

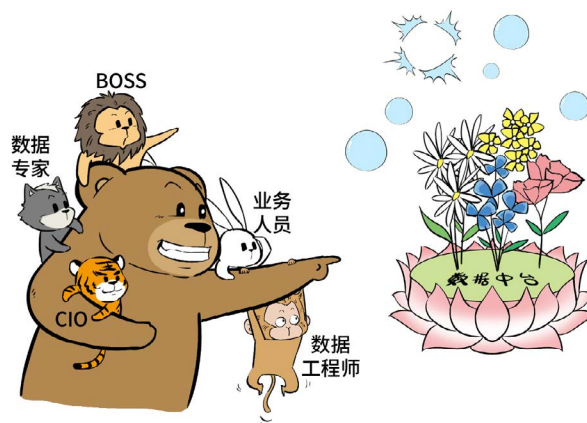
你的数据中台，可能建错了

最近业内流行一种说法：**数据中台逼近炒作峰值**。果真如此么？我觉得其实也差不多



甚至，有些圈里的甲方大佬也产生了一些迷茫：自家的数据中台，是不是建错了，莫非被忽悠了？！

想当年建中台的时候，大家可都是期待满满呀：搭好这个“台”，就能开出美丽的数据之花，企业聚数、用数、赋智的目标都搞得定~



但真正建完上线以后再看看，咦？！也无风雨也无晴。并没有想象中那种化腐朽为神奇、点数成金的效果。

企业的用数之路，依然步履蹒跚...



