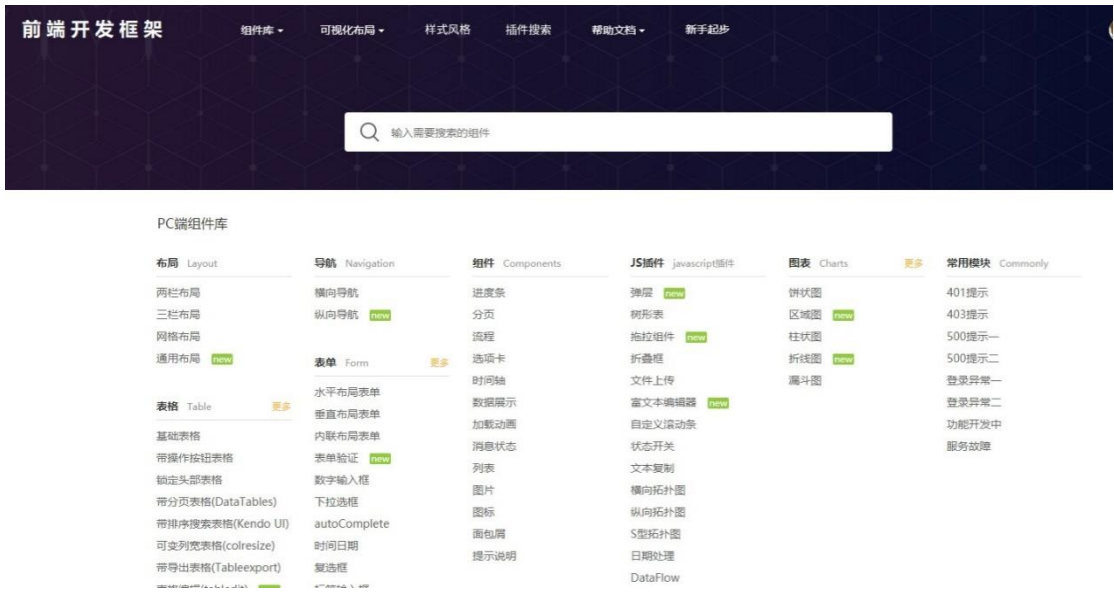
提供组件搜索功能，只要输入组件名称，就能将相应的组件搜索出来，在以列表的方式展示。通过点击组件，可进行组件下载和组件示例查看。

选择框			
插件名称	版本	简介	操作
☐ select2	3.5.2	下拉框美化插件chosen的扩展，它能让丑陋的、很长的select选择框变的更好看，支持搜索，远程数据集。	下载资源包 查看示例
兼容性: IE 6+, Firefox, Chrome 协议: MIT license 下载资源包: select2- 3.5.2.zip			

