亚马逊AWS官方博客

解密数据编织的核心技术 – 数据虚拟化

企业若想真正利用数据资产的价值，数据和业务密不可分。传统的数据平台，无论是数据湖、数据仓库、湖仓一体等架构，虽然在特定场景和历史阶段上解决了企业数据采集、数据清洗、数据分析、数据可视化等问题，但仍旧难以匹配瞬息万变、海量多元的业务需求。如何支撑决策层敏捷用数，进行实时分析，高效数据探索，最及时最准确地掌握有关数据的全面知识？答案就是新一代数据平台架构——数据编织（Data Fabric）。数据编织与传统数据平台的核心差异就在于“数据虚拟化（Data Virtualization）”，几乎无需搬移或复制物理数据，即可通过逻辑层快速实现元数据的实时连接、整合、消费，赋能数据服务，真正实现“数据不动，价值动”！本文将为您详细剖析数据编织架构的核心技术——数据虚拟化。

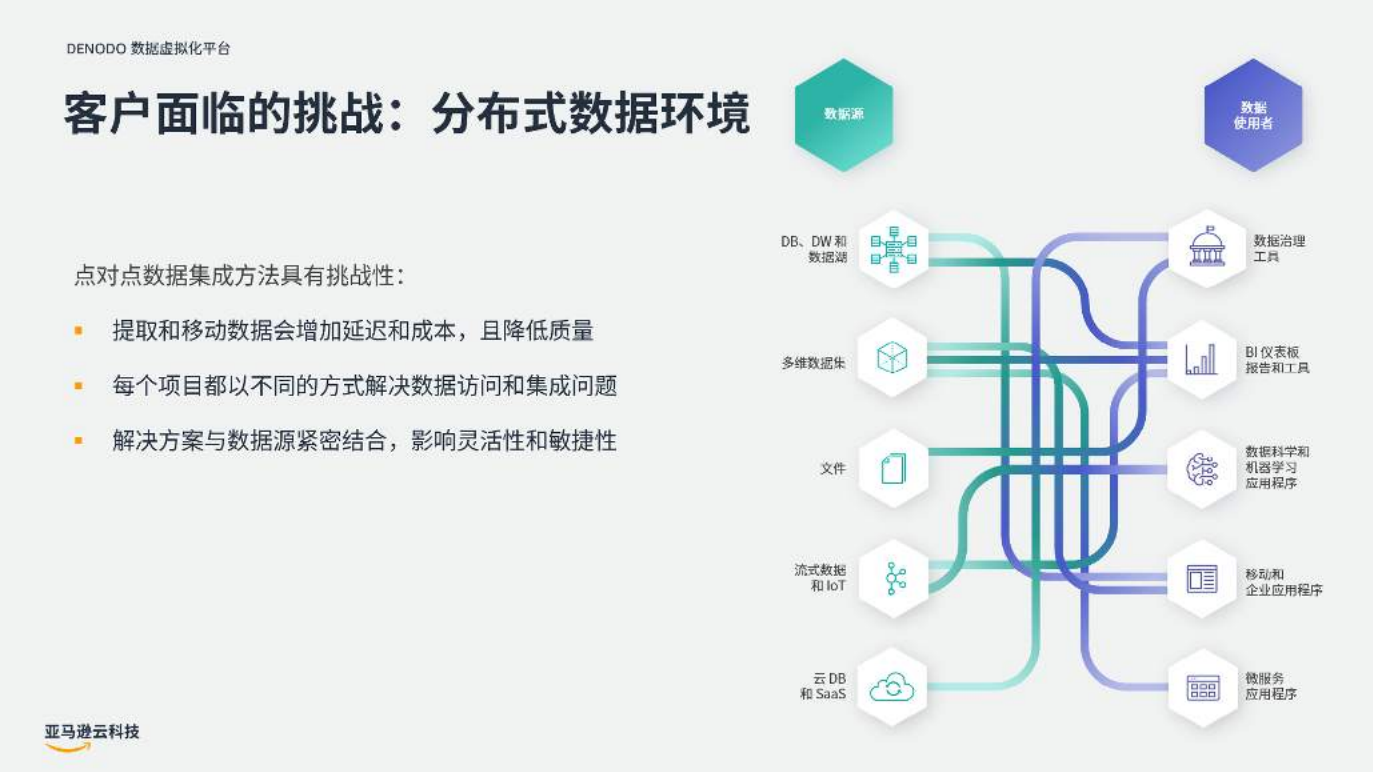
分布式

当今企业数据往往分散在各个异构数据系统之中，比如关系型数据库、NoSQL 数据库、对象存储、数据仓库、关系数据库、Hadoop 集群、SaaS、web服务等... “所有数据都在同一个地方”的设想几乎从未实现。在一个系统中完全复制，集中存储所有相关数据也被证明是不可行的。主要原因一方面是因为公司内部有不同部门，有相对独立的决策权；另一方面不同部门对每项任务都使用最适合的工具，专业的工作交给专业的工具。这就造成了今天大多数大公司在本地和云中维护多个不同的数据仓库和数据湖。如果数据分布在不同的系统中，集成这些数据会很慢，成本也很高。此外，用户不再具有对可用数据的单一访问点。安全和治理也变得更加困难，因为您需要确保在所有系统中应用一致的策略。

数据分散在异构系统里面还有几个常见的原因是：

- 较旧的遗留系统技术难以与更现代的系统互通。
- 本地系统难以与云上系统互通。
- 随着现有系统接近存储容量或性能下降，多年来部署了多个不同的系统。
- 某些系统仅适用于特定应用程序。
- 某些系统被配置为只能由指定的个人或组访问。
- 公司收购具有不同配置系统的其他公司。

