**数据中心-产品需求分析文档**

V1.0.0

张立丹 @ Enlink 数据中心

**目录**

[1. 概述 3](#_Toc528932301)

[1.1. 说明 3](#_Toc528932302)

[1.2. 模块 3](#_Toc528932303)

[1.3. 涉众（用户群） 4](#_Toc528932304)

[1.4. 目标 4](#_Toc528932305)

[2. 接入模块 5](#_Toc528932306)

[2.1. 数据类型 5](#_Toc528932307)

[2.2. 接入方式 5](#_Toc528932308)

[2.3. 性能需求 5](#_Toc528932309)

[3. 存储模块 6](#_Toc528932310)

[3.1. 数据特点 6](#_Toc528932311)

[3.2. 实时存储 6](#_Toc528932312)

[3.3. 离线存储 6](#_Toc528932313)

[4. 分析模块 8](#_Toc528932314)

[4.1. 分析方式 8](#_Toc528932315)

[4.2. 实时分析 8](#_Toc528932316)

[4.3. 离线分析 8](#_Toc528932317)

[5. 预警模块 9](#_Toc528932318)

[5.1. 预测功能 9](#_Toc528932319)

[5.2. 报警功能 9](#_Toc528932320)

[6. 可视化模块 10](#_Toc528932321)

[6.1. 数据无关 10](#_Toc528932322)

[6.2. 数据分析 10](#_Toc528932323)

[6.3. 高可定制 10](#_Toc528932324)

[6.4. 大屏展示 10](#_Toc528932325)

[6.5. 自动报表 10](#_Toc528932326)

# 概述

本文档描述“DataCenter（数据中心）”产品的功能需求，帮助或指导系统设计人员对系统的整体架构进行设计。

## 说明

在编写本需求文档时，DataCenter仅存在对日志形式的数据进行分析、展示的需求，但是系统设计人员需要考虑本产品可接入其他类型的数据源。

DataCenter包含一系列的子产品或模块（以下统称为“模块”），本文档仅对产品中的各个模块进行概括性的描述，每个模块的具体需求不在本文档描述范围之内，请参阅相应模块的需求分析文档。

本需求文档的编写，参考了3个方面内容：

* 编者的实战经验
* 开源社区解决方案
* 大数据相关知识

撰写本需求文档之前，编者缺少对当前市场及竞争对手产品的全面了解。产品如果要做的适应市场、引领市场，则需要投入到市场进行打磨。

## 模块

本小节描述DataCenter所包含的基本模块，每个模块经过“包装”之后应可视为一个独立的产品。如图1-1所示：

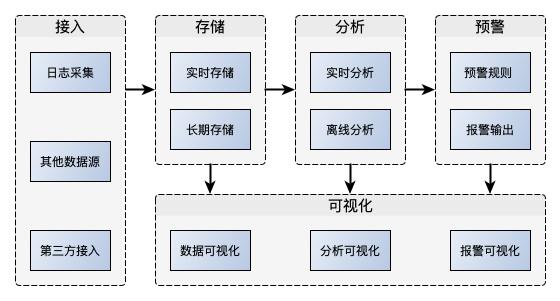


图 1‑1 DataCenter产品模块

本小节描述的模块，是站在产品的角度考虑划分的，与系统设计中的模块没有必然的关系。本文档的后续章节，将对图1-1中所列的所有模块进行详细的描述。

## 涉众（用户群）

本小节描述DataCenter潜在的用户群体，站在用户的角度来分析、描述产品的需求。本产品可能的用户群体包括：

* **系统运维人员**  
  分析系统健康状况、业务延迟、稳定性等；
* **业务分析人员**  
  分析业务趋势、用户群分布、增长预测、审计等；
* **安全分析人员**  
  网络可用性、数据泄漏、用户行为分析等；
* **总监、BOSS**  
  业务报表、大屏展示等。

每一类用户群体，对本产品的需求都不尽相同，如果要使本产品适用于所有潜在的用户，则在设计时需要保证本产品有非常高的灵活性。

## 目标

本小节描述DataCenter产品最终形态的实现目标，如下所示：

* **多样化数据接入**  
  格式多样化：例如，日志格式、数据库格式、网络资源等；  
  内容多样化：例如，用户数据、行为数据、销售数据等；
* **大数据存储**  
  大量、高速、价值等；
* **智能分析和预警**  
  实时分析及预警：异常、安全等；  
  离线分析及预测：BI相关；
* **丰富、灵活的可视化**  
  多维度、可配置、数据无关性等。