8.3 功能展示

8.3.1 前端开发框架

前端开发框架由组件库、可视化布局、样式风格、插件搜索、帮助文档、新手起步等组成。



8.3.2 组件库

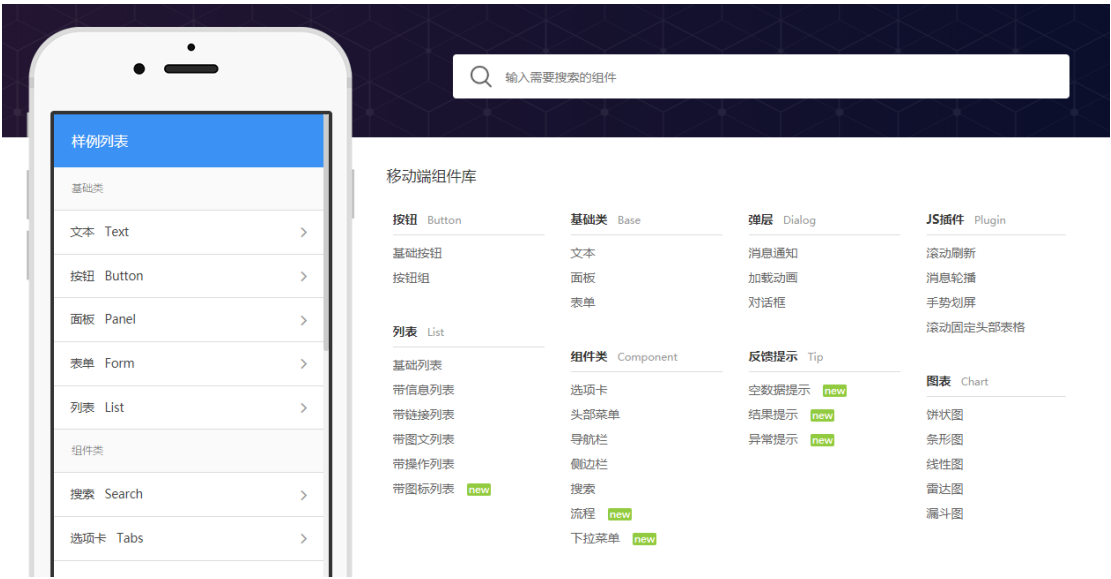
组件库由 PC 端组件库和移动端组件库组成。点击组件库菜单，弹出组件库菜单，PC 端组件库则进入 PC 组件库，点击移动端组件库，则进入移动端组件库。



8.3.2.1 PC 端组件库

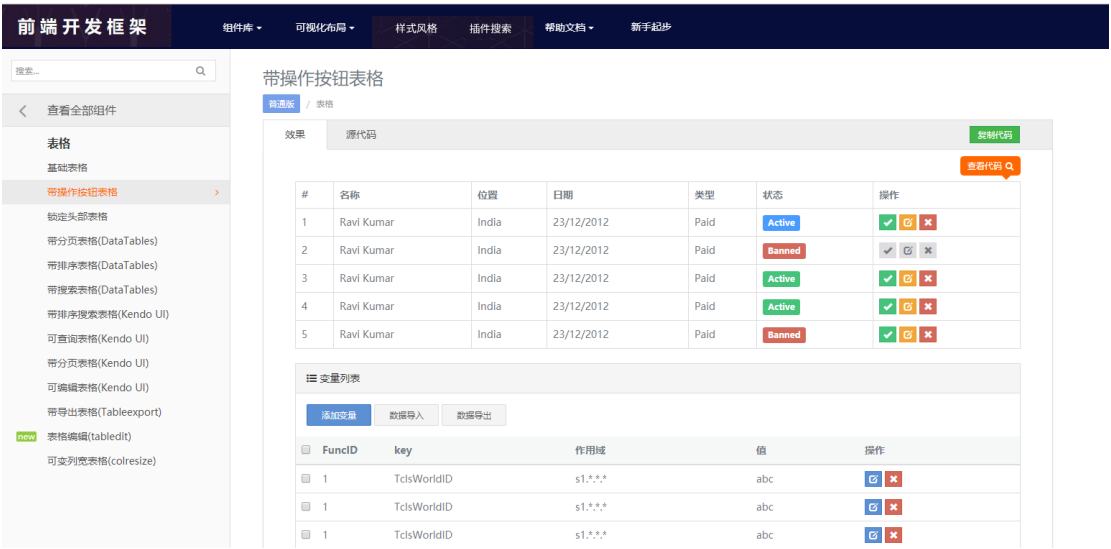
布局 Layout	导航 Navigation	组件 Components	JS插件 javascript插件	图表 Charts	更多	常用模块 Commonly
两栏布局	横向导航	进度条	弹层 new	饼状图		401提示
三栏布局	纵向导航 new	分页	树形表	区域图 new		403提示
网格布局		流程	拖拉组件 new	柱状图		500提示一
通用布局 new		选项卡	折叠框	折线图 new		500提示二
	表单 Form 更多	时间轴	文件上传	漏斗图		登录异常一
表格 Table 更多	水平布局表单	数据展示	富文本编辑器 new			登录异常二
基础表格	垂直布局表单	加载动画	自定义滚动条			功能开发中
带操作按钮表格	内联布局表单	消息状态	状态开关			服务故障
锁定头部表格	表单验证 new	列表	文本复制			
带分页表格(DataTables)	数字输入框	图片	横向拓扑图			
带排序搜索表格(Kendo UI)	下拉选框	图标	纵向拓扑图			
可变列宽表格(colresize)	autoComplete	面包屑	S型拓扑图			
带导出表格(Tableexport)	时间日期	提示说明	日期处理			
表格编辑(tabledit) new	复选框		DataFlow			
	标签输入框	面板 Panel	DataFlow2			
按钮 Button 更多	日期范围选择器	基础面板	json格式化			
基础按钮	数据选择器 new	面板组合一	页面局部刷新			
按钮组		面板组合二	用户指引			
			Tree Multiselect new			
			markdown解析 new			