“蹒跚”也很正常，无论是数据中台，还是数据仓库、数据湖、湖仓一体，这些技术或者架构，都不是解决企业聚数、用数问题的银弹。

他们在不同的历史阶段、针对不同的用数场景，都有积极意义，帮助企业去完成数据采集、数据清洗、数据分析、数据可视化、数据挖掘等各项工作！



甲方“挖湖建仓搭台”，一顿操作猛如虎，也拥有了那么多功能和价值，为什么心里还想要个“自行车”？



那因为，有些用数需求，的确已经解决了，比如针对大规模数据集，集中存储、持续集成，并进行数据展示和数据分析。

可还针对有些诉求，却还差点意思！



比如：决策层敏捷用数、实时分析的需求



“湖仓台”们，往往比较依赖于ETL/ELT手段，进行数据的抽取、转换和加载，然后才能用于分析和挖掘，需要较长的周期。

有时候，很难响应那些高速变化的业务需求，也应付不了某些急不可耐要看数的老板们。



比如：企业数据类型日益复杂，数据源愈发分散



结构化的、非结构化的、实时的、归档的、湖里的、仓里的、云上的、自产的、第三方的.....，企业要面对的是千头万绪的数据整合。

曾经有人说，别管用得上用不上，先挖个湖，全都倒进去再说。实操一下才发现，事情并没有那么简单。



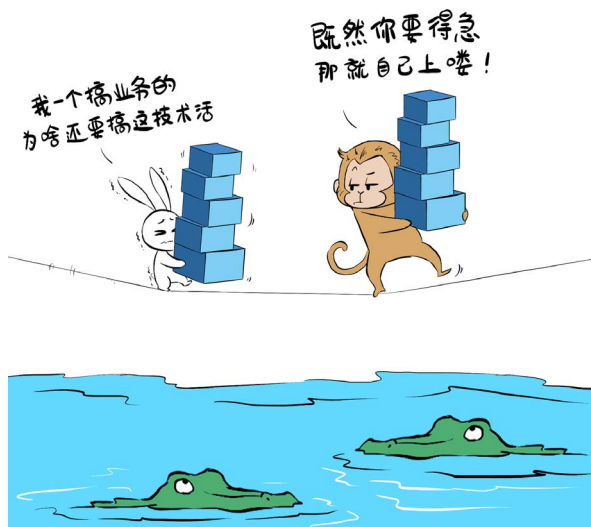
比如：较高的建设和维护成本，
让老板看着不爽的ROI



数仓/湖仓/中台方案，一般都是采用“中央存储库”模式，空间成本、人效、时效，在面对小快灵的敏捷用数需求时，都略显笨重和昂贵。



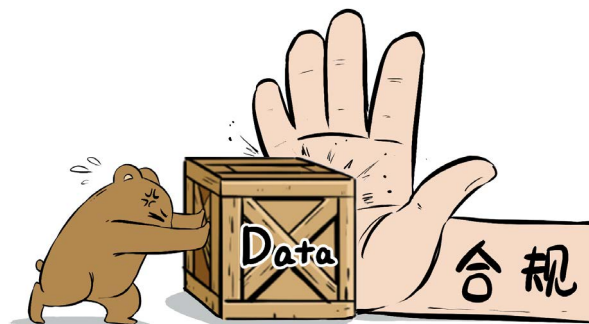
比如：技术门槛相对较高，
业务人员自助分析有点难



这个问题也非常现实，越来越多的业务侧人员，希望自己搞定一些临时性或者个性化的分析需求，不需要太专业的能力，更不需要“求爷爷告奶奶”地去找数据工程师们帮忙。



比如：合规新政策，
给数据移动按下了停止键



这个趋势最近几年愈加明显，随着企业数据安全意识的增强，以及行业监管力度的加大（GDPR、数据保护法...），越来越多的数据，不是你想“搬”就能“搬”的。

因此，很多时候，企业做进行数据分析的时候，只能连接数据（Connect），不能收集数据（Collect）。

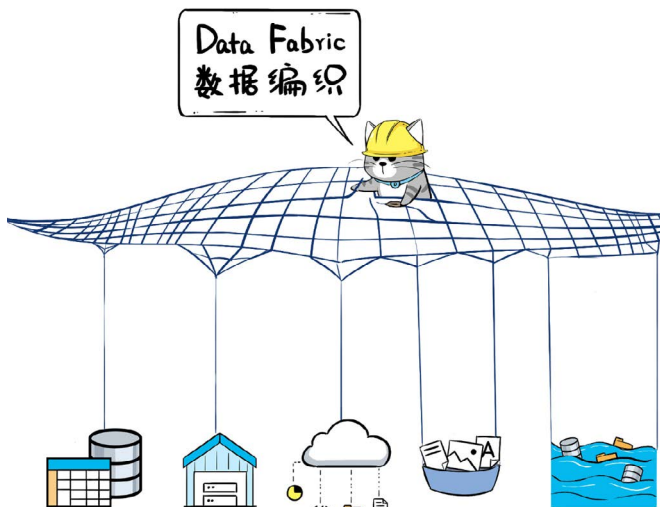
讲到这里，你大概就明白了吧，数仓、湖仓的架构不是不好，国内数据中台的方法论也不是不牛，但企业用数需求千变万化，上面这些新变化，就比较棘手。



那有没有所谓的“自行车”，能够拉上企业一把，把这些新问题一股脑解决了？

还真有一项新兴的热门技术，专门应对这样的场景。

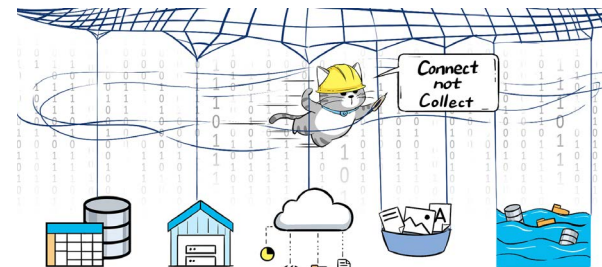
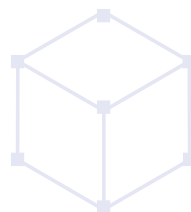
这就是Data Fabric——数据编织技术↓