DataCenter产品设计时，需要考虑最终的交付产品需达到上述的目的。

# 接入模块

本章节描述DataCenter产品中，对数据接入模块的需求。数据接入是指需要分析、展示的数据进入到DataCenter产品中的方式。

## 数据类型

用户侧的数据源类型多种多样，接入模块需支持多样化的数据接入，例如，不同的存储方式、不同的存储格式等。数据类型包括但不限于：

* **文件型数据**  
  用户业务系统产生的日志文件数据，日志格式的多样化、日志种类的多样化；
* **实时型数据**  
  用户设备导出的实时日志，例如用户防火墙日志等，需要考虑标准化及厂商定制；
* **数据库数据**  
  用户存储在数据库中的数据，数据库类型的多样化、数据种类的多样化；
* **用户自定义**  
  用户自定义存储类型。

接入模块设计时，需要考虑到用户行业差异性、业务系统差异性、业务系统研发水平差异性等因素。通过配置或简单的定制，达到支持绝大多数类型的数据接入的目的。

## 接入方式

对不同类型的数据，数据接入的方式也不相同，大致可分为两种：

* **数据采集**  
  对于日志、数据库等数据，接入模块需要提供“主动采集”的功能；
* **数据接收**  
  对于实时型数据，接入模块需提供“被动接收”的功能。

接入模块需具备接入其他更为复杂应用场景下的数据源。

## 性能需求

以“主动采集”的方式接入数据源时，需考虑到采集器（或采集模块）不能占用“过多”的目标主机资源，包括CPU、内存、网络带宽等。

同时，数据采集器需满足目标业务系统数据生成的速度，以保证数据的时效性。

占用系统资源与采集器的性能之间，存在的相互制约的矛盾，系统设计时需要综合考量这两方面的需求。

# 存储模块

本章节描述DataCenter产品中，对存储模块的具体需求。

## 数据特点

DataCenter产品需要适应各种用户、行业的各种类型数据，其数据量已达到“大数据”的基本标准（当然，并非严格意义上的大数据）。根据应用场景及实战经验总结，需要存储的数据具有以下特点：

* **不变性**  
  用户侧接入的数据，基本上为“不可变”数据，即数据生成以后，一般不会出现变更（增加、删除、修改等）操作需求；
* **时序性**  
  数据产生时一般会携带生成数据时的时间戳，即为数据的时序性。系统设计时，同时需要考虑“非时序性”数据的处理方式；
* **实时性**  
  用户接入的数据或者用户需要通过DataCenter分析的数据，一般为最近一段时间内（例如1个月）的数据。而历史数据（例如1年前的数据），一般用来作为参考。

根据数据特点、存储压力、检索性能以及特殊用户的需求，产品设计时需要考虑以下几个方面：

* **横向扩展**  
  大数据存储的基本要求，满足不同用户的需求；
* **实时检索**  
  低检索（搜索）延迟、精确检索结果，满足数据实时性及有效性的需求；
* **离线存储（长期存储）**  
  满足数据回溯、恢复、对比分析等需求。

## 实时存储

对于时序性数据，一般分析、检测的对象为三个月之内的数据，而三个月之前的数据一般用作对比分析中的参考值。所以对“最近”的数据需要实时存储，缩短数据检索的响应时间，满足分析结果的时效性。

实时存储可以提供较高的性能，但同时对系统资源要求也较高（例如大内存），用户侧的使用成本也会相应提高。系统设计时要充分结合实时存储的需求及应用场景，在性能与资源之间进行取舍，达到某个产品与用户都能接受的平衡点即可。

实时存储模块需要同时满足“时序性”与“非时序性”数据的存储要求。

## 离线存储

离线存储又称长期存储，其目的是对数据进行持久化，在数据需要回溯、恢复时提供数据来源。离线存储的数据访问频率较少，一般在业务出现问题或者需要追溯某个具体问题时，才会有访问需求。

