技术白皮书

DATAFOR
数据可视化与分析

Pentaho PBA Plugin

上海数为信息技术有限公司
marketing@datafor.com.cn

目录

Datafor 数据可视化与分析

适用于 Pentaho 的“数据可视化与分析”

1.简介

1.1. 背景

随着信息技术的发展和应用，人类进入了一个大数据时代。数据，已经渗透到当今每一个行业和业务的职能领域，成为重要的生产要素。物联网的持续壮大、数据量的快速增长，越来越多的企业迎来业务数字化转型，对海量数据分析的需求越来越多，数据分析技术也迎来了一些新的挑战和变革。

2013 年前，以 IT 部门为主导的传统 BI 一直是市场的主流。Gartner2012 年、2013 年、2014 年、2015 年连续 4 年的数据分析的研究报告，提到的唯一的一个共性，探索性分析已经成为 BI 选型的唯一选择。Gartner2019 年报告指出，增强型数据分析、持续型智能和可解释的人工智能(AI)是数据和分析技术的趋势。

1.2. 设计目标

■ 面向未来的智能型数据可视化与分析工具

✓ 增强型数据分析

利用机器学习与人工智能改变分析内容的开发、使用与共享方式，以提高数据分析的效率和准确性。

✓ 实时分析

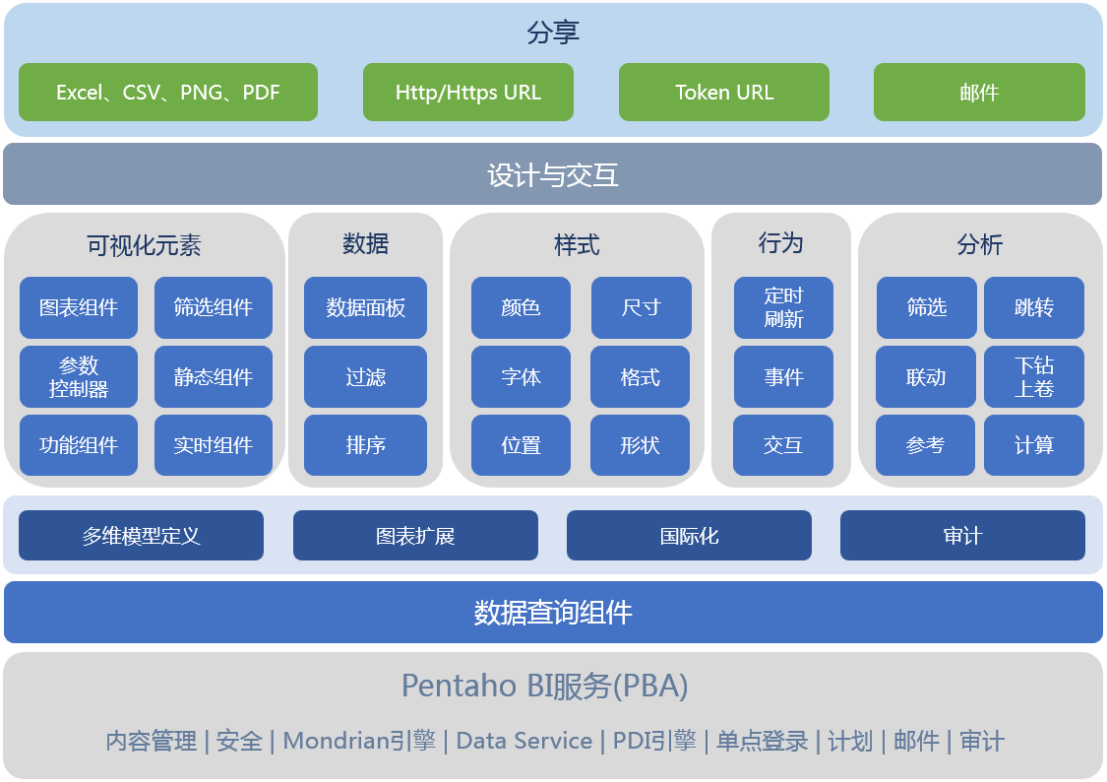
实时摄取设备、传感器、日志文件、金融交易系统等海量数据，并将数据实时可视化，发现问题并分析问题原因。

■ 适合每个人的自助分析

✓ 具有开放性、可嵌入性和可扩展性。实现快速部署、数据源集成、高性能计算、探索式分析，确保开发人员和业务用户都能轻松地将数据转化为价值。

1.3. Datafor 的组成模块

Datafor 基于 Pentaho PBA 设计，以插件形式与 Penaho PBA 集成。主要有数据查询组件、服务组件、数据可视化、分析、嵌入和分享等模块组成。

