ALEIYE

让 | 大 | 数 | 据 | 更 | 简 | 单

某 P2P 公司

大数据业务系统整合平台项目

北京数介科技有限公司

2015-12-28

目录

第1章 技术方案

1.1 项目背景

某 P2P 公司，目前已有 p2p 业务，也正在开展电商、分发业务。因前期项目发展需求，目前的数据分析采用的是友盟 / 百度的模块。现在遇到几个问题

1）数据可控性。用户数据的来源包括微信公众号、App、Web 登陆。而目前对微信公众号、web 登录等数据的入口都没有自己的平台进行收集。生成的数据报告也主要集中在友盟 / 百度的已有功能。无法进一步根据自身业务进行扩展开发。

2）数据安全性。由于采用友盟 / 百度等分析平台，而这些平台是第三方的开发平台。因此，某 P2P 公司本身的业务是不会和该平台的业务进行统一分析的。需要建立统一的内部平台来实现外部数据、入口数据、内部数据的统一存储、分析。从而在实现更智能化的业务系统下保证系统的数据安全性。

3）数据扩展性。外部数据（指 web 端、微信公众号、app 等的访问数据）、入口数据（指微信 API 数据、web server 的接口数据等）、和内部数据（内部平台、内部系统和其他数据）的量是非常庞大的，而且随着客户数的增长还将不断变大。因此内部建立的系统需要考虑大数据的建设问题。

因此，某 P2P 公司需要建立自身的大数据平台，来实现所有数据的整合、现有业务的支持和未来业务的扩展性支持。

1.2 建设目标

本期为初步实施大数据业务整合系统，主要目标包括：

1）建立大数据数据采集平台

通过分布式架构，和多种采集手段采集系统分析所需的所有数据（外部、入口、内部等全量数据）。

2）建立大数据数据存储分析平台

考虑结构化、非结构化等多种数据情况，具备良好的数据兼容性。系统需具备 TB 级处理能力，PB 级存储能力。并行式高效处理，前台秒级响应，后台高倍速处理。

3）建立大数据业务分析平台

具备灵活的分析手段，全文搜索、全文关联，任意维度的分析统计。实现对 p2p 业务的数据整合和基本业务分析。包括实现友盟 / 百度已有的部分业务分析功能，数据画像功能（特征分析、地域分析、用户统计、质量分析、转化率分析、渠道建议等）。同时，需充分考虑后续业务扩展的接口问题。

1.3 建设原则

总的原则：遵循标准、立足需求、以运营为目的、总体规划、分步实施。

1.3.1 标准性原则

数介科技提供的系统解决方案及 Aleiye 系统完全满足相关国内标准；国内没有标准的则参照相应国际标准。对当前正在制定和即将制定的国内标准，数介科技承诺在标准出台后能够平稳接轨。

1.3.2 可扩展原则

数介科技所提供的系统可以在保证初期业务的前提下，留有充分的扩展空间，保证将来各种新业务的开展。数介科技所建议采用的 Aleiye 标准系统在后期扩展方面拥有极大的优势，某 P2P 公司可以在业务需要的时候方便的添加设备。不仅可以添加系统设备，而且可以添加符合标准的增值业务设备。

1.3.3 可升级原则

随着技术的飞速进步，现有的设备和系统肯定在不同时期需要进行升级和不同程度的更新，数介科技本次的组网建议能够实现可预见的平滑升级，确保在系统不作大的变更前提下，平滑升级到更高的层次。在升级过程中，能够确保业务不间断，同时保证原设备能正常使用，在未来尽可能的保护原有投资，减少二次投资。

1.3.4 全开放性原则

数介科技本次组网方案采用开放式设计，保证系统与其它各大厂商设备、系统的良好集成性能，能够确保对符合相关标准的第三方厂家设备进行兼容，以便于第三方设备能公平地进入已经部分搭建完毕的系统。

1.3.5 安全性原则

数介科技本次提供的设备系统均按照国家自主可控平台标准建设，系统能够有效地杜绝、限制黑客非法进入系统，以确保系统安全；并且可以根据需要加入系统级备份，可以根据需要选择对系统进行冷备份和热备份。

1.3.6 稳定性原则

数介科技研发的系统能保证单个设备的长期稳定运行，从而保证整个平台的稳定与安全。在某个模块出现问题的时候，也可以很方便的进行更换和维修，因而最大限度的缩小了波及的范围。

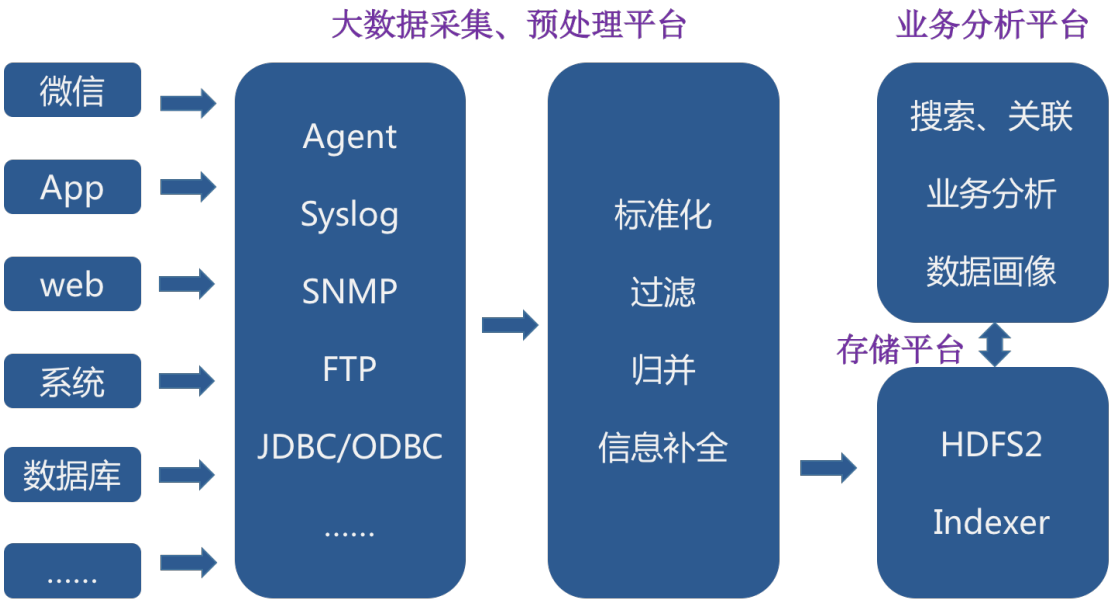
1.3.7 可管理性原则

数介科技本次提供的系统具有完善、健全的网络管理接口，可通过网管的统一控制对设备进行实时全面的监测和控制；可对业务、用户进行方便快捷的查询和管理。并且根据软件系统的重要程度，数介科技提供了不同级别的访问权限设置。同时，Aleiye 平台是部署在企业内部的系统，所采集、存储和分析的数据可以完全对外隔离。