数据孤岛使业务用户难以访问和分析组织内的所有可用数据。可能导致不准确的洞察，以及数据不完整或决策延迟。缺乏单一的“真相来源”也会对数据的真实性产生不确定性。



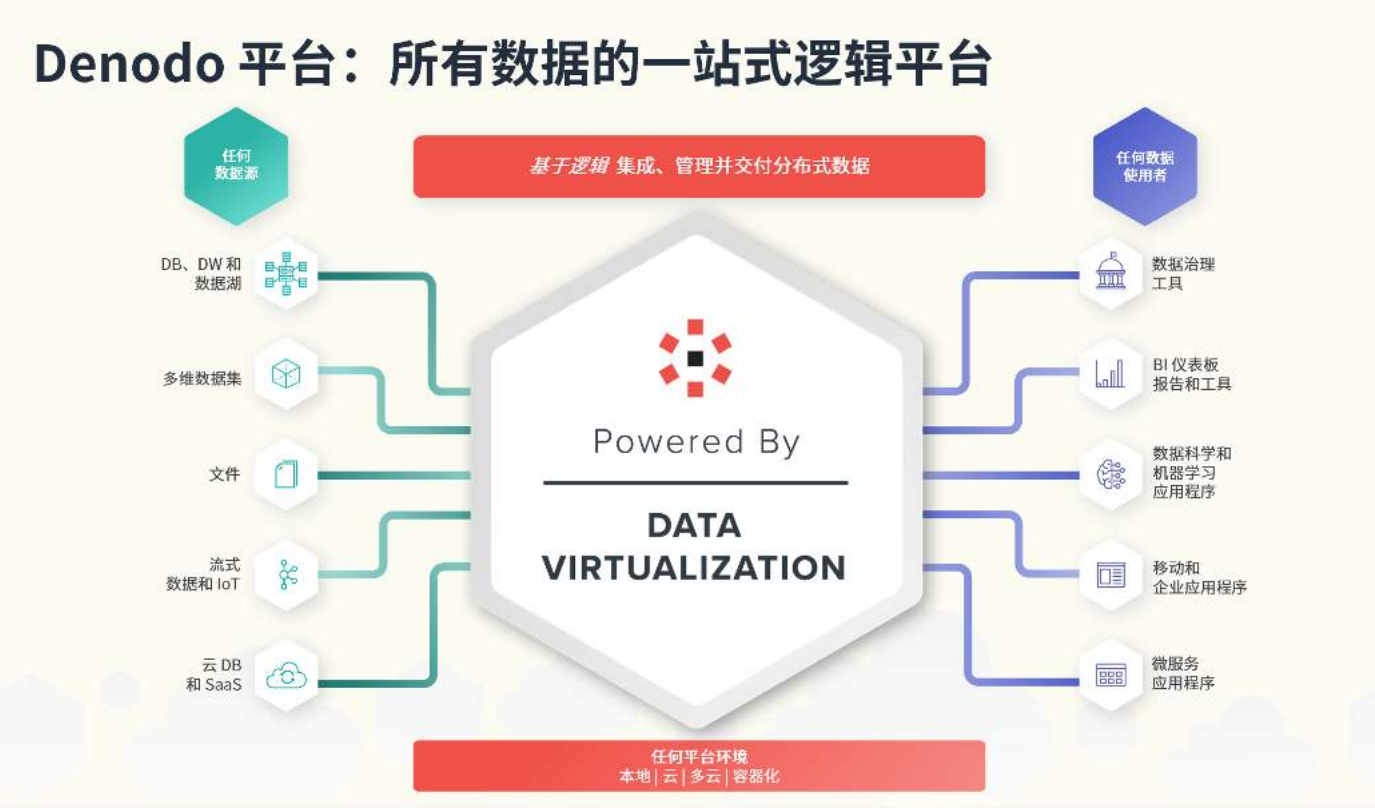
数据治理

数据治理定义了企业如何使用一组预定义的规则和流程来管理其数据的可用性、完整性和安全性。

大中型组织在建立一个集中的、控制良好的数据管理框架方面面临着许多挑战：

- 分散的数据治理。随着公司的发展和数据集的多样化，本地和云系统的数量激增。每个孤立的数据源或消费系统都有自己的数据治理和数据安全机制，没有一个在整个企业之间共享的数据平台。
- 跨部门的数据不一致。大中型组织中的许多部门得使用自己的工具来访问相同的信息，这会导致分析数据沿袭以及信息不一致的问题。
- 数据访问控制。由于有太多孤立的工具和分散在系统中的不同数据访问规则，因此无法为内部和外部用户制定适当的数据访问规则。
- 地理位置挑战。几乎每个分布在多个国家/地区的公司都会多次复制其数据，这会增加成本、影响数据沿袭和质量，并可能违反区域数据保护规则。

通过使用数据虚拟化作为业务用户和应用程序的数据交付平台，可以大大简化这个维护问题。简言之，数据虚拟化平台公开了“逻辑视图”，可以组合和转换来自一个或多个物理源的实时数据，而不需要数据复制。这些“逻辑视图”可以公开符合业务术语的“虚拟数据集”，同时隐藏使物理系统中的数据适应业务约定所需的复杂转换。消费应用程序不需要知道数据所在的位置或每个源系统的技术细节（如查询语言或数据模型）：所有数据在他们看来都属于一个具有一致查询界面的单一系统。