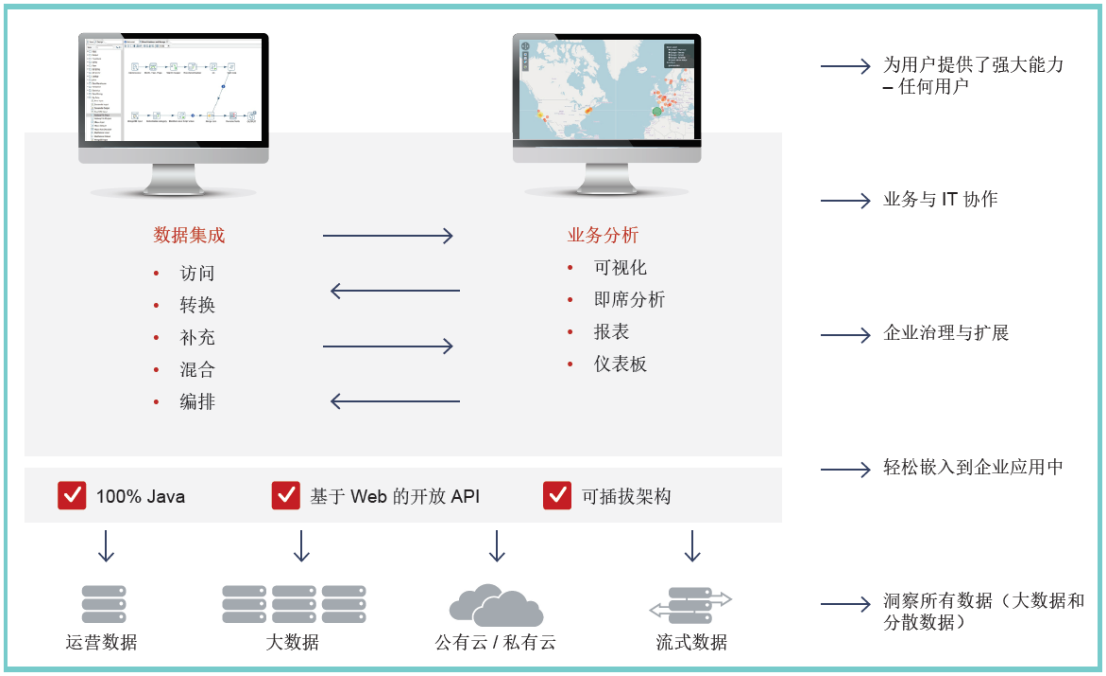


- **数据查询组件**：封装数据查询对象、解析分析引擎返回的结果。
- **服务组件**：定义多维模型、可视化组件扩展服务、语言国际化服务、日志和审计输出等。
- **可视化**：可视化元素渲染、页面布局等。
- **分析**：数据多维分析、预警、分类、趋势等
- **嵌入和分享**：导出功能、嵌入 URL 生成、自适应设置、发送邮件等功能。

1.4. 为什么基于 Pentaho 设计

Pentaho---创新的大数据集成分析平台，是世界上最流行的开源大数据集成分析平台。Pentaho 的开放性，可嵌入特性，既可以有效利用企业现有的数据基础架构，也能够适应企业未来数据分析应用和技术的变化。

Pentaho 主要由两部分组成：Data Integration（PDI）和 Business Analytics（PBA）



图来源：hitachi vantara 公司

■ Data Integration（PDI）

PDI 是一个端到端的实时数据集成工具，可以支持不断融合不同数据形态，并且支持不断进化的大数据架构。PDI 支持网络数据，位置信息，网页数据，社交媒体数据等结构化和非结构化数据的访问和数据集成处理。PDI 图形化设计器和丰富的预置组件简化了数据管道的创建。通过 PDI，无需任何编程，轻松实现大数据集成，提升团队生产力。

■ Business Analytics（PBA）

Pentaho 业务分析软件 PBA 领先的数据可视化与分析套件，PBA 提供了现代化、高响应能力和基于 Web 的直观界面，可帮助业务用户发现并探索几乎任何数据。借助全面的分析工具，用户可以创建报表和仪表板，并且查看和

分析不同维度的数据，而无需寻求 IT 人员或开发人员的帮助。同时，IT 部门可对整个企业进行安全、可扩展和受管制的分析。PBA 可以部署在本地或云端，并且可以无缝地嵌入到其他软件应用中。

Pentaho 的优势

■ 数据集成

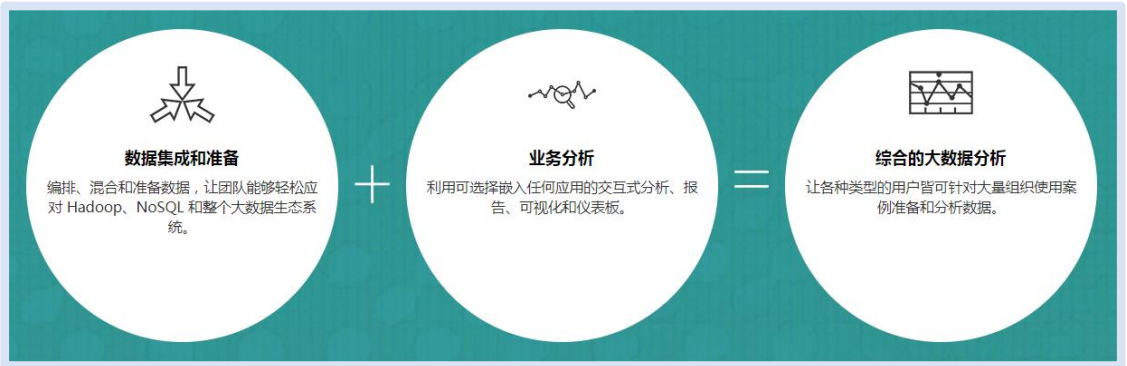
在从边缘到云端的基础架构中管理和提取日益增多的数据量方面，企业面临的挑战越来越大。借助 Pentaho 数据集成 (PDI)，企业可以访问复杂和异构数据源中的数据，并将其与现有关系数据相结合，以生成高质量、随时可用的分析信息。