8.3.2.2 移动端组件库

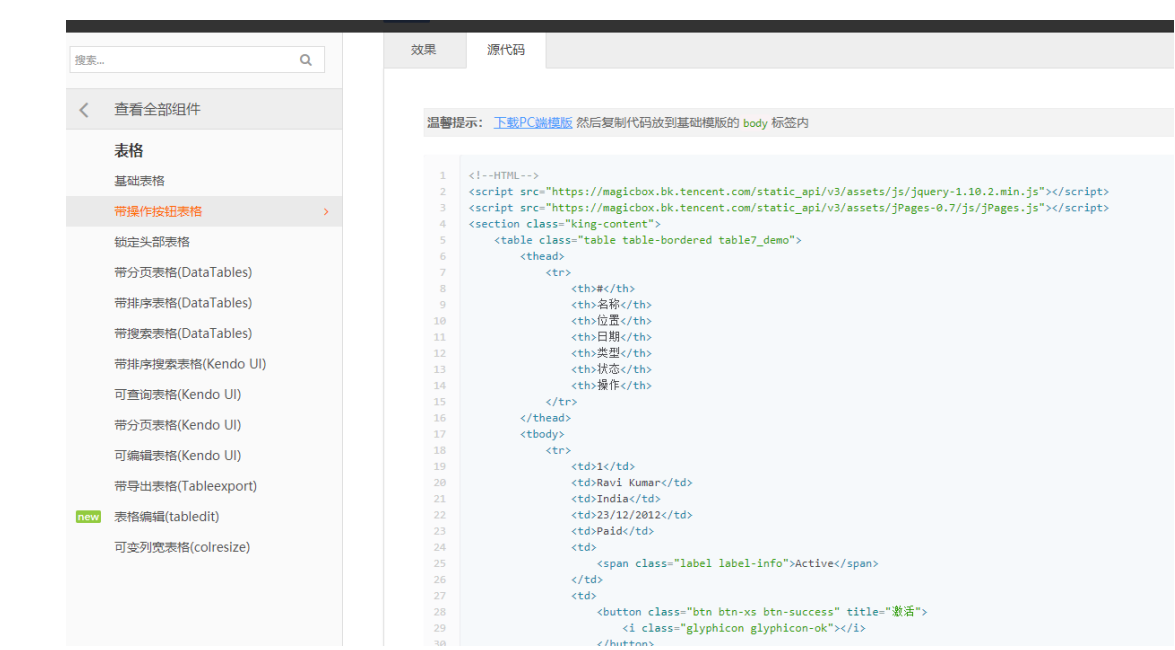


8.3.2.3 组件查看

点击组件名称，则进入相应的组件，可查看组件形状，添加组件变量，查看组件源代码、复制组件源代码等。



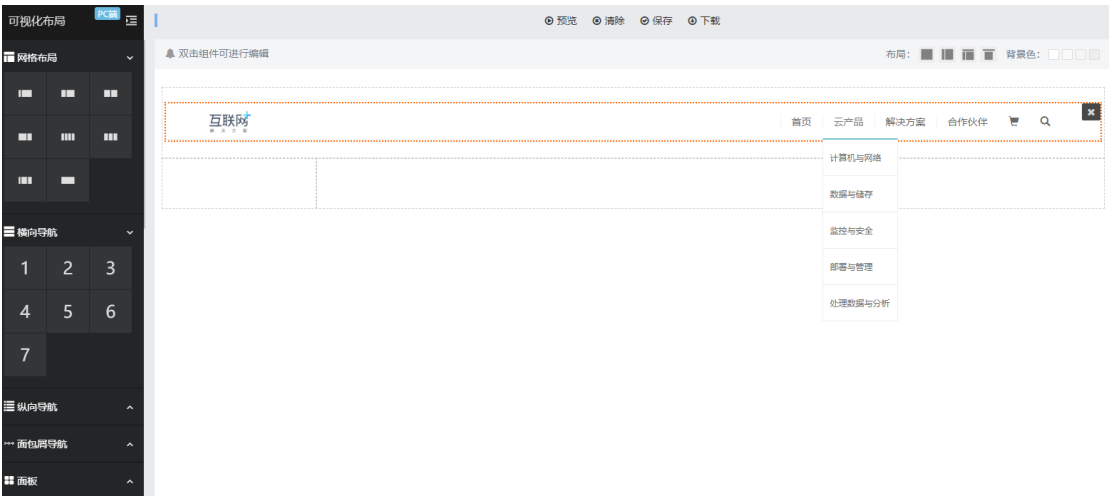
8.3.2.4 组件源代码查看



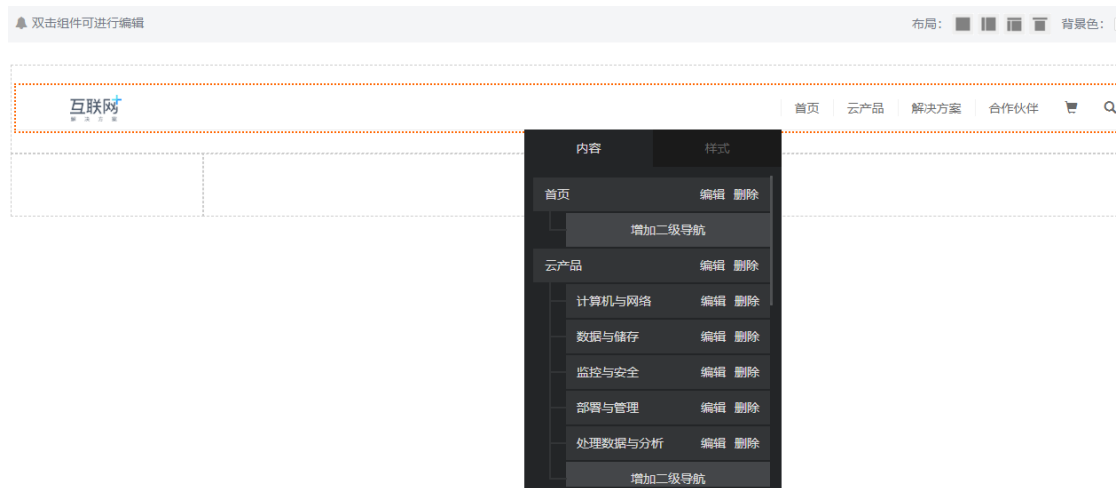
8.3.3 可视化布局

8.3.3.1 PC 端可视化布局

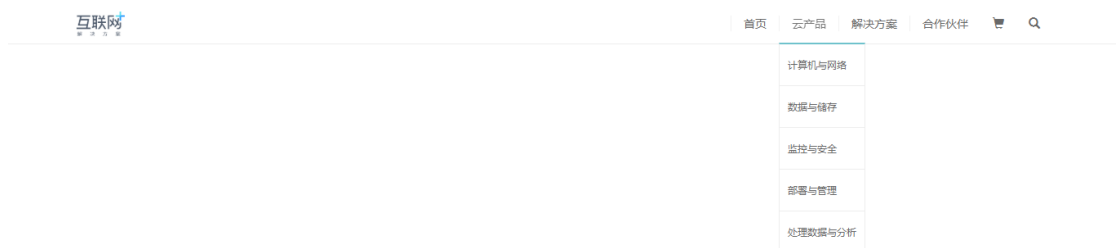
先选择布局样式，然后将相应的组件拖入到布局栏中，双击组件，可对组件进行编辑。



可视化编辑组件



点击预览，就可看到生成的页面。



点击保存，就将该页面保存下来。



点击下载，生成页面代码，复制下来形成 html 保存。

预览清除保存下载

代码

```
1 <!DOCTYPE html>
2 <html>
3
4 <head>
5   <meta charset="utf-8" />
6   <meta http-equiv="X-UA-Compatible" content="IE=edge">
7   <meta name="viewport" content="width=device-width, initial-scale=1">
8   <title></title>
9   <!-- 若您需要使用Kendo UI Professional, 请联系版权人获得合法的授权或许可。 -->
10  <!-- Bootstrap css -->
11  <link href="https://magicbox.bk.tencent.com/static_api/v3/assets/bootstrap-3.3.4/css/bootstrap.min.css" rel="stylesheet">
12  <!-- kendo ui css -->
13  <link href="https://magicbox.bk.tencent.com/static_api/v3/assets/kendoui-2015.2.624/styles/kendo.common.min.css"
    rel="stylesheet">
14  <link href="https://magicbox.bk.tencent.com/static_api/v3/assets/kendoui-2015.2.624/styles/kendo.default.min.css"
    rel="stylesheet">
15  <!-- font-awesome -->
16  <link href="https://magicbox.bk.tencent.com/static_api/v3/assets/fontawesome/css/font-awesome.css" rel="stylesheet">
17  <!-- 蓝鲸提供的公用样式库 -->
18  <link href="https://magicbox.bk.tencent.com/static_api/v3/bk/css/bk.css" rel="stylesheet">
19  <link href="https://magicbox.bk.tencent.com/static_api/v3/bk/css/bk_pack.css" rel="stylesheet">
20  <!-- 如果要使用Bootstrap的js插件, 必须先调入jQuery -->
```

