专注构建企业大数据平台

ALEIYE Data Engine 企业级大数据引擎

技术白皮书

北京数介科技有限公司

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。

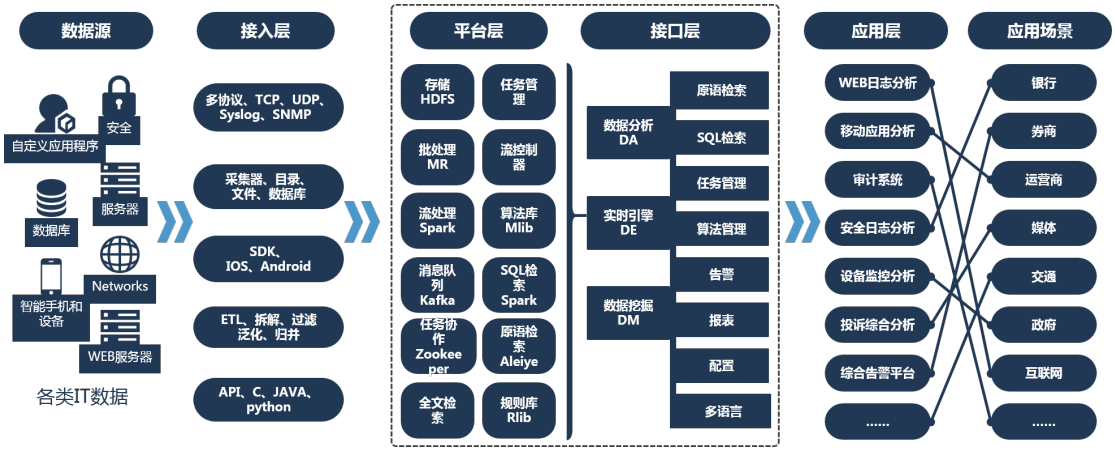


图 1：Aleiye 大数据引擎流程框图

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

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。

Aleiye Data Engine 体系架构

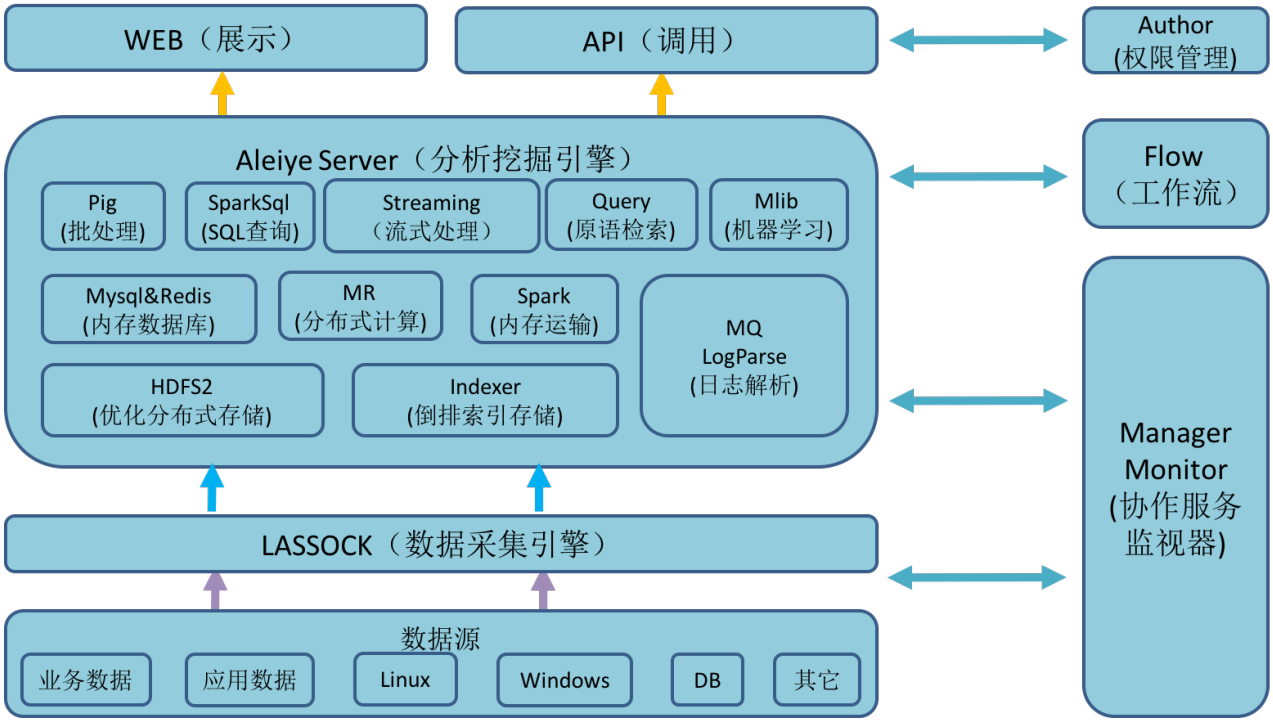


图 2：Aleiye 大数据引擎技术架构

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。