■ 大数据

Pentaho 平台通过大幅减少设计、开发和部署大数据分析所需的时间并降低复杂性，使企业能够从大量不同数据中获取业务价值。Pentaho 涵盖整个大数据生命周期，即从数据提取和各种数据的准备，再到 Spark 和 Hadoop 的可扩展处理，从而实现端到端的分析解决方案。

■ 多云支持

增强开放式、可扩展的 Pentaho 平台的优势，以满足多云、混合和私有云部署的全面需求。Pentaho 的现代数据架构使用单一数据管理工具而简化日益分散的数据架构的管理。



图来源：hitachi vantara 公司

2.功能介绍

2.1. 数据源

2.1.1. 文件和数据库

支持连接几乎所有关系型数据源，如数据库（MySql、Oracle、SQL Server 等）、文本数据源（CSV 文件）、大数据分析引擎 Kylin、Spark、Impala 等。

2.1.2. 实时流式数据

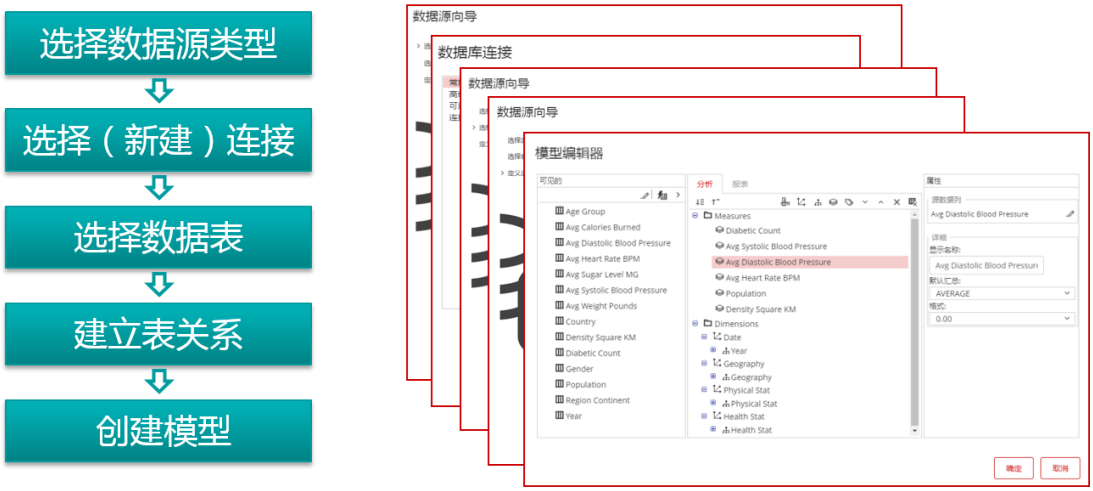
通过 Pentaho 实时流式处理，可以流式处理数据并实时更新仪表板。 通过流式分析深入了解业务和客户活动，例如计费率，服务器活动，网站点击次数以及设备，人员或地理位置服务的使用情况。企业可以通过持续监控和分析，根据需要快速作出响应。流式数据源：