数据虚拟化允许两个或多个数据库作为一个数据库的形式出现，无论它们是在本地还是在云中。

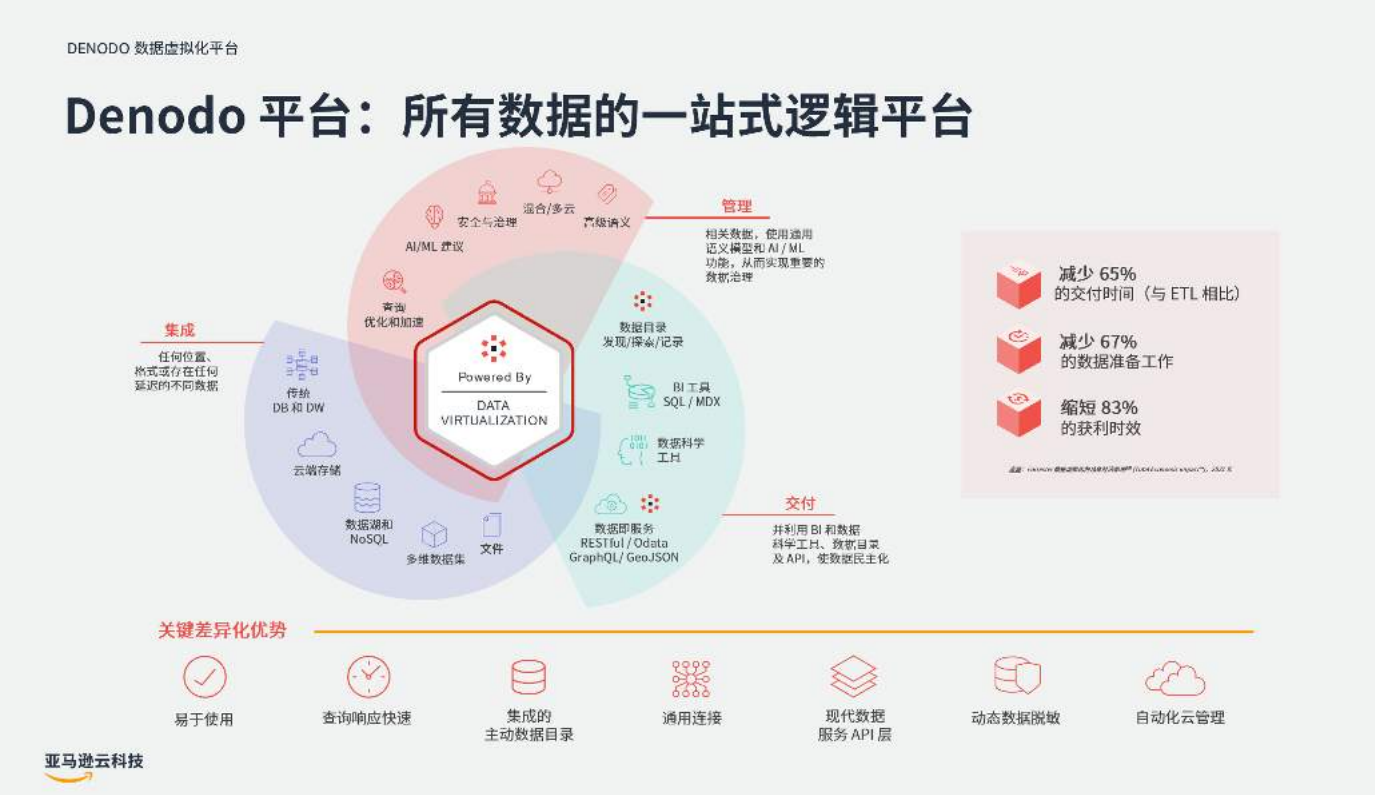
数据虚拟化提供了一个统一的数据集成和交付层，充当 IT 数据管理基础设施与业务用户和应用程序之间的桥梁。它提供了一个单一的入口点，以安全的方式访问任何数据，无论它位于何处或其原生格式。一个统一的语义层，使组织能够创建虚拟模型，以每种类型的消费者所需的形状、格式和结构呈现数据，并使利益相关者能够分层组织虚拟模型，从而鼓励语义定义和数据转换的可重用性。更重要的是，数据虚拟化可以以最适合用户需求的格式向每种类型的用户和应用程序提供数据，与基于数据复制的传统方法相比，成本几乎可以忽略不计。

使用 Denodo 数据目录确定当前数据洞察的潜在来源。

- 搜索可以从元数据开始（“我在哪里可以找到社会保险号码”、“我在哪儿可以找到交易的详细信息”），也可以从数据本身开始（“我们有’约翰·史密斯’的数据吗？”）
- Denodo 的数据目录公开了技术元数据（数据类型、列名等）和业务元数据（描述、类别、标签等）
- Denodo 提供了一个数据索引器和搜索引擎（也可以与 ElasticSearch 集成），以实现对表格内容的基于关键字的搜索
- 可以直接在目录中预览数据，以快速验证数据是否有用
- 用户可以添加自己的评论，以便将来帮助他人

Denodo 平台利用数据虚拟化提供全面的数据和元数据发现和管理功能，包括数据治理、数据沿袭和变更影响分析。数据虚拟化使组织能够跨结构化和非结构化数据源的异构系统创建中央数据访问、数据治理和安全策略。无论数据源或消费应用程序是分布在不同地区还是在本地和云之间，Denodo 平台都可以无缝地促进对数据治理和安全性的集中控制。

- 跨不同系统分布数据的文件数据集成，比基于数据复制的传统替代方案更快、更便宜
- 消费者能够使用任何技术（如 SQL 或者 RestFul API）访问数据 — 数据服务 API 可以在几分钟内创建，无需任何代码
- 跨所有数据处理引擎应用语义、安全和治理策略的单一点，以及数据消费者的单一真相来源
- 实现“数据市场”的能力，业务用户可以在其中找到并访问相关数据，无论数据位于何处
- 数据抽象层，使组织能够将数据从一个位置或系统移动到另一个位置，而不会影响数据使用者
- 智能缓存和加速，支持小数据子集的选择性复制，加速对慢速数据源的查询，以及加速分布式查询的能力