关系型数据库 数据仓库 云上湖仓 非结构化数据 数据湖

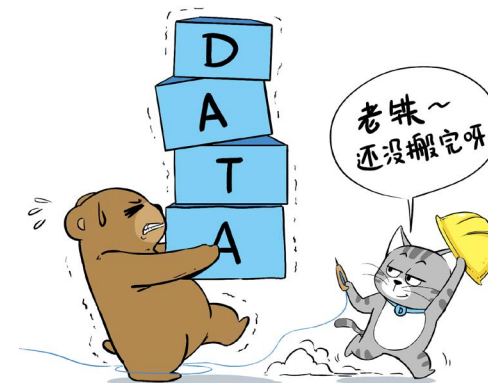
啥？又整出了神乎其技的小词儿，莫不是炒作的新概念？莫慌，听我仔细掰扯下，到底怎么个编织法，究竟能不能解决那些问题。

数据编织技术，最核心的一点是“不搬数据”，而是“连接数据”，通过数据虚拟化手段，实现快速供数、用数。



“不搬数据”，意味着，省去复杂、耗时、耗神的ETL/ELT过程，直接从数据源头“搞事情”。

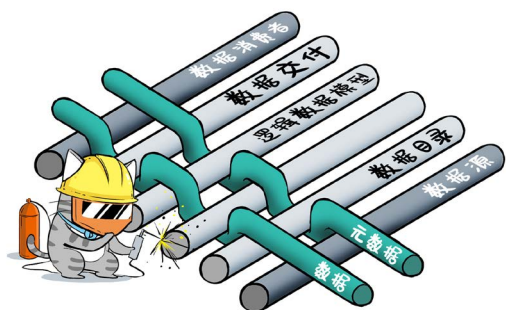
同时，也不需要建设中央存储库，存储庞大的数据集，无论从合规还是成本效率上，都有明显优势。



数据编织改进了数据仓库和数据湖的概念，使用基于网络的架构而不是点对点的连接来处理数据，实现了从数据源层面到分析、洞察力生成、协调和应用的一体化数据结构。

上面说的有点难懂，还是看图吧

通过“**编织**”，编出一个“**Fabric**”，这个Fabric相当于一个虚拟的取数和供数层，屏蔽了各种数据源的复杂性和差异性（位置、类型），然后给上游部门统一的数据供应接口。



据说，采用**数据编织技术**，可以将数据准备时间最高缩短67%，相比ETL速度最高提升65%，6个月内达到盈亏平衡点。（来源：Forrester 2021年报告《**数据虚拟化的总体经济影响**》）