- 应用程序、电子商务网站、游戏应用等系统的日志文件
- 设备、传感器和数据中心仪器的数据。
- 社交网络，金融交易系统和地理空间服务收集的数据。

2.2. 灵活的建模（Pentaho）

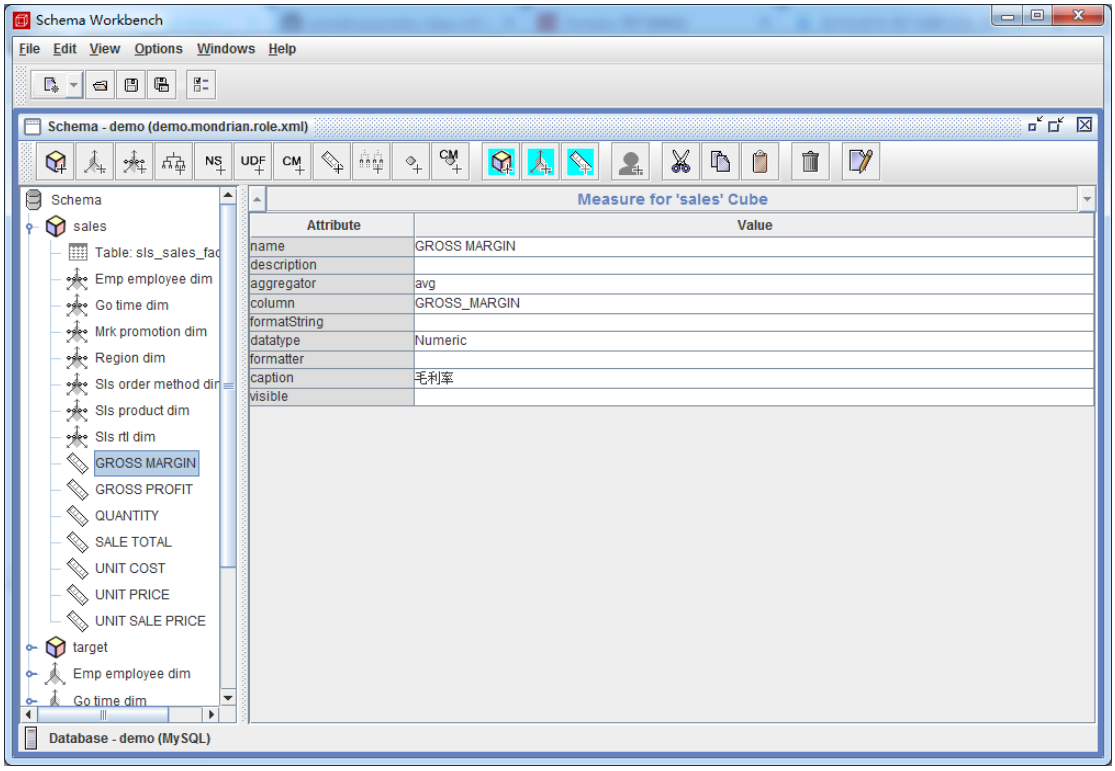
2.2.1. 向导式建模

向导式建模适合非技术人员进行多维数据建模。通过 Pentaho 控制台选择“管理数据源”，通过新建数据源进入“数据源向导”。只需 5 步，轻松建立多维分析模型。



2.2.2. Schema Workbench 建模

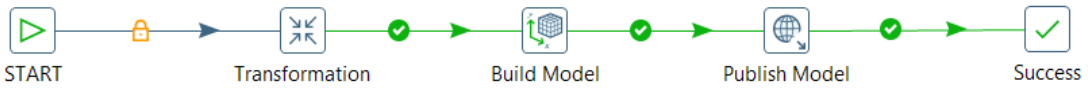
Schema Workbench 适合技术人员进行高级数据建模。支持雪花模型设计、共享维度设计、虚拟模型设计、创建计算指标、数据权限分配等。



2.2.3. Pentaho Data Refinery

在 Kettle 或 PDI 通过 Transformation 和 Job，结合业务创建多维模型。适用于自助式数据服务，结合业务自动化建模等高级应用场景。

- **创建模型**：创建分析模型
- **发布模型**：将模型发布到指定的 BA 服务器
- **标注数据流**：设置模型字段
- **共享维度**：创建共享维度表



2.3. 数据可视化

数据可视化是用图形来表示信息和数据，是当前计算机科学的一个重要研究方向。借助图表、图形和地图等可视化设计元素，可以让您更便捷的查看、解释和理解数据，发现数据中的价值。Datafor 的数据可视化特点：

- 简单易用、智能

- 炫酷的数据图表，丰富的可视化元素，可扩展
- 布局灵活，个性化定制，可配置的主题和样式
- 多屏应用、自适应嵌入

易学易用，丰富的分析组件，拖拽分析与设计，只需几分钟就能轻松制作出精美的个性化数据文档。

2.3.1. 拖拽式设计

拖拽式设计模式又被称为画布设计模。Datafor 通过拖拽方式将可视化元素拖放到画布中，调整位置和大小、绑定数据、设置样式和行为，所见即所得。Datafor 设计器组成部分：画布、可视化组件面板、数据面板、样式面板、行为面板、工具条、组件菜单等。

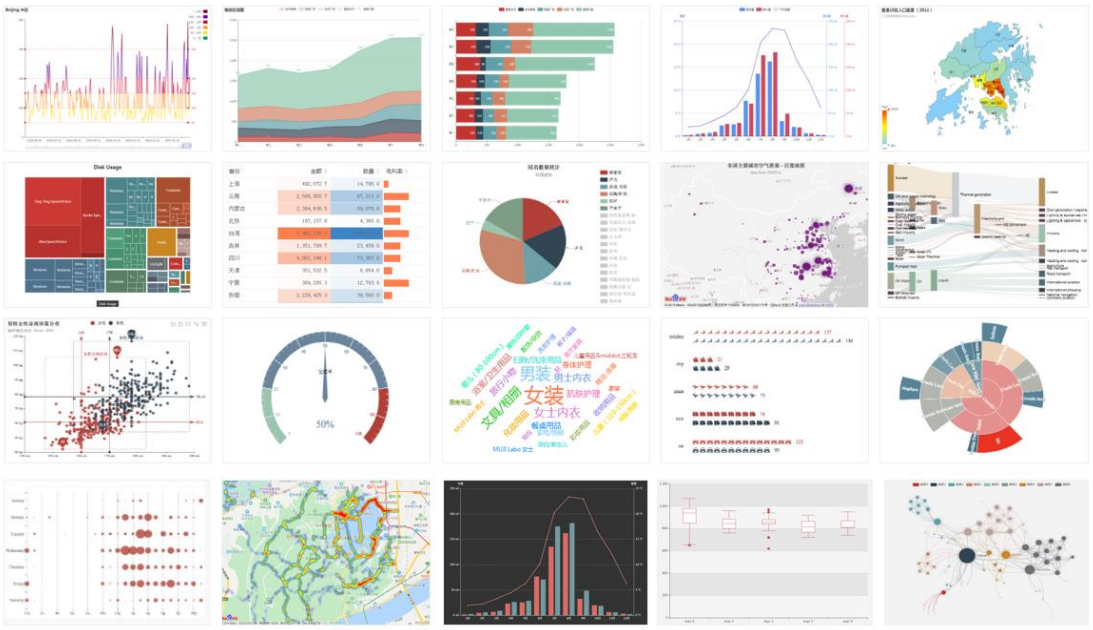
内容	功能
画布	像素级精确定位，层叠布局，布局可视化组件
工具条	刷新数据、布局操作、图表切换、嵌入分享、复制、粘贴、保存、模式切换等
数据面板	可视化组件绑定数据、数据格式设置、排序、筛选、预警设置
样式面板	可视化组件样式设置，包括字体、颜色、尺寸、位置等个性化元素设置
行为面板	自定义事件行为、数据刷新、图表交互行为、组件菜单行为等设置
组件菜单	下钻、数据导出、分析、查看过滤等