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。

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。

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。

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。

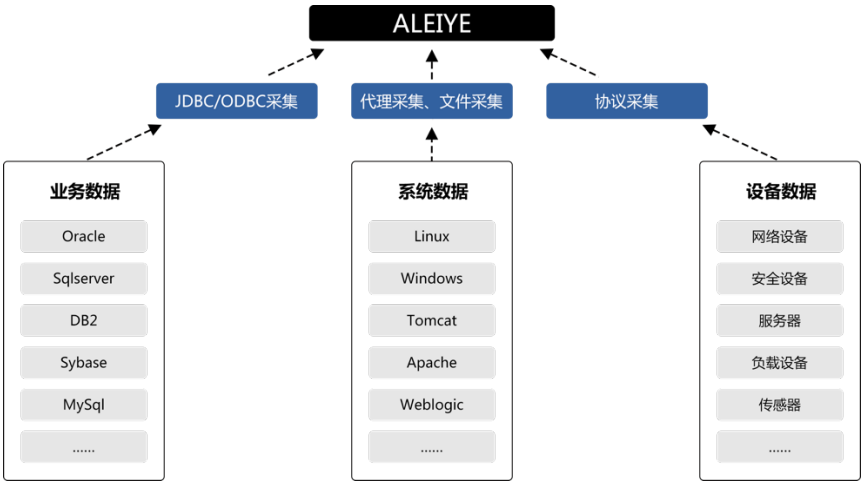


图 3：数据采集过程

- 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。

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。

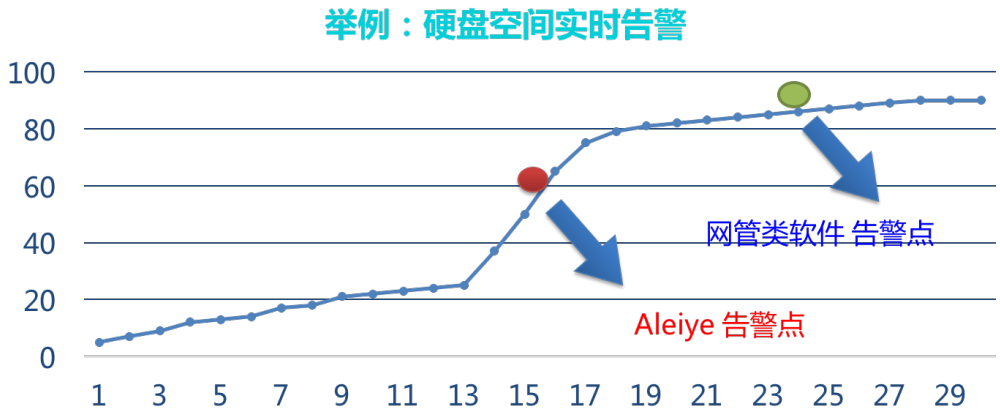
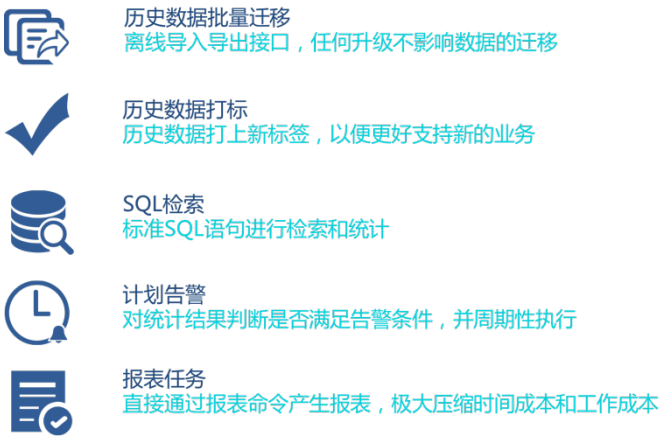