您可以复制代码, 保存为html文件, 并设置文件编码为utf8, 即可预览效果

复制

点击清除, 就将生成的页面进行删除。



8.3.3.2 移动端可视化布局

将右边的组件拖到对应位置，即可将控件展示出来。点击对应的控件，右边即显示控件的可编辑信息。



预览、保存、清除、下载等功能通 PC 端可视化布局功能。

8.3.3.3 微信小程序可视化布局

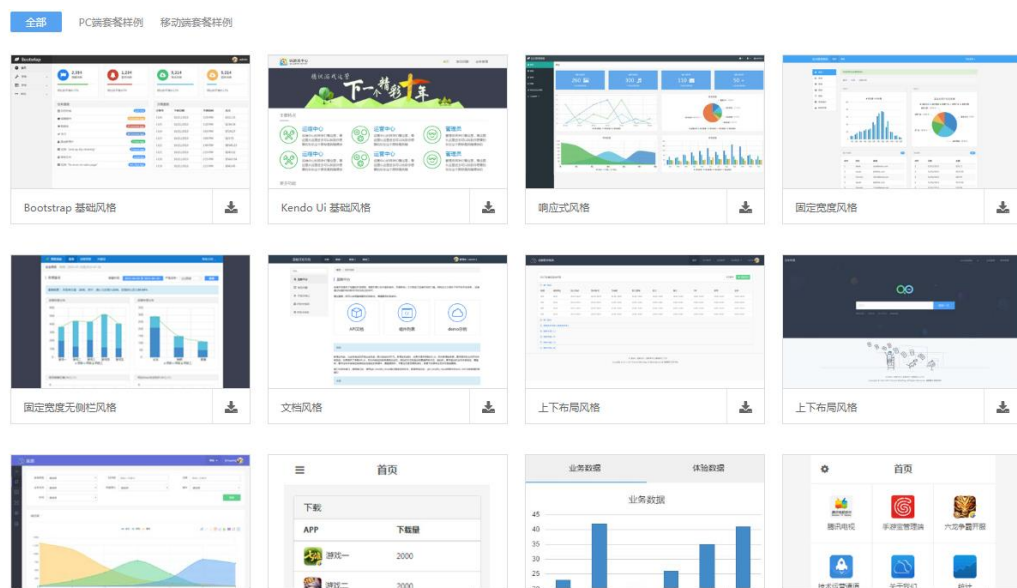
微信小程序可视化布局操作同移动端布局，只是组件是微信小程序的相关组件。



8.3.4 样式风格

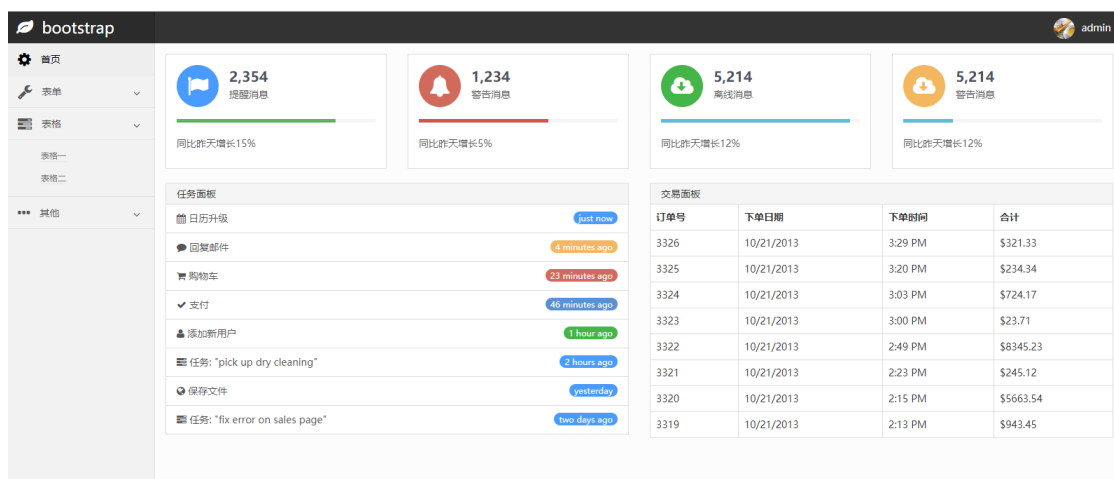
提供多种样式风格选择，样式风格以套餐的形式提供。点击指定样式风格套餐，则进入选定的风格样式。看该种风格样式是否符合我们需要。

风格样式提供 PC 端、移动端选择。



点击指定套餐，进入具体的样式套餐。

Bootstrap 套餐



响应式风格套餐



8.3.5 插件搜索

在插件搜索框中输入插件名称，则搜索出指定插件。

输入想要查找的内容，按Enter搜索

插件名称	版本	简介	操作
artDialog	6.0.4	经典的网页对话框组件，内外皆用心雕琢。	下载资源包 查看示例
icon	2.0	图标库	下载资源包 查看示例
DataSelector	1.0	DataSelector是一款数据选择插件，支持自定义模版，拉取远程数据等功能	下载资源包 查看示例
IDropdownMenu	1.0	基于jQuery实现的下拉菜单插件	下载资源包 查看示例
Switcher	1.0	基于jQuery实现的开关状态插件	下载资源包 查看示例
TimePicker	1.0	TimePicker是一款时间范围选择插件，支持单选，多选两种模式	下载资源包 查看示例
bootstrap	3.3.4	Bootstrap 是最受欢迎的 HTML、CSS 和 JS 框架，用于开发响应式布局、移动设备优先的 WEB 项目	下载资源包
bootstrap-switch	1.2	这是一个针对Bootstrap实现的开关(switch)控件	下载资源包 查看示例
bootstrap-tagsinput	0.6.1	标签输入框，用于管理标签,主要用于将输入的内容自动生成标签。	下载资源包 查看示例