真有那么好用吗？随便“**编织**”一下就能如此高效？！不仅甲方不信，传统湖仓台们，也是相当不服气。



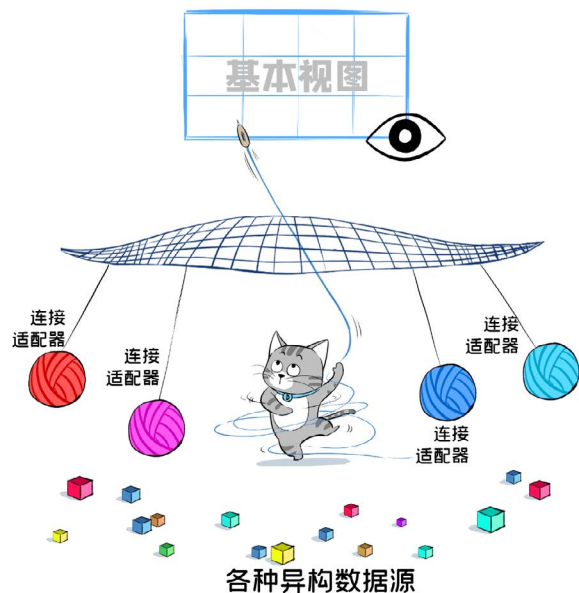
其实，在「**数据编织**」化繁为简的背后，是一系列的关键支撑组件。

通过这些关键组件，数据编织搞成了一张大网，向下连接「**任何数据源**」，向上连接「**任何数据使用者**」。

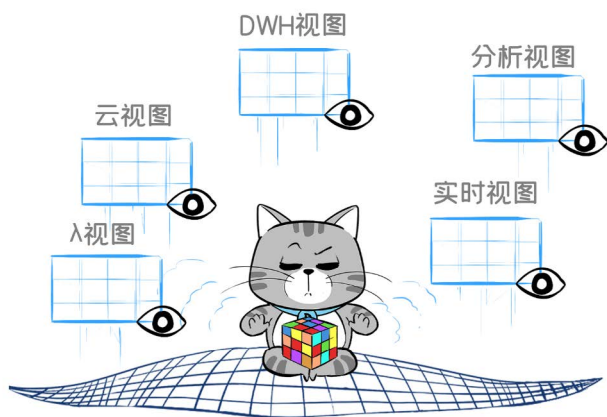


这其中，「**数据虚拟化**」是最关键的一个组件，作为整个Data Fabric的核心，它负责着数据的**连接、整合和发布**。

它在**连接**各种数据源时，不会受位置和数据类型的影响，屏蔽底层差异，并提供针对每种数据源的特定适配器，让用户可以生成“基本视图”，并以表状结构提供给上层使用。



接下来，可以将这些提取自不同数据源的对象整合起来，创建出“虚拟数据模型”，这个模型，对业务侧“友好”，便于数据消费者可以轻松理解和使用。

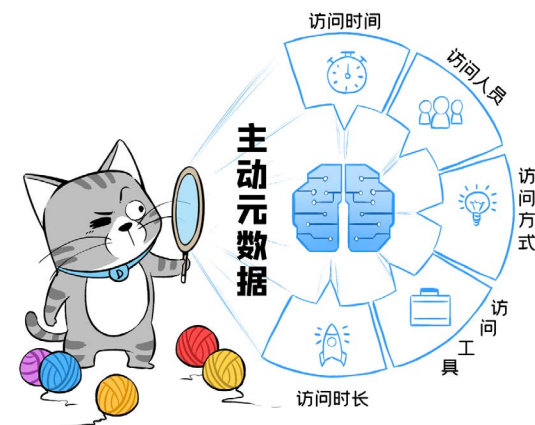


随后，关键一步，是完成数据**发布**，上层数据消费者可以不必关心数据的原始位置，通过统一的“任意门”（基于自己熟悉的API），就可以安全的调用、查询、分析，进行各类用数行为。



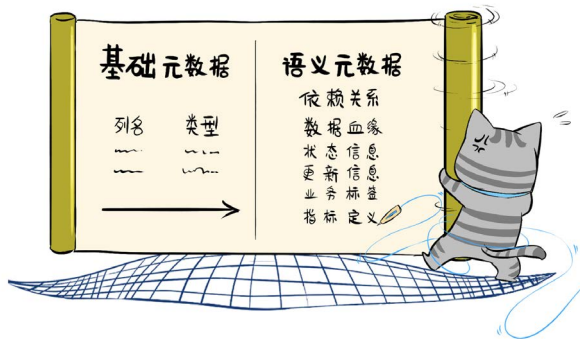
在实际应用中，「数据虚拟化层」拥有“上帝视角”，对所有数据源的访问，都要通过这一层来实现，所以，它可以“捕获”各种访问活动。（数据访问人员、时间、方式、工具……）

这些访问规律经过沉淀和总结，比如引入机器学习和知识图谱，就可以将传统元数据扩展和增强，形成**主动元数据**。



主动元数据管理是Data Fabric的另一个核心组件，用来支撑更智能的数据集成和数据分析。比如数据发现建议、查询加速建议，以及更精细的安全审核、数据治理和管理等等。