1.3.8 实用性原则

数介科技的系统设计、设备选型均符合中国国情，充分考虑到了某 P2P 公司对自身需求和后续扩张的分析，性能价格比极佳，并通过合理科学的设备选型和网络搭建最大可能的降低买方的长期成本。

1.4 Aleiye 方案设计



1.4.1 数据采集

数据采集是针对目前某 P2P 公司的数据采集不全（很多数据如系统日志、业务日志分散采集或未采集）、数据存储在第三方平台（友盟、百度）等问题。Aleiye 的数据采集会更加某 P2P 公司的实际业务需求和未来开展的业务需求，做统一、全面、可扩展的数据采集、整合、存储。主要技术实现如下：

1.4.1.1 采集方式

数据采集层，主要是针对不同的业务系统和不同的安全设备中的日志进行采集，作为后续的数据处理和关联关系的信息来源，其采集方式重要包括 syslog、snmp、snmptrap、FTP、代理采集和数据库等几种方式。

- Syslog 方式：支持 syslog 内容解码。

- Snmp 方式：支持 snmpV1、V2、V3，内容解码。
- FTP 方式：支持 FTP 协议方式进行日志文件获取。
- 数据库方式:支持当前主流数据库并从中获取日志,其中包括:Oralce、Sybase、DB2、Informix、MySQL、Postgresql 等
- 代理采集：系统须具备在通过安装代理软件实现原始日志的采集功能。

1.4.1.2 采集管理

在分布式采集或单点采集的状况下，Aleiye 数据平台提供采集节点集中管理，实现对采集状态、采集规则和采集起始进行统一管理。

- 批量操作：在分布式采集的状况下，可以对各采集节点进行统一管理，如批量关闭、批量启动、批量暂停、批量下发操作。
- 策略复制：在分布式采集的状况下，可以将单采集节点中的采集策略复制到其他采集节点。
- 采集状态监控：可实时监控不同采集节点的采集状态，在数据传输过程中出现异常，系统会给与采集异常提示。

1.4.2 数据预处理

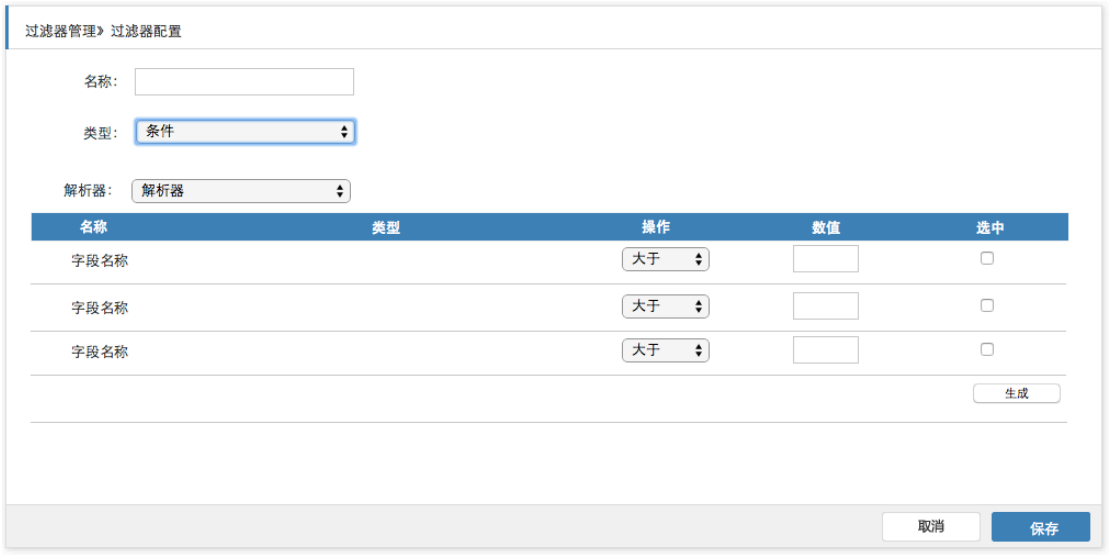
原始日志采集之后，需要进行数据预处理的过程，通过标准化配置，对数据源进行明确的数据类型划分，将日志格式进行统一转化和分类，根据划分好的数据类型进行过滤、归并、补全等规则操作，为后续的关联分析提供信息。最终输出明确的事件类型和各字段属性及补全后的安全对象信息等内容的标准事件

1.4.2.1 数据标准化

根据数据源的内容和格式，对应相应的事件类型进行字段提取、命名等操作，最终形成结构化数据。

1.4.2.2 事件过滤

事件过滤功能通过自定义设置，可对不影响后续分析的安全事件进行过滤，减少不可信、不重要的事件，过滤策略可根据字段间的条件进行有效过滤，字段件条件包括：大于、小于、等于、大于等于、小于等于、等于、不等于；还可以通过关键字和 IP 段进行过滤规则的配置。



1.4.2.3 事件归并

对于重复发生、大部分属性相同的疑似安全事件，在不影响后续事件分析的前提下，应对个体进行合并，减少事件个体数量，并可以对合并后的数据进行新事件的创建。

1.4.2.4 信息补全

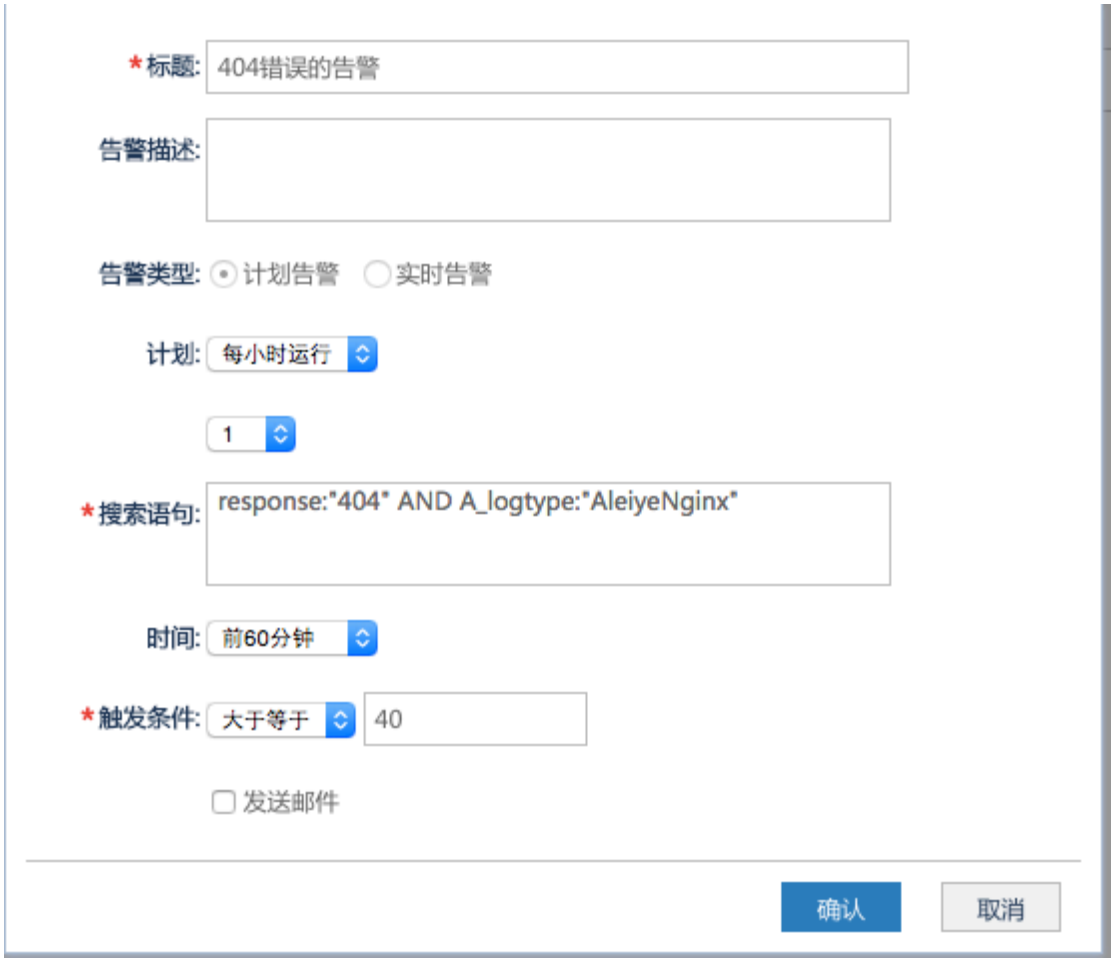
对于未直接体现在原始日志中的必要信息，事件管理模块应具备补全功能，主要为与事件相关的安全对象信息。