对话框搜索

对话框

插件名称	版本	简介	操作
artDialog	6.0.4	经典的网页对话框组件，内外皆用心雕琢。	下载资源包 查看示例

查看示例

点击查看示例，进入对话框示例页面。

搜索...

< 查看全部组件

弹层

弹窗(artDialog6)

提示层(Bootstrap)

弹出框(Bootstrap)

new 确认框(Bootstrap)

new 右键菜单(contextjs)

提示层(OpenTip)

顶部消息层(Topbar)

消息提示(Toastr)

树形表

树形表(jsTree)

树形表(Kendo UI)

jQuery多级列表树

拖拉组件

弹窗(artDialog6)

普通版 / JS插件 / 弹层

效果

源代码

复制代码

常用效果示例

普通对话框

模态对话框

带按钮弹窗

定时提示框

无关闭按钮

气泡浮层

不同状态示例，可以和消息状态结合使用

成功提示

错误提示

其他提示

消息状态

等待提示

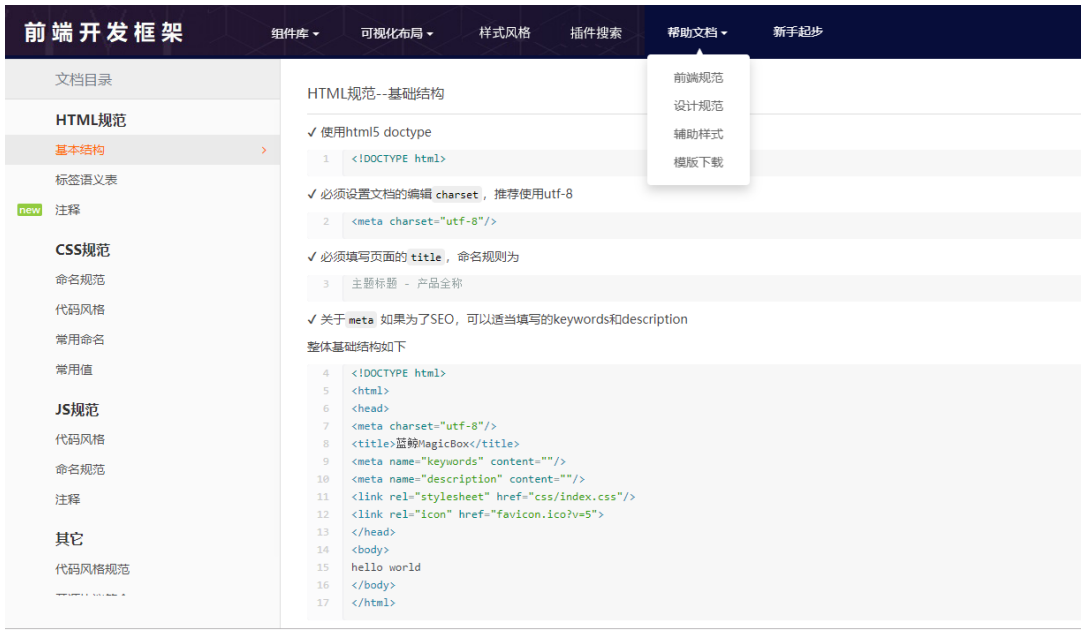
成功提示

失败提示

153

8.3.6 帮助文档

帮助文档提供前端设计的相关规范和样式、模板等。帮助开发人员进行前端开发。



8.3.7 新手起步

指定新手如何使用前端开发框架进行前端开发。里面提供各种布局、插件使用指导教程。