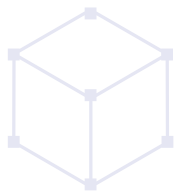
同时，传统元数据还会被进一步进行「语义扩展」，变成**语义元数据**，以便于让上层业务用户更一致地理解底层数据。



说白了，主动元数据和语义元数据，都是对传统元数据的扩展和增强，前者促进了智能化和自动化，而后者则提升了数据的可理解性。

而最终，在面向数据消费者的统一“窗口”处，Data Fabric一般会提供一个强大的**数据目录**，供广大“用数群众”直观、轻松地找到数据。

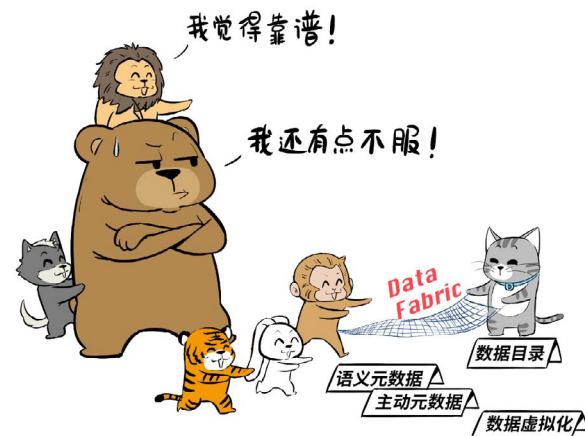
一个优秀的数据目录，可以清晰的展示可用的数据全景、数据血缘，提供高级搜索和个性化建议，数据受欢迎程度以及数据集预览等等。



总之，这种增强版数据目录，可以让非专业人员（比如业务人员），也可以快速上手，降低对数据工程师的依赖。

小结一下，**数据编织技术**是一种逻辑数据集成架构，通过「数据虚拟化层」完成各种异构数据的连接、整合与发布，大大减少数据搬运量和ETL/ELT操作。

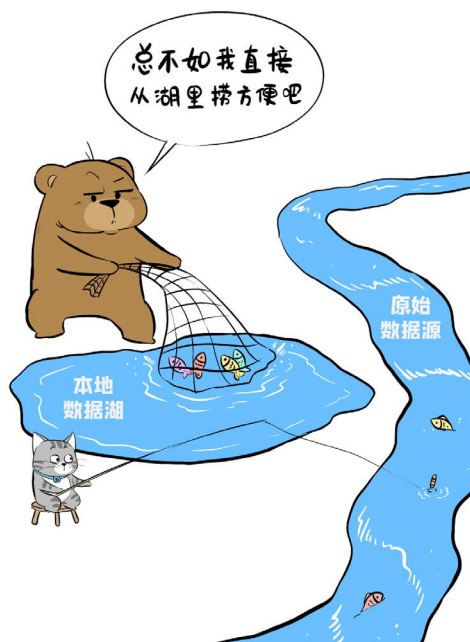
同时，主动元数据、语义元数据和增强版数据目录，可以大大提高数据集成的智能化和简易性，为上游用数者提供最大的便利性。



大体的架构，我们已经讲了个七七八八，可是，有些小伙伴还是将信将疑。

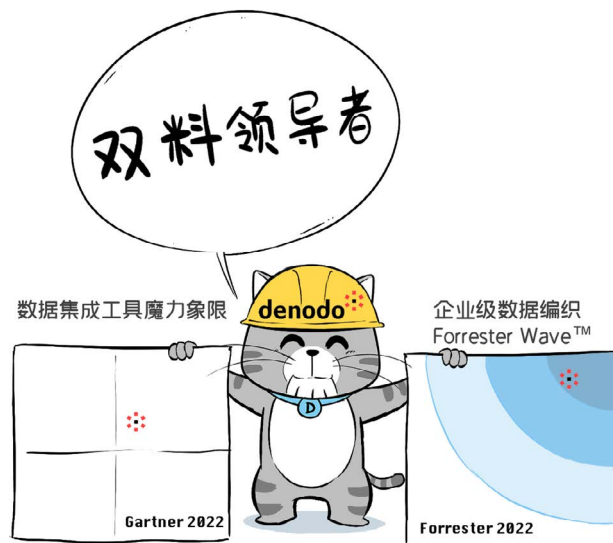
比如：数据编织要远程访问各种异构的数据源，性能怎么能保证？少量数据集还行，面对大数据集能搞定？远水如何解近渴，总不如从本地大湖里直接捞方便吧。