2.3.2. 炫酷的可视化元素

■ 图表组件

Datafor 内置了最流行的开源图库 :incubator-echarts。提供了常规的折线图、柱状图、散点图、饼图，用于地理数据可视化的地图、热力图、线图，用于关系数据可视化的关系图、矩形树图、旭日图，多维数据可视化的平行坐标，还有用于表示到达率的漏斗图，仪表盘。Echarts 图库的特点：

- 丰富的可视化类型
- 移动端优化
- 跨平台使用
- 绚丽的特效



■ 辅助组件

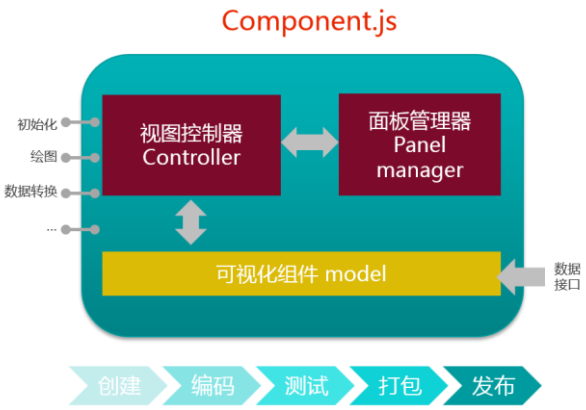
Datafor 提供的辅助组件包括：图片、线条、文本、矩形框、标签页、SVG 图片、超链接、字体图标等。

■ 过滤组件

页面级别的过滤组件帮助用户筛选页面数据。Datafor 的过滤组件和可视化元素采用订阅模式进行数据筛选，并且可以跨模型进行订阅。过滤组件自动将过滤条件传递到订阅此组件的可视化组件的数据查询对象中。

■ 自定义可视化元素

开发人员可使用自定义图表扩展 SDK 创建自定义可视化元素。集成其它第三方图库，例如：D3、Highchart、Fusioncharts；Google 地图、矢量地图、商圈地图；IOT 设备、线路等。



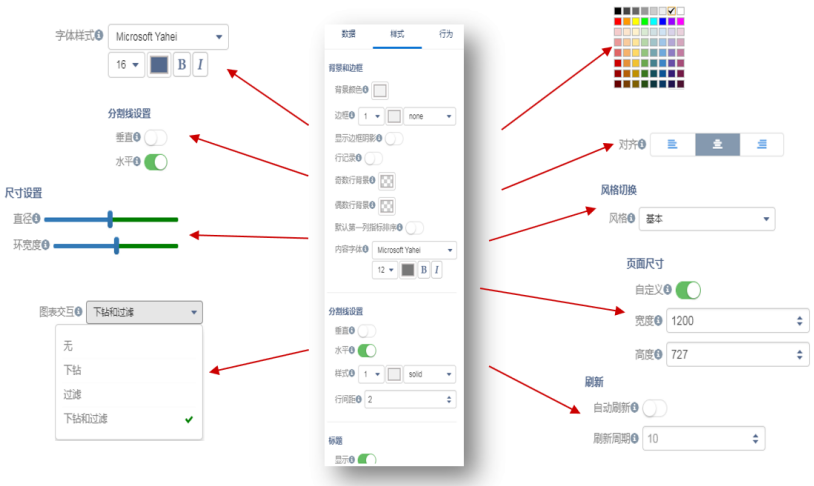
2.3.3. 个性化设置

■ 页面主题样式

使用页面主题样式 ，可以将样式更改应用于整个页面，如使用公司颜色、更改图标集或应用新的默认图表样式。 在你应用页面主题后，页面中的所有视觉元素都会使用选定主题中的颜色和格式设置。

■ 可视化元素样式

每个可视化元素都可以单独调整样式设置，包括颜色、大小、位置、形状、格式、字体等。



■ 自定义 CSS 样式

页面和可视化元素的样式面板上可以选择的样式选项是有限的。当需要某个特定的效果视图时，可以在页面中插入自定义 CSS 样式代码，实现超出样式面板选项的样式调整。

2.4. 数据分析

Datafor 为用户提供了一个易用的、高度交互且直观的数据分析平台，可用于发现和洞察数据。借助全面的分析工具，用户可以跨多个数据源进行多维度可视化和分析，而不依赖于 IT 或开发人员。