1.4.3 业务分析

业务分析是对某 P2P 公司的外部数据、接口数据、内部系统数据的分析方法，以事件触发为基础，对实际业务情况进行深度挖掘，从而实现高可信度的信息。关联分析基于统计关联、模式关联两种方式进行组合。具体的业务场景可参考 1.5 节《功能列表》。

1.4.3.1 统计关联

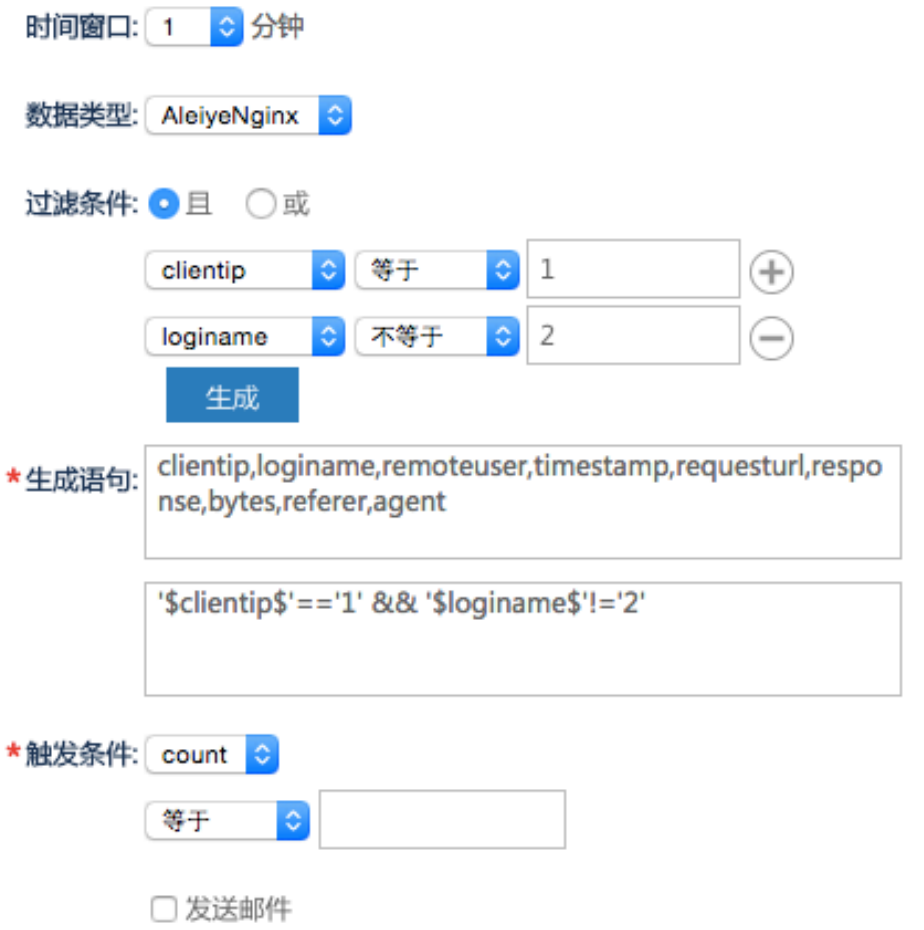
使用计数器来统计某类事件发生的次数，并设定可以接受的数值范围，一旦在统计过程中发现事件超出了正常设定的阈值，就认为系统出现了异常，而生成告警。基于统计关联的方法适应于检查统计量发生次数有明确限制情况



1.4.3.2 模式关联

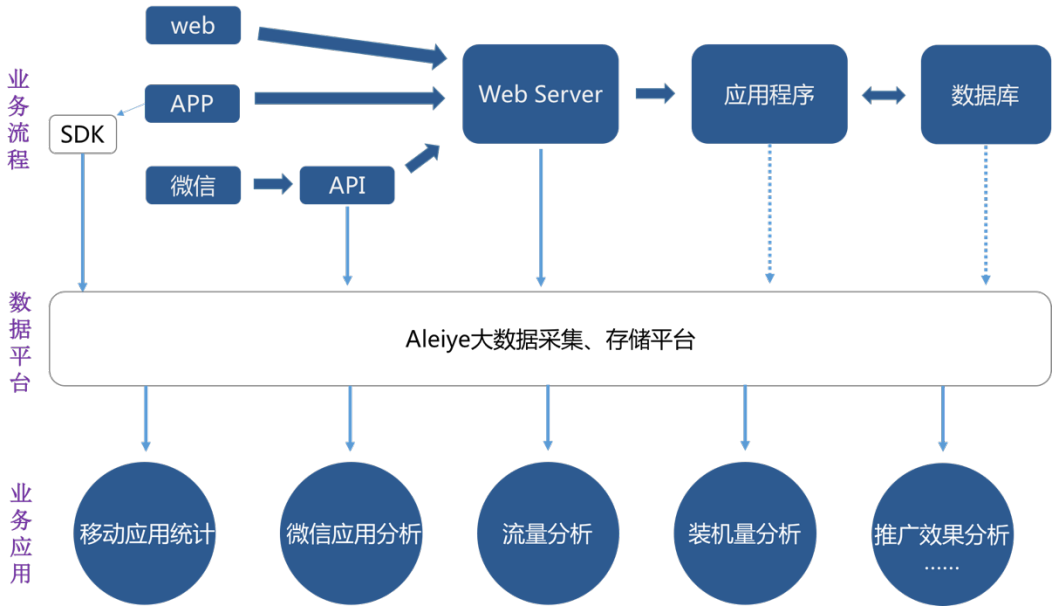
基于模式的关联分析是指将业务活动场景加以预先定义，对收集到的业务事件进行检查，确定该事件是否和特定的模式匹配。