离线存储的数据，格式上不必与实时存储统一。只要满足当用户需要访问时，能通过一定的方法将数据恢复到实时存储模块即可。DataCenter产品允许，同时也提倡对“实时存储”与“离线存储”提供差异化的解决方案。

# 分析模块

本章节描述DataCenter产品中，对分析模块的具体需求。

## 分析方式

分析模块需提供足够的智能化分析功能，充分种用“大数据”优势，满足专业、非专业人事的分析需求。包括：

* **定制化**  
  分析内容、分析步骤等的定制化，满足各类人员的分析需求；
* **自动化**  
  根据预定的规则进行自动分析，减少用户的工作量；
* **可视化**  
  对于非专业人员（或业务员），提供可视化的分析工具是最明智的选择。

根据数据的实时性，分析方式又可分为“实时分析”和“离线分析”两种，后续章节会作出相应的描述。

## 实时分析

实时分析是指对“实时存储”的数据进行分析，其分析方法、粒度、维度、范围以及参考值的选取等都存在一定的限制。一般而言，实时分析使用“规则与大数相结合”的方式进行，不妨称之为“规则2.0版”。

实时分析所得出的分析结果，一般有三种用途：（1）实时监控；（2）实时统计；（3）实时报警。实时分析需满足以下需求：

* **定制分析规则**  
  用户可参与分析规则的制定，或修改分析规则以适应不同的应用场景；
* **定制分析模型**  
  用户参与分析指标、分析维度以及时间模型等的定制；
* **定制输出类型**  
  用户可以自定义分析结果的输出，例如报警通知某负责人；通过大屏进行展示等。

实时分析模块在方法（或规则）的制定上，需要考虑增加最大化的人工参与；而在分析的执行上则考虑避免人员的参与。

## 离线分析

离线分析的数据规模要远大于实时分析的数据，而在实时性上并没有太高的要求。离线分析就同时支持“规则”与“机器学习”两种模式，如果技术允许，建议通过分布式计算（云计算）应用深度学习技术。

具体需求根据实际用户的需求而定，本模块可脱离DataCenter产品而单独存在。

# 预警模块

本章节描述DataCenter产品中，预警模块的具体需求。其中“预警”包括预测和报警两个方面的需求。

## 预测功能

用户通过DataCenter产品分析所得的结果，能够对安全、业务等进行预测，同时产品的可视化模块需对预测功能有一定的展示。

例如，可根据过去两年同期的用户访问量，来预测当前可未来几天的用户访问量。

## 报警功能

对于异常、安全等分析结果，提供必要的报警功能，包括：

* **通知责任人**  
  以邮件、短信等通知运维人员或相关责任人；对接用户的报警管理系统等；
* **执行用户控制策略**  
  对接用户系统的认证中心，执行预设的控制策略等；
* **执行防火墙策略**  
  对接用户的防火墙，执行网络安全策略等。

以上列出了常见的报警输出需求，详细或更具体的需求，需要根据用户类型、系统运行环境等因素而定。

# 可视化模块

本章节描述DataCenter产品中，对可视化模块的具体需求。

## 数据无关

产品对数据进行可视化展示时，必需具备数据无关性，即用户接入的任何格式、任何内容的数据，均可通过DataCenter产品进行可视化。为满足这一需求，可视化模块在设计时必需满足：

* **不依赖**  
  不依赖用户接入的任何数据；
* **不假设**  
  不对用户接入的数据作任何假设；
* **不限制**  
  不对用户接入的数据作任何限制。

当满足“不依赖”、“不假设”、“不限制”三个要求时，数据可视化才有可能达到“数据无关”的要求。

## 数据分析

可视化模块提供必要的可视化分析功能，满足“非专业”人员的使用。数据分析功能提供最基本的两种分析方法：

* **关联分析**  
  使用相同的过滤条件，分析不同维度、不同数据源、不同时间范围等的功能；
* **下钻分析**  
  提供一种“自上而下（或总-分）”的分析能力，为分析人员提供从统计数据一步一步分析到详细数据的功能。

同时可视化模块也需关注，如何辅助用户进行分析规则的制定。

## 高可定制

可视化界面的“所有”内容均支持定制化，包括：图表、菜单、页面跳转等，详细内容请参阅可视化功能需求文档。

## 大屏展示

支持大屏展示及定制。

## 自动报表

根据配置自动成生报表。