因此，拥有统一的数据交付基础架构（如数据虚拟化提供的基础架构）可以大大简化这种类型的数据目录的创建，这并不奇怪。如果您有一个向业务提供信息的通用平台，那么您可以去一个地方了解哪些数据集可用，其他用户和应用程序如何使用这些数据集，并探索它们提供的数据。此外，一旦您发现了一个有趣的数据集，就可以直接通过数据虚拟化层访问它，而无需担心它位于何处或数据源本机使用什么访问协议。

- 它使用逻辑方法提供对所有数据资产的访问，而不考虑位置和格式，无需复制。复制数据成为一种选择，而不是必要的。
- 它允许定义复杂的衍生模型，这些模型使用来自任何连接系统的数据，并跟踪其沿袭、转换和定义。
- 它以大数据系统（物理数据湖）为中心，可以更智能地利用其处理能力和存储能力。

查询优化

对于这样的数据虚拟化系统来说，最大的挑战之一就是即使必须从多个源传输数据，也要获得良好的性能。因此，有一个强大的查询优化器能够为每个查询选择最佳的执行策略是至关重要的。

查询优化器的工作方式对用户来说是透明的，它负责分析初始查询，并在不同的等效执行计划中进行探索、比较和决策，以获得所需的结果。

当用户向数据虚拟化工具发送查询时，它将向数据源发送一些处理，检索这些部分结果，并将其组合以构建最终答案。这意味着在性能方面，我们需要考虑三个主要因素：数据源中的处理、从这些系统到虚拟层的数据传输以及虚拟层中的处理。

这些技术可以包括：重新排序查询操作以最大限度地下推到数据源（聚合下推、联合下推、连接重新排序），删除或简化操作（分支修剪、外部到内部），甚至在一个数据源中创建临时表，其中一部分数据位于另一个数据来源中（数据移动、MPP 加速）。优化器可以考虑表的统计信息，如行数或不同列值的数量，以及其他有用信息，如索引或引用约束的存在，以估计每个备选方案的成本并做出最终决定。

使用这种方法，数据可以保留在其原始源中，当虚拟化层接收到查询时，它识别与每个数据源对应的操作，向这些分散和异构的数据源发送必要的子查询，检索和组合数据并返回最终结果。这提供了一个要维护的单一访问点和安全性，以及一个将存储数据的实际硬件架构与业务应用程序断开连接的抽象层。这不仅降低了成本，而且为新项目的开发节省了大量时间。

自动评估不同的可执行计划，估计与每个计划相关联的时间和内存消耗成本，以决定哪一个更适合该特定查询。为了实现准确的预估，不仅必须考虑需要从每个源传输的估计行，还必须考虑数据源是否将使用索引或数据源是否会并行处理某个操作信息。

最后，DV（数据虚拟化）系统还需要考虑每个数据源的相对数据传输速度。例如，从本地数据库获取 100000 行的成本与从 Salesforce 等 SaaS 应用程序获取这些行的成本非常不同。Denodo 允许指定数据传输因子，以考虑 DV 系统和不同数据源之间传输速率的相对差异。

完整和部分聚合下推	• 特定查询重写，比如通过分布式、分区式星型结构进行的分析联结-分组查询，经过大幅优化可最大限度减少数据传输。举例来说，多维查询中典型的分组运算将下移至联结和并集以最大化委托。
关系运算优化	• 关系运算、选择、投影、函数等下推至关系源。
数据移动优化	• 采用非常先进的数据移动优化来加速查询，对于一个数据源中的小型数据集需与另一个源中的超大型数据集组合计算的情况而言，这是最佳选择。
自动连接重排	• 自动联结重排以加速执行效率。
查询重写	• 查询重写（消除星型架构中不必要的联结、消除冗余过滤器、约束条件等）
多种联结策略	• 多种联结策略，联结策略由高级启发式方法基于源约束和执行成本来选择，合并(Merge)、嵌套(Nested)、嵌套并行(Nested Parallel)、哈希(Hash)。
基于成本的优化(CBO)	• 基于成本的优化，统计信息：行数，行大小；字段：最大值、最小值、不同值数量… • 索引：可用索引、索引类型（群集、哈希…）数据源信息 等 。

总结

Denodo 与 Amazon Redshift 以及其他本地和云数据源的原生连接可帮助公司利用 Denodo 的性能优化器实时集成 PB 级数据。Denodo 的高级优化器将尽可能多的处理下推到 Amazon Redshift，以提高整体性能。