其中，条件为业务事件中某些属性的限制条件，具有检测事实存在与否、比较事实、根据标志检验事实等功能。条件可以由单个检测属性组成，也可以由多个检测属性组成，且各属性用逻辑符号 OR、AND、NOT 来表示多属性的逻辑关系。结果是新告警的输出，同时指定此告警的严重程度。



1.5 功能列表

某 P2P 公司的大数据平台业务流程整合示意图如下：



1.5.1 大数据采集、存储平台

模块	功能	功能点	功能描述
添加数据	上传方式	文件上传	文件上传格式支持：zip、gz、rar、txt、tar.gz、tar、bzip2、gzip、pack200、xz、z、lzma、7z、sz
		采集器	采集器支持格式：文本文件
		FTP 上传	FTP 上传格式支持：gz、zip、文本文件
		Syslog 上传	Syslog 文件
	数据预览	自定义数据	对于无标准格式的数据进行时间戳和换行格式的交互性配置。并可以在数据配置的过程中，能够实时看到配置后的结果数据。
		表格数据	对于表格数据的标准格式数据，无需填写正则表达式，通过交互实现字段的切分和字段名称的匹配。并可以在数据配置的过程中，能够实时看到配置后的结果数据。
检索	SQL 搜索	告警	支持通过 SQL 语句进行查询功能，对搜索结果可以进行可视化编辑。并保存到告警及报表，同样可以通过仪表盘，查看实时结果。
		报表	
		仪表盘	
	搜索功能	告警	使用 Aleiye 搜索命令，可通过使用关键字、短语、字段、布尔表达式和比较表达式来准确指定您想要从 Aleiye 索引检索到的事件。并保存到告警及报表，同样可以通过仪表盘，查看实时结果。
		报表	
		仪表盘	

数据管理	数据源管理	采集器管理	对通过采集器采集的数据源进行管理，可以对采集器下的路径进行开始/停止和删除的操作
		文件管理	对文件上传的记录进行删除操作。
		FTP 管理	对 FTP 上传的数据进行编辑修改和删除操作。
用户管理	修改密码		可以对已有密码进行修改。

1.5.2 业务分析-流量分析

功能模块	子模块	功能描述	统计维度
实时分析		网站分析最基本的应用就是实时监控网站的运营状态。收集网站日常产生的各类数据，包括浏览和访客数据等进行统计和分析，并以可视化形式进行直观的展示。	从以下四个维度进行统计： 1、今日浏览量 PV（从当日零点到当前时间的 pv 数） 2、今日独立 IP 数 3、今日传输总量 4、今日唯一访客
实时搜索		对网站访问日志提供实检索	
趋势分析	实时访问		pv 数 、IP 数

	当前统计	趋势分析模块提供网站访问量数据分析，通过访问量的趋势变化形态，可帮助用户分析出网站访客的访问规律、网站发展状况等。	pv 数 、IP 数
用户画像	地域分布	地域分布功能将网站访客按区域进行统计，其中访客按访客 UI 数、IP 数、跳出率、平均访问时长和平均访问页数等指标进行统计。帮助用户了解网站访客的地域分布，特定地域用户偏好可进行针对性的运营和推广	UV、IP 数、跳出率、平均访问时长、平均访问页数
	ISP 统计	以运营商的纬度分析网站浏览量（PV）、访问次数和访客数（UV）	pv、UV、访问次数
	访客活跃度	分析网站指定时间段内各个时间周期访客的 PV、UV 以及客户忠诚度，进而把握网站的整体访客活跃度趋势；	pv、UV、忠诚度
	访问热点	网站访问热点页面的浏览量（PV）进行统计。其中热点访问页面是指该页面受访次数达到一定频次。热点访问功能帮助用户了解访客的访问行为，从而有针对性的进行网站运营维护工作	浏览量
	访问时长	将访客根据访问网站的时长进行分类统计并计算其所占比重，例如今天在页面停留时间为 1－3min 的访问次数为 54。其中访问次数指有效	访问次数、所占比例

		的会话次数，即同一用户访问间隔不小于 15min 计算为有效次数。	
	来源	将分析该网站不同来源的访客所占比重以及访客 UV 在不同时刻的变化趋势，访问来源包括搜索引擎、外部链接和直接访问。帮助用户了解哪些来源为网站带来了更多的有效访客，为用户合理优化搜索推广渠道提供数据支持；	pv、浏览量占比、访问次数、UV
	网络类型	提供不同网络（移动、联通、电信）、不同类型（2G、3G、4G）引入浏览量的比例情况。可帮助用户了解不同网络类型所带来的访客情况，有针对性的进行维护和优化。	pv 占比、UV 占比
	平均保留时长		访问次数、所占比例
站点质量	带宽流量		区域流量、ISP 流量
	状态码	状态码是指网站异常所对应的网面中显示的数字代码，WLA 目前支持经常出现的五种状态码，分别为 302、200、404、304 和 499。	
	流量分析		地域、浏览量
流量质量	入口分析	入口页即为访客访问网站的第一个入口，即每次访问的第一个受访页面。入口分析分别统计各个页面的	贡献浏览量、IP 数、平均访问时长

		贡献浏览量、访问 IP 数和各访问 IP 的平均访问时长。	
	退出分析	退出分析模块是对各个页面的贡献浏览量、退出浏览量、IP 数、平均访问时长进行统计分析，其中退出浏览量是指访问会话最后一个浏览页面的浏览量。	贡献浏览量、退出浏览量、IP 数、平均访问时长
自定义报表	实时报表	用户可自定义某一时间点，系统根据该时间点定时执行报表规则，并生成定时报表。	
	计划报表	用户可自定义一个时间周期，系统根据该时间周期执行报表规则，并生成计划报表。	
告警信息		通过设定告警阈值，对需要实时告警的事件进行邮件告警	

1.5.3 业务分析-移动应用统计

功能	维度	指标	描述
应用数据	时间（当天、7 天、30 天、90 天）	新增用户、活跃用户、新增用户占比、启动次数、平均单次使用时长、累计用户、累计启动、使用时长、活跃用户、累计人均启动次数、平均使用时长	通过应用数据，可实时产看全面的应用指标数据，也可以根据时间段，查看各个指标趋势数据。

版本数据		时间（今天、昨天、自定义时间）、应用版本	截止今日版本累计用户、新增用户、新增加升级用户、启动次数	通过不同时间段内，对应用中所有版本进行多指标数据查询。
地域分布		时间（自定义时间）、应用版本	新增用户、新增用户占比、启动次数	通过不同时间段内，查看对应用中所有版本在不同地域下的指标数据。
终端数据		时间（自定义时间）、应用版本、机型、分辨率、操作系统、联网方式	新增用户、占比	在不同时间段内，查看应用在不同版本下不同设备、系统、联网方式等指标的数据。
错误报告		时间（自定义时间）、错误摘要、应用版本、错我详情、设备、操作系统	错误次数	在不同时间段内，查看应用中不同版本下发生错误的信息，同时还可以查看该错误信息的详情和错误信息发生的次数、所在的设备和系统。
路径分析		起始页面、全部版本、时间（自定义时间）	访问次数、访问占比、平均访问时长、平均访问时长占比	在不同时间段内，基于起始页面，可以追踪其访问路径，最多支持层页面路径，起始页面可随意设置。
添加数据	安卓系统			基于 SDK 对安卓平台中的应用进行数据采集