2.4.1. 多维分析

多维分析功能使分析人员能够迅速、一致、交互地从各个方面观察信息，以达到深入理解数据的目的。多维分析功能包括透视、下钻、过滤、联动、跳转等。

■ 透视

支持类似 Excel 数据透视表的功能，通过拖拽绑定行列维度，可以迅速的实现数据透视功能，并且可以实现小计、总计、行列转换、颜色背景等功能，数字突出显示、热力图显示、支持将透视的结果导出 Excel 等。

Sales by city

Years		2003		2004		
Country	City	Quantity	Sales	Quantity	Sales	Quantity
Australia		**2,514**	**253,134**	**2,232**	**232,397**	**1,50**
	Chatswood	266	28,397	803	79,202	53
	North Sydney	874	88,984	0	0	59
	South Brisbane	336	37,739	0	0	20
	Glen Waverly	447	37,879	94	12,335	16
	Melbourne	591	60,136	1,335	140,860	
New Zealand		**1,015**	**89,947**	**2,537**	**256,298**	**1,84**
	Auckland	1,015	89,947	1,781	183,977	1,54
	Wellington	0	0	756	72,321	29
Singapore		**0**	**0**	**1,169**	**112,911**	**6**
	Singapore	0	0	1,169	112,911	6
Austria		**872**	**82,118**	**491**	**51,694**	**61**
	Graz	430	43,489	0	0	10
总计		**36,439**	**3,677,384**	**49,417**	**4,987,740**	**19,47**

■ 下钻

多维模型中配置了层级后可以进行数据钻取。层级可以理解为一个顺序有关的文件夹，每个文件夹包含不同粒度的汇总数据，通过高层级文件夹向低层级文件夹按照顺序进行下钻。

■ 过滤

Datafor 提供了多种数据过滤方式，满足不同场景的数据过滤需求。

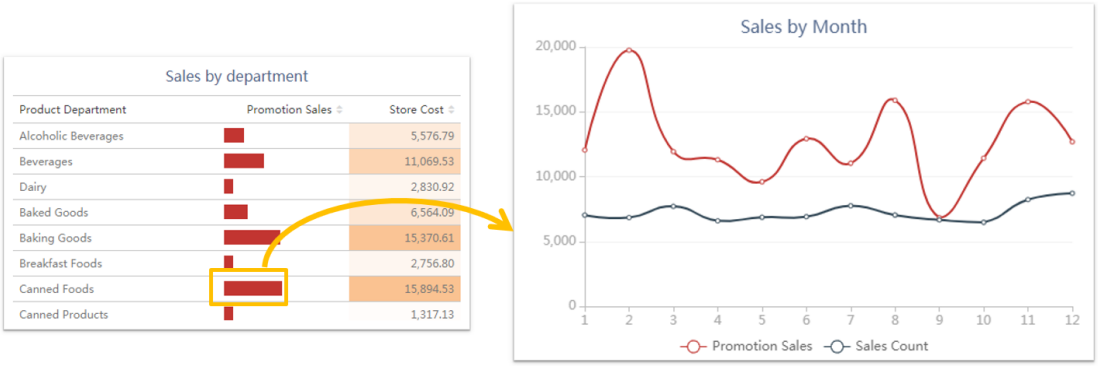
内容	过滤方式•

文本字段	• 列表选择（包含、排除） • 高级过滤（包含、不包含、前置、后缀） • 搜索筛选
时间	• 列表选择（包含、排除） • 动态日期过滤 • 高级过滤（之前、之后，等于）
数字	• 范围过滤 • Top n、Bottom n 过滤

■ 跳转

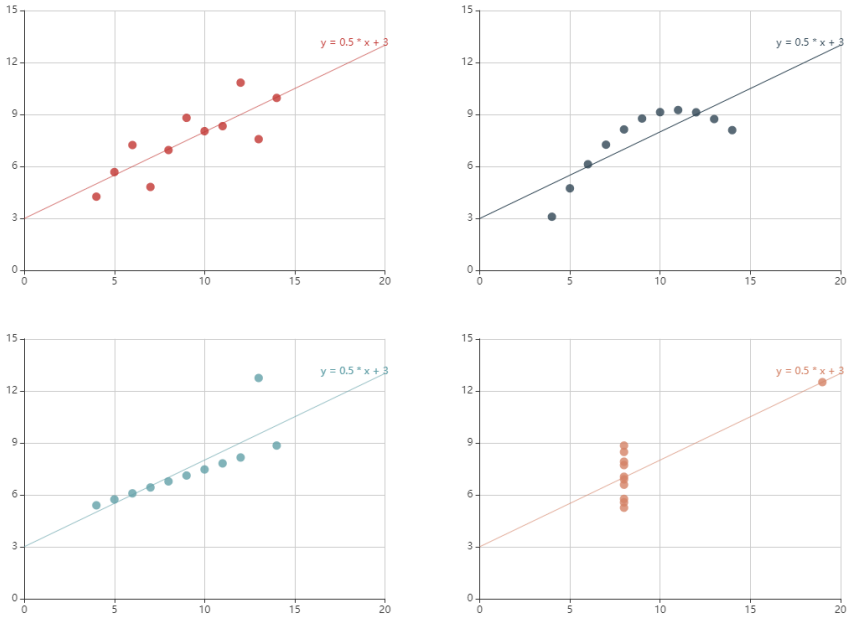
页面跳转功能通常应用在不同视角和粒度查看数据或者从数据跳转到业务操作，从业务操作跳转到数据查看。跳转功能在嵌入式分析中发挥重要作用。