所以接下来，不讲理论，讲讲实战，我们拿数据编织领域的招牌公司举例子，看看人家是如何帮助客户解决实际问题的。

这家公司叫做**Denodo**，数据集成领域的王牌公司，也是逻辑数据编织技术的引领者之一。



D家的数据编织平台，具体如何高效整合数据呢？

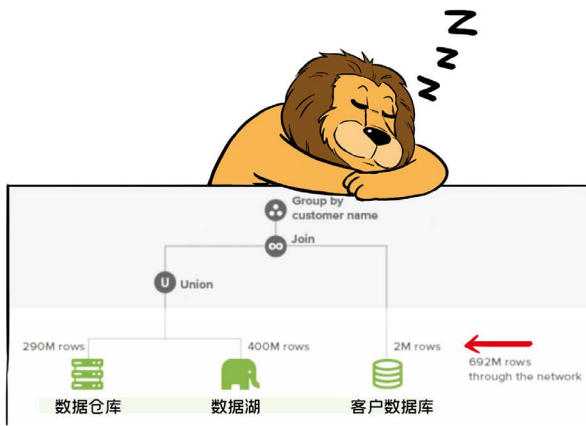
首先，数据编织技术，采用的一般都是“schema-on-read”模式，本身具备非常快速的初始数据加载，更重要的是，Denodo在数据虚拟化层，特别提供了专门的**执行引擎和优化器**。



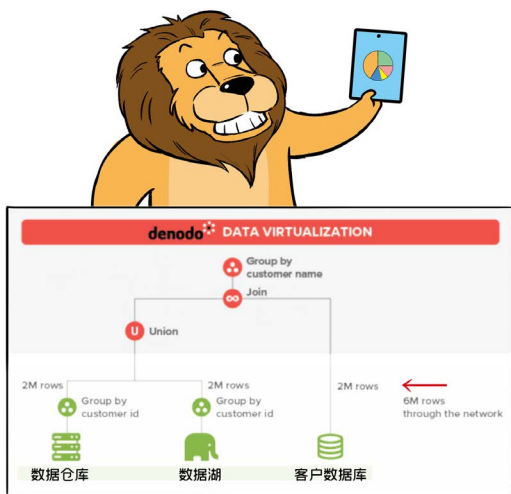
这个关键组件负责制定查询执行计划，并以最优方式检索数据，面对多个数据源，可以连接/聚合和查询重写。

举个例子，某公司有200万客户信息存在CRM中，同时在数仓中存了2.9亿行年度销售数据，Hadoop系统中还有历史销售数据30亿行。

如果老板要求通过客户姓名，汇总查询过去两年的销售额，普通的方法，需要通过网络传输**6亿9200万行**数据，这将是漫长的等待...