	IOS 系统			基于 SDK 对 IOS 平台中的应用进行数据采集
应用管理				可以对已添加的应用进行增删改查操作

1.5.4 业务分析-微信公众号分析

端口	数据源	可实现功能
微信端	1.1 微信分组数据： 包括分组 ID、名称、总人数	分组特征分析
	1.2 用户基本信息： 是否订阅公众号、账号、昵称、头像、微信名、地区（国家、省份、城市）、性别，设置备注、标签、关注时间、用户所在分组	用户特性统计
	1.3 用户地理位置： 用户同意上报地理位置后，每次进入公众号会话时，都会在进入时上报地理位置，包括地理位置纬度、地理位置经度、地理位置精度	用户地域分析
	1.4 用户列表： 关注该公众号的总用户数、新增关注用户数、取消关注用户数	用户特性统计
	1.5 微信自身的消息分析数据接口	用户质量分析
注册模块	2.1 在微信链接的注册过程中（自身平台）对接微信帐号（1.2 的 openID），进行用户列表交叉分析，统计微信公众号到注册用户的转化率	微信转化率统计用于提升转化率

	2.2 在注册过程中对前一级链接进行分析，增加字段（如来自微信公众号、来自直接注册、来自好友推荐等其他渠道）。实现对用户注册来源的统计	用户注册来源分析
	2.3 非注册用户的关注情况，通过 mib 号，UUID 等方式，实现对访客到注册用户的转化	非注册用户转化成注册用户
应用内部	3.0 需了解应用是内嵌 Aleiye 的 SDK 或应用生成访问日志，然后通过 Aleiye 平台采集日志。内嵌 Aleiye 的 SDK，转到 3.1-3.3，访问日志方式，转到 3.4	
	3.1 用户访问页面记录	
	3.2 页面访问时长记录	
	3.3 Refer 的跳转情况记录	
	3.4 Server 端情况了解（需知道是 Apache、IIS、Nginx）	
	3.5 App 的业务信息	后续功能

1.6 方案价值

Aleiye 数据平台可以整合某 P2P 公司的外部数据、接口数据、内部数据等多个业务系统数据，统一设备输出的事件，帮助管理人员从全局角度保证整体业务态势。该平台目前可以实现对所有数据的采集、预处理、整合、存储，并实现了部分业务分析工作，如流量分析、移动应用统计、微信公众号分析等，并具备良好的业务接口，可快速实现后续业务的分析。该平台采用最新的大数据架构方式，在数据量和数据快速处理都具备很高的先进性。

第2章 数介科技简介

北京数介科技有限公司是一家专注于企业大数据分析、挖掘和应用的高新科技企业，国内领先的大数据解决方案解决商。核心理念是为企业在大数据变革中提供技术支撑平台，真正实现企业数据的可见、可用，可挖掘价值。

数介科技依托自主知识产权的 Aleiye 实时大数据分析引擎，形成数据平台+应用服务+行业解决方案的综合大数据产品。可充分应对行业多样化和企业个性化的大数据需求，为企业在 IT 运维、业务运营、系统安全以及合规审计等多方面提供深度服务。

目前，数介科技的大数据服务已深入金融、运营商、广告、政府等多个行业，并协助客户收集并整合海量业务数据，提供多维度的数据分析图表，预测业务发展趋势，为经营决策提供直观、精确、实时的数据支撑。

第3章Aleiye 技术白皮书

3.1 Aleiye Data Engine 简介

ALEIYE 是企业交付式大数据开放平台，为企业提供大数据服务，并协助客户收集并整合海量业务数据，提供多维度的数据分析图表，预测业务发展趋势，为经营决策提供直观、精确、实时的数据支撑。

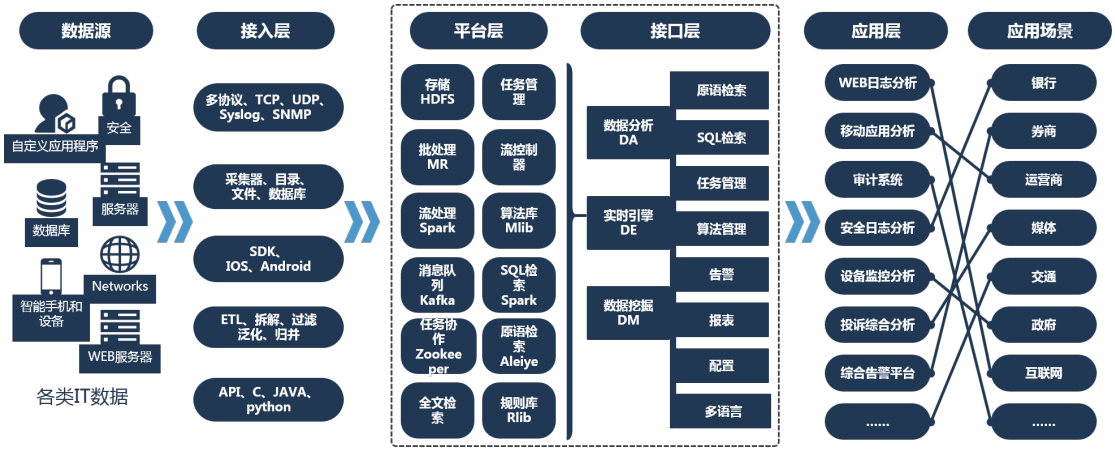


图 1：Aleiye 大数据引擎流程框图

3.1.1 Aleiye Data Engine 部分服务

- **银行：**负载分析、数据整合、安全分析、资产管理、系统告警、审计等
- **证券：**证券交易日志统计、企业安全日志分析、交易报表、审计等
- **保险：**审计、数据整合、安全分析、数据挖掘、潜在客户分析等
- **电信运营商：**ISMP、SIM 平台替换、资产管理、数据整合、安全分析等
- **IDC 行业：**数据整合、资产管理、设备监控、负载分析、关联分析、自动告警等
- **媒体：**系统访问日志分析、数据整合、关联分析、数据挖掘等
- **移动互联网：**移动 APP 用户行为分析、仪表盘、关联分析等
- **政府：**数据整合、安全分析、资产管理、数据挖掘等

3.1.2 Aleiye Data Engine 技术优势

- **简单**：通过 ALEIYE 平台提供的交互界面，使用者在无需了解底层技术的前提下即可对自身的企业数据进行泛化、入库、检索、分析以及挖掘等多种操作。
- **灵活**：数据接入方式灵活多样，并可通过 SQL、原语、脚本等多种检索方式满足不同对数据检索的多样需求，并生成报表；内置算法库，方便用户挖掘数据潜在价值。
- **高效**：TB 级别的数据处理能力；热点数据秒级响应；实时规则告警；动态调整实时报表纬度；
- **开放**：支持多平台集成，快速安装、简易配置。API 支持 C/C++、Java、Python 等主流开发语言直接调用。
- **交互**：企业内部独立部署，确保数据绝对安全性；通过 API 完成应用的快速开发；应用插件化，实现业务快速迁移。