- 页面和页面跳转：点击图表数据，用“模态框”或者“新打开页面”方式跳转到另一个分析页面。
- 页面和外部应用跳转：从外部应用页面跳转到 Datafor 页面，或者从 Datafor 页面跳转到外部应用页面。



2.4.2. 参考线分析

Datafor 为不同的图表提供参考线分析功能，包括常量线、最大最小线、平均线、中位线，聚类线、预测线等。



2.4.3. 高级计算

Datafor 提供近 300 个计算函数，包括：逻辑函数、时间函数、数值函数、类型转换函数、文本函数、统计函数和自定义函数。某些数据组件自带一些常用的计算，包括同比、环比、最小值、最大值、平均、累加等。同时，也支持用户进行自定义计算函数编写。

2.4.4. 预警

根据数据的条件比较结果和阈值设定，通过颜色的变化、高亮显示、标记的变化等方法，以直观的方式标记当前数据的特征，起到数据的预警作用。

2.4.5. 地理分析

Datafor 可利用强大的可视化功能和地图数据相结合，自动将您已有的位置数据和信息转化为可缩放的丰富交互式地图。Datafor 内置了世界地图、中国地图（区县级）矢量地图，并可连接百度、google 等在线 GIS 地图。Datafor 支持色块图、点分布图、热点图等，能够让您的数据在地图上灵活展现。

2.4.6. 实时监控分析

Datafor 使您可以在实时数据流上持续进行数据分析。Datafor 通过收集日志文件、传感器、设备和社交网络等产生的实时数据，并按照逐条记录或基于时间的滑动窗口顺序和递增地处理，进行数据分析和监控。

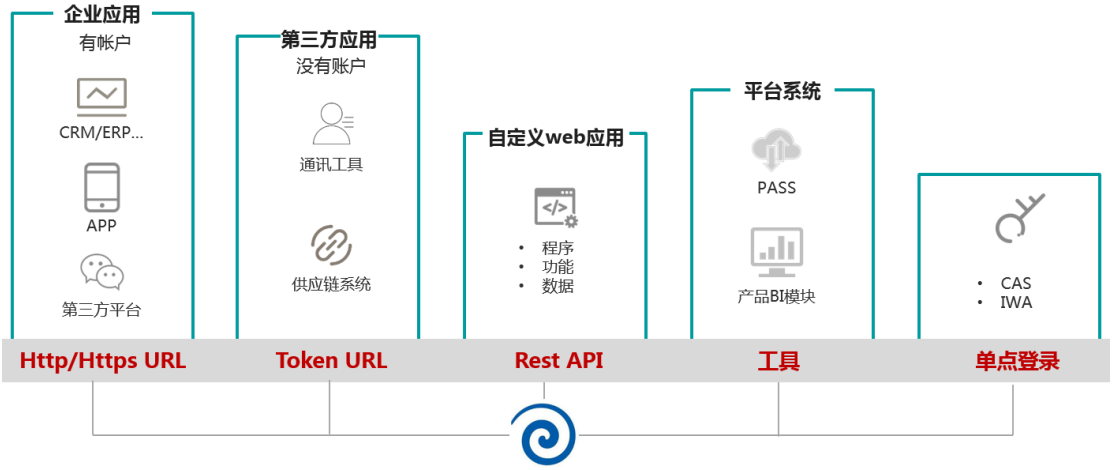
2.4.7. 自定义分析行为

页面和数据图表组件提供“事件接口”，包括：页面加载完成事件、图区单

击事件、渲染前事件、数据获取前事件、数据获取完成事件、渲染完成事件，在合适的事件接口中嵌入程序，加入分析逻辑，实现更高级的自定义分析。

2.5. 嵌入式和分享

根据 Gartner 最近的一项关于数据分析的声明，现在 25%的数据分析活动对应的是嵌入在业务应用程序中的分析。嵌入式分析通过赋予使用者按需获取数据、自主设计报表、自主可视化数据分析的能力。企业内部也可以使用嵌入式分析来提高内部运营效率。Datafor 利用 Pentaho 的开放性和可扩展架构，根据不同业务场景提供灵活的嵌入能力。



■ 分析结果页面集成

将报表和可视化仪表板等分析结果，完美的嵌入到软件厂商的自有产品中，供最终用户使用和查看。数据分析和软件融为一体。最终用户可以同时操作业务系统和做数据分析，毫无违和感。

■ 设计器集成

能够将报表和仪表板设计器直接嵌入到软件产品中，用户在业务系统中，便能创建、编辑和预览文档，真正的实现最终用户自主分析。

■ 移动端集成

将分析分析报表和仪表板集嵌入到厂商自有的 APP 软件中，提供良好移动端的使用体验，在移动端可以实现自适应布局。

■ 用户身份集成

可以直接使用核心业务系统软件用户认证体系，登录系统后，自动根据账号的角色，查看响应权限的报表和仪表板数据。

2.6. 权限

2.6.1. 系统和文件权限

Datafor 利用 Pentaho 权限管理功能授权用户和角色对系统和内容的访问权限。我们支持两种不同的安全性选项：Pentaho Security 或第三方安全管理系统。