而基于Denodo数据编织平台，通过查询优化器的聚合，只需要传输**600万行数据**，省了超过100倍，几乎是立等可取。



如果查询需求进一步复杂，数据集进一步庞大，Denodo平台还有更细致的优化操作。

比如使用缓存或者聚合感知摘要来进一步增强性能，如果合规要求允许数据副本，Denodo平台还可以即时运行ETL/ELT作业。

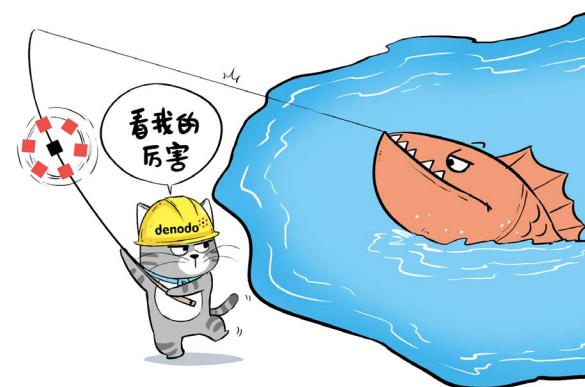


so，在Denodo平台采用了一系列的措施（智能优化和访问加速），即便面对分布式的数据源和大型数据集，也可以表现出优秀的实时性能，让用数人“立等可取”。

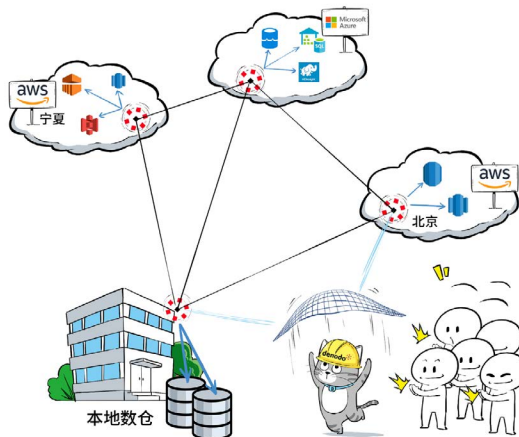
另外，业界诸如湖仓一体化（LakeHouse）等方案，都是存算一体架构，在应对较高的分析负载时，不容易单独扩展。