3.2 Aleiye Data Engine 体系架构

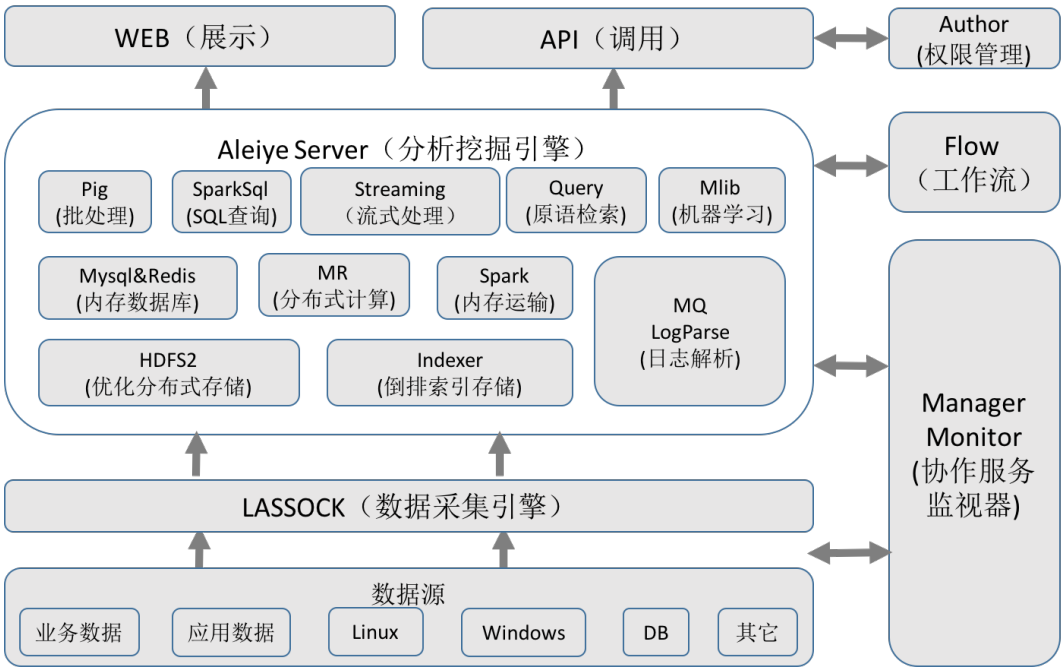


图 2：Aleiye 大数据引擎技术架构

3.2.1 Aleiye LASSOCK

Aleiye LASSOCK 数据采集引擎是无视数据结构的数据整合基础。通过如数据采集器、文件上传、协议传输，脚本采集，API 等手段将分散的、异构的数据进行实时的收集、拆解并整合进入平台。企业通过定义的采集规则，通过对数据进行拆解、过滤等手段进行预处理，并保证数据的实效性，完整性及准确性，为数据的关联、分析以及挖掘打下基础。

3.2.2 Aleiye OpenSource

Aleiye OpenSource 包括 Apache 开源项目和基于开源的 Aleiye 优化项目。Apache 开源项目主要包括：优化的分布式存储 HDFS2、内存运输 Spark、批处理 Pig 和结构化查询 SparkSQL 等。Aleiye 优化项目包括：内存数据库 Redis、关系型数据库 Mysql、分布式计算 MapReduce、倒排索引存储 Indexer 和协作服务监视器 Manager Monitor 等。Aleiye 通过优化大幅度提高了系统的性能和稳定性，从而保证了 Aleiye 大数据引擎的安全可靠。

3.2.3 Aleiye Server

Aleiye Server 是 Aleiye 自主开发的分析挖掘引擎，提供从 Aleiye LASSOCK 采集数据的解析、流式处理、存储、原语检索、关联分析、机器学习等多种数据处理手段，通过工作流的控制保证整个分析挖掘过程的安全稳定。Aleiye Server 内置电信级安全规则库，可适合于相关规则的各种业务场景；同时提供了规则引擎模块，能快速适用于其他业务应用场景。

3.2.4 Aleiye BasicAPI

Aleiye BasicAPI 提供丰富的、完善的 API 接口，使得 Aleiye Data Engine 可以成为真正的平台产品，提供类似于操作系统的功能，第三方可以在其上做相关的业务开发。目前在 Aleiye Data Engine 上调用 BasicAPI 实现的业务应用已有几十种。

3.3 Aleiye Data Engine 数据整合

企业数据一般都分散存储在不同的业务系统中，企业规模越大业务系统越多，数据类型也就越多越复杂。所以多数据类型的整合是构建企业大数据平台的第一

步。Aleiye LASSOCK 采集器可直接在服务器中运行，通过 web 控制台，对运行在多台设备上的的采集器进行控制管理。可以支持同时监控多个文件的变化情况，并将变化后的数据实时采集提交到平台。

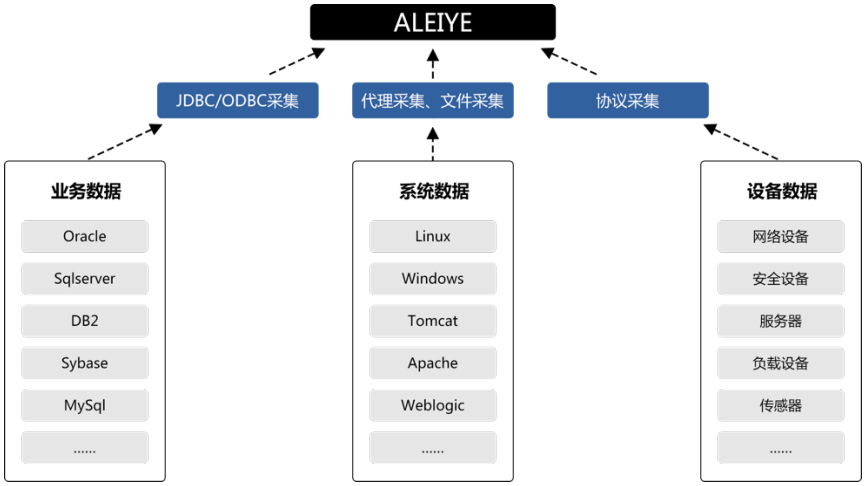


图 3：数据采集过程

- 文件上传

  通过 web 界面直接将文件上传至平台。支持格式包含文本文档、csv、rar、zip、7z、tar 和 tar.gz 压缩文件。

- 协议传输

  数据可以通过协议进行传输。支持以下传输协议

  - FTP：支持系统获取固定 FTP 的文件，也可以通过 FTP 协议进行上传。
  - Syslog：通过 syslog 将数据直接发送到平台。
  - SNMP：主要针对运维信息，可以通过标准 snmp 进行采集。