■ Pentaho security

使用 Pentaho 用户控制台，使您可以定义和管理用户和角色，并控制存储库中的资源访问权限。

系统权限	管理安全性
	计划内容
	阅读内容
	发布内容
	创建内容
	执行任务
	管理数据源
文件权限	完全控制
	删除
	写
	读

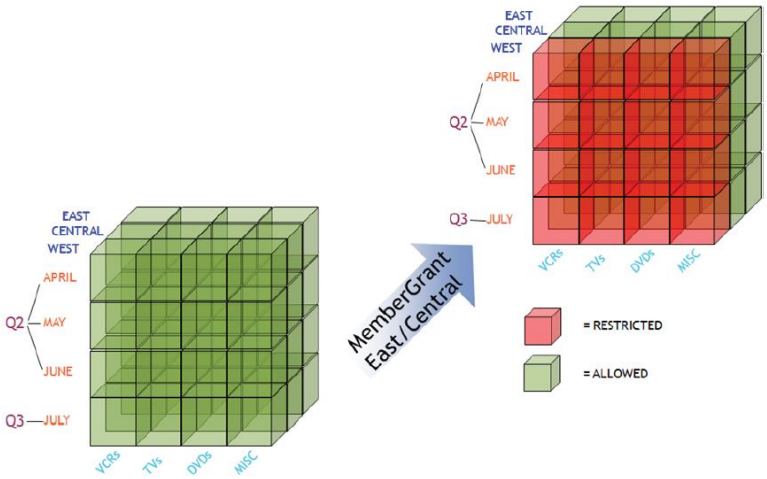
■ 第三方安全管理系统

如果您已经在使用第三方安全管理系统，例如：LDAP， Microsoft active Directory，则可以使用第三方安全管理系统里的角色和用户控制存储库中的资源访问权限。

2.6.2. 数据权限

企业的数据分析报表通常需要进行权限控制，根据报表使用者的角色，决定他可以看到的数据。例如，A 角色的人只能查看 A 部门的数据，B 角色的人只能查看 B 部门的数据，而领导层则可以看到所有的数据。Datafor 支持对角色设置数据访问权限，权限控制粒度可达到行和列级别。

模型	控制角色可以访问哪些模型或者不能访问哪些模型。
维度、层次和层级	控制“数据列”的访问权限，层级中的可见层次粒度。
成员	控制“数据行”的访问权限，成员值可见或不可见。



2.7. 性能

查询引擎（Mondrian）是一个 ROLAP 引擎，在实现 OLAP 查询前你不必做任何处理以生成特殊的数据结构，因此它是“实时 OLAP”引擎。其性能取决于数据源的性能(是否加主键 ,是否使用 SSD 存储 ,hadoop 集群节点数等等)。同时 Mondrian 提供了“汇总表”和“缓存”机制，支持海量数据查询性能的优化。

- **聚合表**

通过创建大量的聚合表，当用户进行比较高级的分析时，无需访问数据量庞大的基础表，只需要在已经形成的实体化视图或聚合表上作进一步的聚合就可以了，这样能够大大提高查询分析的效率，并且减少占用的系统资源。

- **缓存**

为了提高海量数据下的查询响应速度，Mondrian 自动将首次查询的结果缓存到内存中，之后的查询如果命中缓存内容，则不再访问数据库。

2.8. 部署

2.8.1. 集群部署

Pentaho PBA 可以通过创建一个 CDA 集群来支持高可用性和负载平衡。每个 PBA 服务器拥有各自的内容仓库，每个 PBA 对本身的内容仓库做了增加、删除、修改的操作，都会在内容仓库集群汇聚点做个记录，然后其他的内容仓库会同步集群汇聚点上的操作。面对大量用户访问、高并发请求，通过负载均衡配置，将负载（工作任务，访问请求）进行平衡、分摊到多个 PBA 服务器上进行执行。

2.8.2. 多租户部署

Pentaho PBA 支持多租户解决方案工作，或作为多租户服务的一部分进行嵌入。 Pentaho 足够灵活，可以支持各种多租户方法。 利用 Pentaho 的多租户功能可以提供复杂的分析，同时降低复杂性和成本。

3.结论

本白皮书中，我们介绍了 Datafor 数据可视化与分析的各项功能特性和设计理念。Datafor 依托 Pentaho PBA 全面且完全集成的平台优势，结合自身的先进设计和强大的功能实现，是满足企业未来智能分析需要，人人可用的综合商业智能解决方案。

优势

1. 数据集成与业务分析紧密结合。集成 Pentaho 分析数据管道，融合增强型分析，流式数据，适应未来。
2. 使用范围广。满足数据湖分析、物联网实时监控、嵌入式分析、日常数据分析报告制作。
3. 简单易用，炫酷的可视化，可扩展
4. 智能的数据分析、人人会用的数据分析工具

上海数为信息技术有限公司

邮箱：marketing@datafor.com.cn

网站：datafor.com.cn

电话：021-5043 3178