图 5：Aleiye 实时处理优势举例

离线分析



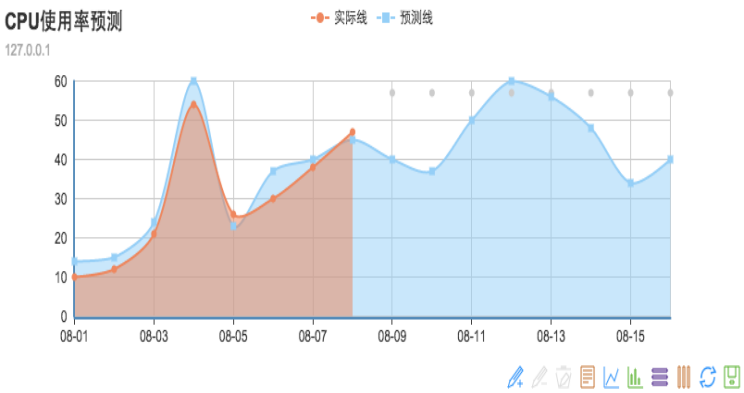
- 历史数据批量迁移：ALEIYE 数据平台提供数据离线导出和导入的接口，因此，任何平台升级、硬件升级或业务系统升级改造都会不影响数据的迁移。
- 历史数据打标：历史数据打标：当新业务产生后，有可能需要对历史数据进行新的业务分类。系统提供了历史数据批量打标的功能，满足历史数据添加新标签以知更好的支撑新业务的需要。
- SQL 检索：通过 SQL 语句对已经存储到数据库的数据进行检索和统计。
- 计划告警：通过周期性的任务定义告警手段。ALEIYE 平台可以通过关键字或 SQL 语句对统计的结果判断是否满足告警条件，并按照指定的周期执行。
- 报表任务：报表可以将业务最直观展现。传统的数据报表需要编写代码、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。

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

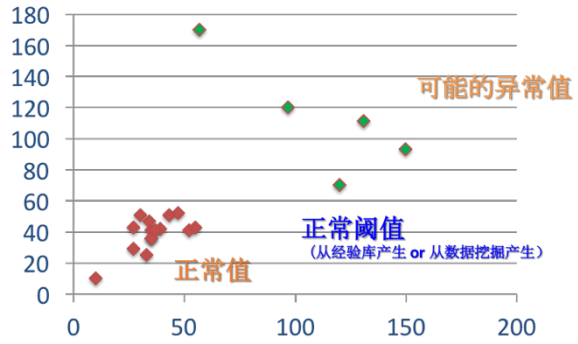


- 预警

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。

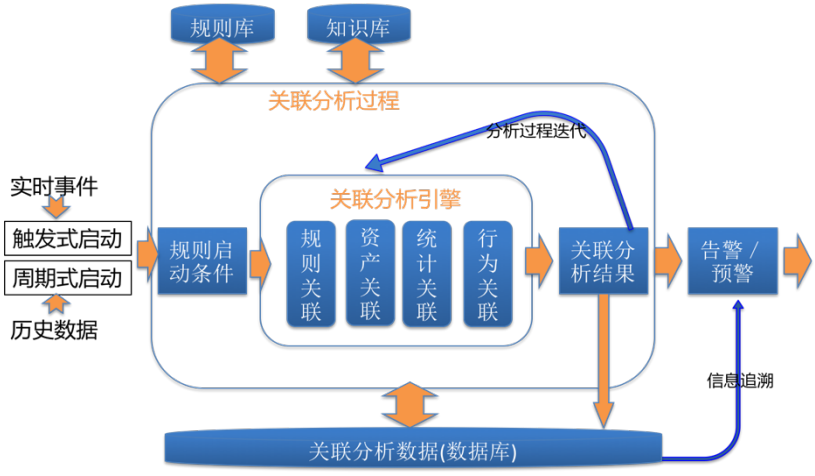
- 基于业务

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。



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：

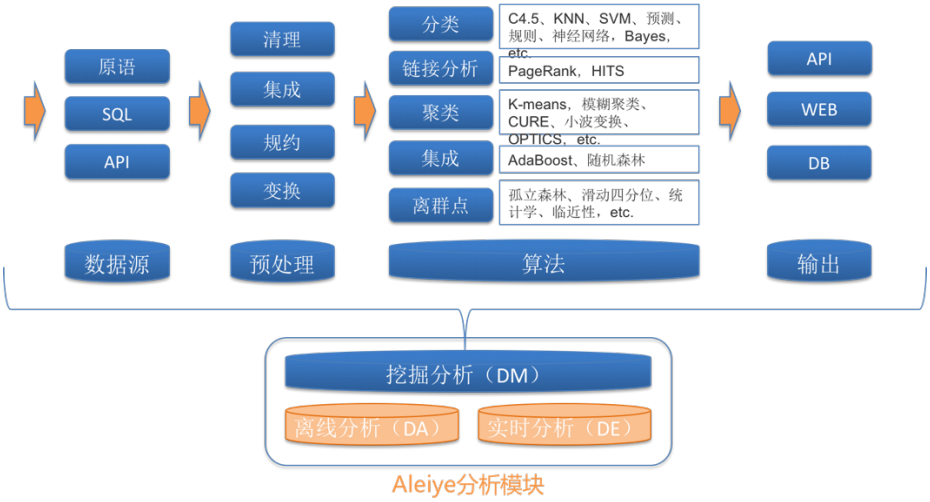


图 6：Aleiye Data Engine 数据挖掘模块

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。

关于我们

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。

应用案例

目前，数介科技的大数据服务已深入银行、证券、运营商、广告、政府、移动互联网、军队等十多个行业，30 多个使用客户，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。

北京数介科技有限公司
网址：www.aleiye.cn
邮箱：service@aleiye.cn
电话：010-82053991
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