- 脚本采集

  平台提供数据上传脚本，可以通过指定的用户名参数，将命令的执行结果发送给系统。

- API

  提供入库工具包，可以兼容 java、python、php 等脚本语言。用户可以直接通过编程直接将数据发送给数据平台。

- 其他方式

- 数据库
  - 支持 mysql、oracle、sqlserver 等常见关系型数据库。
  - 支持历史数据直接导入。
  - 支持增量数据的导入
- 其他。

3.4 Aleiye Data Engine 数据处理

预处理

原始日志采集之后，需要进行数据预处理的过程，通过标准化配置，对数据源进行明确的数据类型划分，将日志格式进行统一转化和分类，根据划分好的数据类型进行过滤、归并、补全等规则操作，为后续数据处理提供信息。最终输出明确的事件类型和各字段属性及补全后的信息等内容的标准事件。

数据预处理（data preprocessing）是指在主要的处理以前对数据进行的一些处理。现实世界中数据大体上都是不完整，不一致的脏数据，无法直接进行数据挖掘，或挖掘结果差强人意，为了提高数据挖掘的质量产生了数据预处理技术。

数据预处理的主要步骤：数据清理、数据集成、数据规约和数据变换。具体实现步骤主要包括过滤、归并和补全过程。

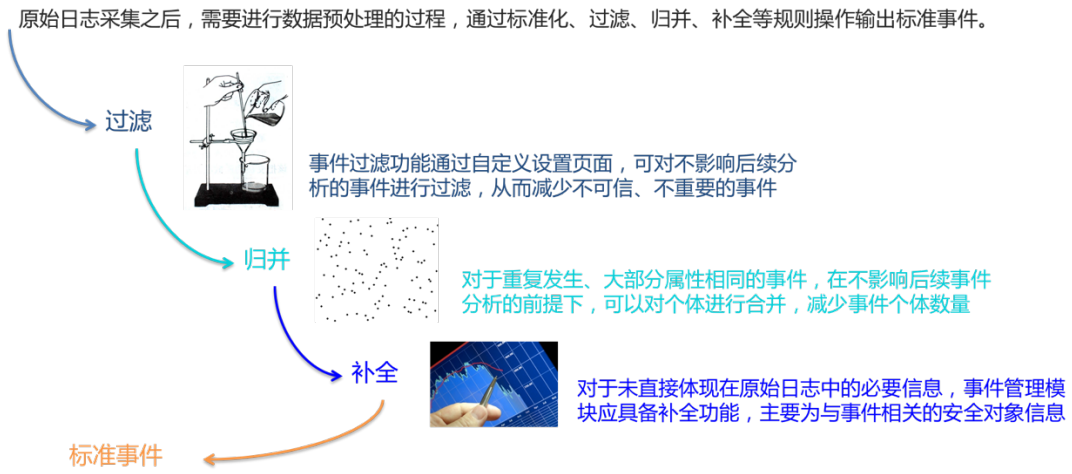


图 4：标准事件生成步骤

过滤：

事件过滤功能通过自定义设置页面，可对不影响后续分析的时间进行过滤，从而减少不可信、不重要的事件，过滤的策略可根据字段间的条件进有效过滤，字段间条件包括：大于、小于、等于、大于等于、小于等于、等于、不等于；还可以通过关键字和 IP 段进行过滤规则的配置。

归并：

对于重复发生、大部分属性相同的事件，在不影响后续事件分析的前提下，可以对个体进行合并，减少事件个体数量，支持供事件归并规则配置功能，定义事件归并的条件和方法。

补全：

对于未直接体现在原始日志中的必要信息，事件管理模块应具备补全功能，主要为与事件相关的安全对象信息。该功能以规则驱动方式实现。

实时分析

- 数据流处理：Aleiye 结合数据属性以及用户需求，对实效性要求的较高的数据进行实时的数据流处理。
- 实时检索：类似百度和谷歌的关键字检索，并可以使用布尔代数 AND、OR、NOT 及括号任意组合关键字进行数据的实时检索。
- 实时告警：对于时序数据，可根据业务规则定制告警规则，对在数据流动的采集过程中指定时间窗口内发生了满足业务规则的数据流触发告警。

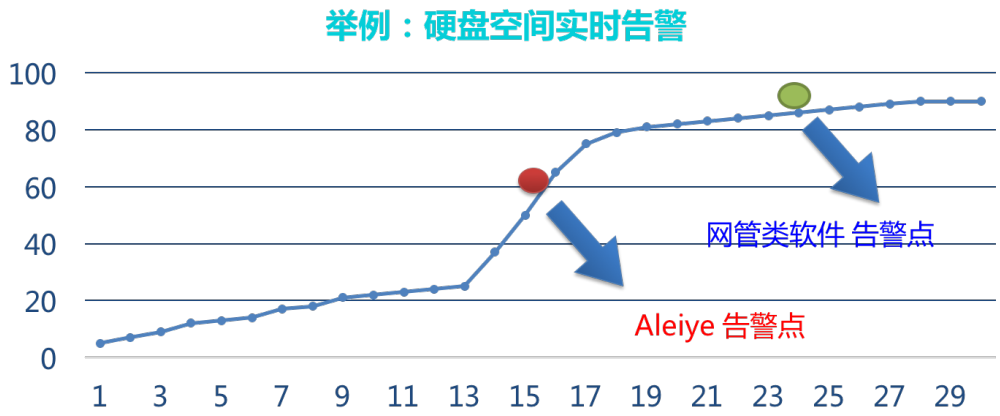
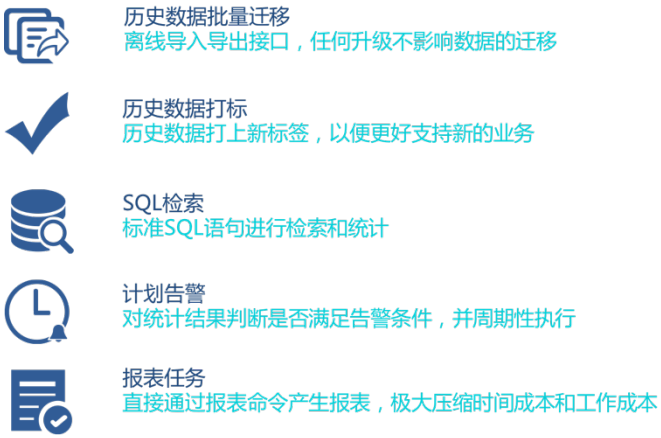


图 5：Aleiye 实时处理优势举例

离线分析



- 历史数据批量迁移：ALEIYE 数据平台提供数据离线导出和导入的接口，因此，任何平台升级、硬件升级或业务系统升级改造都会不影响数据的迁移。
- 历史数据打标：历史数据打标：当新业务产生后，有可能需要对历史数据进行新的业务分类。系统提供了历史数据批量打标的功能，满足历史数据添加新标签以知更好的支撑新业务的需要。