Denodo 平台架构

Denodo 平台提供 150 多个云和本地数据源的开箱即用连接，包括与以下亚马逊云科技服务的优化连接：Amazon Redshift, Amazon Athena, Amazon-Aurora, Amazon-S3, Amazon-Aurora, Amazon RDS, Amazon EMR 等。

Denodo 是 AWS APN 合作伙伴，加入了 Amazon Redshift Ready 计划。

借助 Denodo 与 Amazon Redshift 以及其他本地和云端数据源的原生连接，公司可利用 Denodo 性能优化器实时集成 PB 级数据。 Denodo 的高级优化器会尽可能将大量处理过程下推至 Amazon Redshift，旨在提升整体性能。

Denodo 平台可即时连接到 150 多个云和本地数据源，优化了与以下 AWS 服务的连接：

亚马逊云科技和 Denodo 的联合方案支持：

- 全球元数据/数据发现。全球信息搜索功能允许任何用户或应用程序通过虚拟数据服务发现、搜索、浏览并最终查询元数据和数据，以检索信息。
- 混合查询优化。最好的 DV 平台结合了实时查询优化和重写、智能缓存和选择性数据移动，以实现对按需拉取和定时批推送数据请求的卓越响应和性能优化。
- 综合业务信息。数据虚拟化提供了集成的信息，同时隐藏了访问不同数据的复杂性。用户和应用程序以他们想要的格式，以实时高性能获得他们想要的内容。
- 数据治理。DV 层作为一个灵活统一的中间层，向用户公开业务元数据。同时，它有助于通过数据分析、数据沿袭、变更影响分析和其他工具了解底层数据层，并揭示底层数据源中数据规范化/质量的需求。因此，DV 可以成为管理信息的“单一参考点”。
- 安全和服务级别策略。从源级别到规范业务视图再到数据服务的所有数据视图都可以在高粒度的视图行列级别上对用户、组和角色进行安全保护和身份验证。进一步的自定义安全和访问策略可以限制或管理服务级别，以防止源系统被过度使用。
- 监测和管理。领先的 DV 平台将包括多个监视器、仪表板、审计日志和管理控制台，以确保 DV 解决方案的顺利运行。它还提供了用于管理集群、高可用性、用户/角色以及在开发、测试和生产之间迁移虚拟数据的工具。
- 数据虚拟化层可以访问其原始位置的数据，这意味着不需要同一数据的多个副本。
- 在直接源访问由于性能原因而不是最佳的情况下，像 Denodo 平台这样的数据虚拟化技术可以轻松地将数据加载到物理关系数据库中，从而实现完全无缝的转换。
- 类似的方法可以用于更高级别的清理和转换。如果需要，数据可以很容易地持久化。Denodo 的引擎将按需安排计算。这将利用数据湖引擎作为一种 ETL 过程，或者更准确地说，是一种智能优化的 ELT 过程。
- 原始数据可以保留在原始源中，只需要将有用的数据带入系统。数据可以在逻辑模型中进行整理、转换、聚合和组合，以便最终只将所需的部分持久化到数据池中。

了解更多

欢迎下载白皮书，快速了解亚马逊云科技合作伙伴 Denodo 领先的逻辑数据编织解决方案：
https://www.denodo.com.cn/document/whitepaper-logical-data-fabric?utm_source=partner_lead_AWS

本篇作者

张元涛

亚马逊云科技高级架构师。负责亚马逊云科技合作伙伴相关解决方案的建设以及合作伙伴生态合作。与合作伙伴一起，根据客户需求，分析其在技术架构层面所遇到的挑战和未来的方向，设计和落地基于亚马逊云科技平台和合作伙伴产品的架构方案。曾在知名外企以及国内领导企业任解决方案架构师。在云以及网络等领域有丰富的经验，对于公有云服务以及架构有深入的理解。

刘闯

Denodo大中华区首席架构师。负责Denodo大中华区的销售和技术服务。有15年以上的工作经验专注于数据管理和数据分析领域，在能源、金融、汽车、制造行业有深入的研究。加入Denodo公司之前，曾在多家知名外企、大型能源央企从事数据相关管理工作，在云计算，大数据，商务智能，机器学习领域有丰富的项目实战经验。