- SQL 检索：通过 SQL 语句对已经存储到数据库的数据进行检索和统计。
- 计划告警：通过周期性的任务定义告警手段。ALEIYE 平台可以通过关键字或 SQL 语句对统计的结果判断是否满足告警条件，并按照指定的周期执行。
- 报表任务：报表可以将业务最直观展现。传统的数据报表需要编写代码、数据入库、前端展现等多步骤实现，而 ALEIYE 可以直接通过报表命令产生报表，并且组成用户自己报表群支撑日常工作，极大的压缩是时间成本和工作成本。

3.5 Aleiye Data Engine 关联分析

趋势分析

- 预测

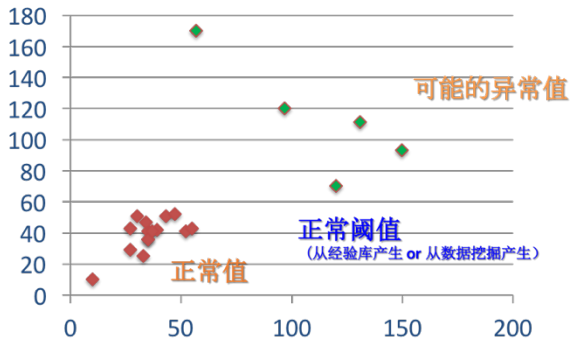
时序数据是具有流动性的，而且一般业务都存在周期性。平台通过对历史数据进行抽象，形成模型。形成的模型结合当前数据的表现，可以预测下一个阶段数据趋势

- 预警

基于预测结合告警阈值的设置，就可以达到预警的目的，提前发现系统或是业务可能出现的爆发点。

关联分析

不同的业务场景，关联分析的内容会有比较大差异。平台提供基于时间和基于业务两种机制。



- 基于时间：

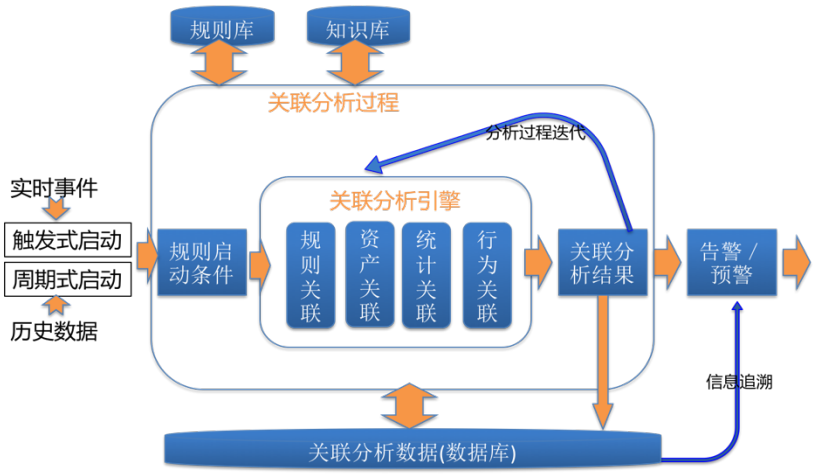
根据时间进行的关联分析。当某个业务出现异常时，可以帮助用户找到问题之间关系，如先后顺序，影响范围等。

- 基于业务

结合业务情况，用户选择数据源和需要分析的指标，选择不同的算法定义任务，提交给平台进行分析。

关联分析效果的关键是在分析过程中使用适当的分析引擎、提取适当的数据、输出适当的告警，这些关键因素均是由关联分析规则进行约定。

以安全关联为例，安全告警关联分析过程是规则驱动的综合信息分析活动，可采用事件触发方式启动也可周期性触发，以 Aleiye 平台掌握的各类安全信息为输入，以输出安全告警为目标。



Aleiye Data Engine 安全关联模型如右图所示。

3.6 Aleiye Data Engine 数据挖掘

Aleiye 通过 LASSOCK 采集后的数据，经过预处理、实时分析、离线分析和关联分析的所有数据都可以成为 Aleiye 数据挖掘模块的数据源。Aleiye 的主要挖掘算法包括：

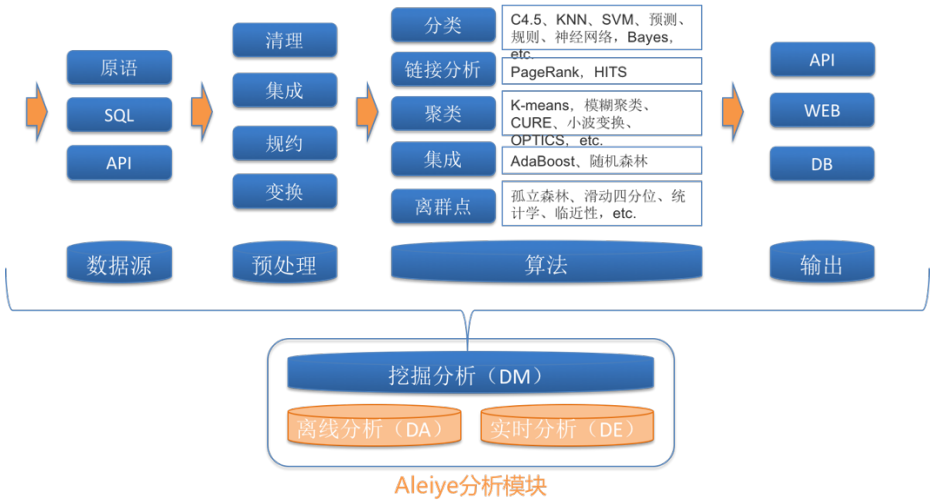


图 6：Aleiye Data Engine 数据挖掘模块

分类算法

分类与预测是两种数据分析形式，它们可以用来抽取能够描述重要数据集合或预测未来数据趋势的模型。分类方法（Classification）用于预测数据对象的离散类别（Categorical Label）；预测方法（Prediction ）用于预测数据对象的连续取值。主要的分类算法包括：决策树、KNN、SVM、Bayes、神经网络、VSM、预测和基于规则的分类（关联分析）。

聚类算法

聚类分析是把一个给定的数据对象集合划分成不同的子集的过程，每个子集是一个簇。聚类是一种无监督分类法：没有预先指定的类别；遇到要分析的数据缺乏描述性信息时，或者无法组织成任何分类模式时，采用聚类分析。

聚类作为一种典型的数据挖掘方法，一直以来都是人工智能领域的一个研究热点，被广泛地应用于人脸图像识别、股票分析预测、搜索引擎、生物信息学等重要领域中。目前主要的聚类算法包括 K-Means、小波变换、CURE、模糊聚类等。

链式分析

典型的链式分析算法包括 Google 的 PageRank 算法和应用于小规模数据的 HITS 算法。

集成算法

主要包括 AdaBoost 迭代算法和随机森林算法。

离群点检测

根据实际项目业务需求，Aleiye Data Engine 已经实现了基于统计学的离群检测、基于邻近性的离群检测、孤立森林离群检测和滑动四分位差距离群检测算法等。

ALEIYE

让 | 大 | 数 | 据 | 更 | 简 | 单

公司地址：北京市西城区新街口外大街28号普天德胜大厦A座405

邮　　编：100088

联系电话：010-82053991

电子邮箱：service@ALEIYE.com