Denodo的逻辑数据编织平台是典型的存算分离架构，可以方便的独立扩展性能，处理高工作负载，应对上层各种分析和数据挖掘需求。

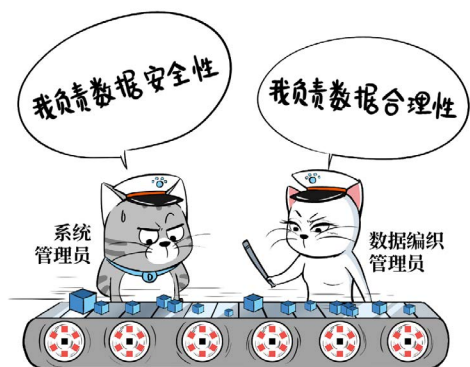


同时，Denodo平台完美适配多云混合架构，无论数据源分布多么“零散”，都可以轻松拿捏。通过内置的150+连接适配器，平台能够搞定各种异构数据。



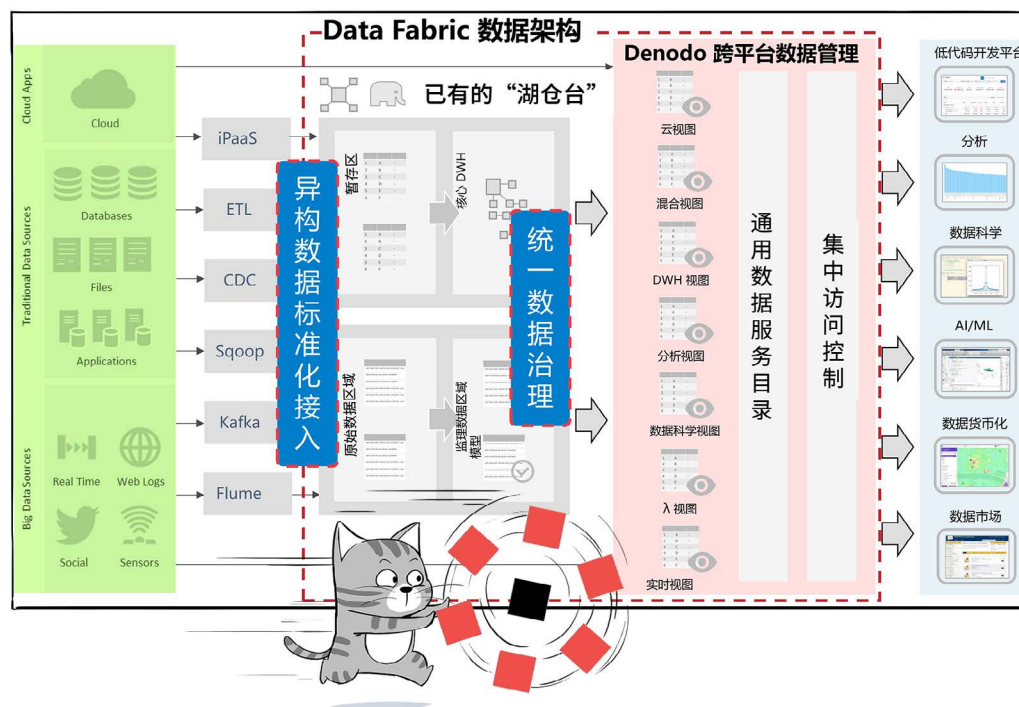
另外，在数据合规和安全层面，企业也完全不必担心，Denodo平台提供集中式安全和治理。由数据编织管理员和系统管理员来进行统一的权限控制（安全权限和业务权限）。

各种业务角色和上层应用，按需取数、用数即可，既不会影响正常业务，也不会造成安全风险。



我们来看一下Denodo数据编织平台的完整架构↓

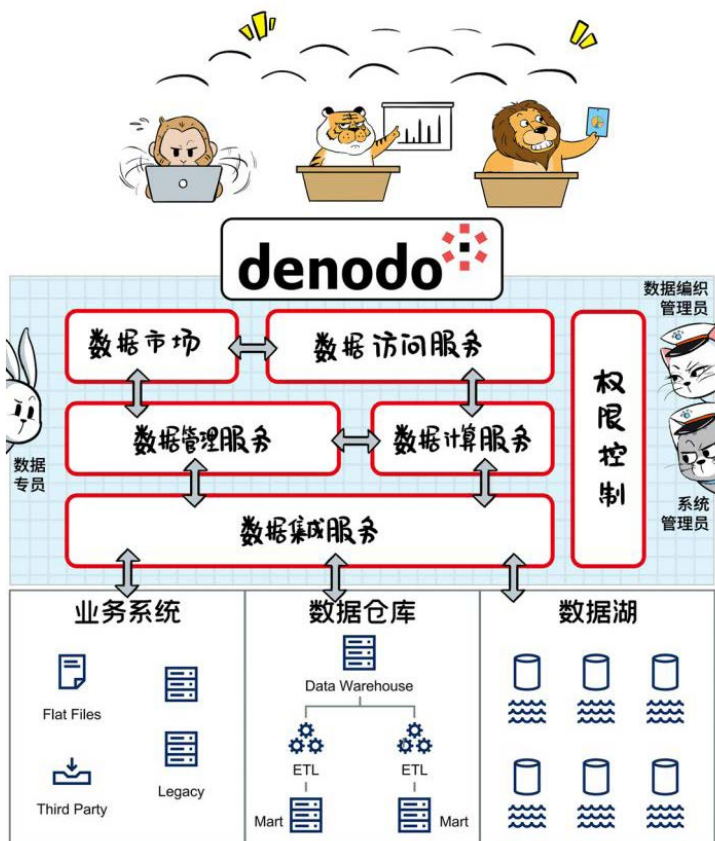
不管数据源如何分散和异构（云上、多云、混合），不管客户的数据基础设施建设现状如何（已有湖仓台相关建设或者相对空白），不管上层的用数需求如何（BI、数据科学、开发、数据交易），都能「编织」在一起，完成高灵活、智能化、自动化的数据集成。



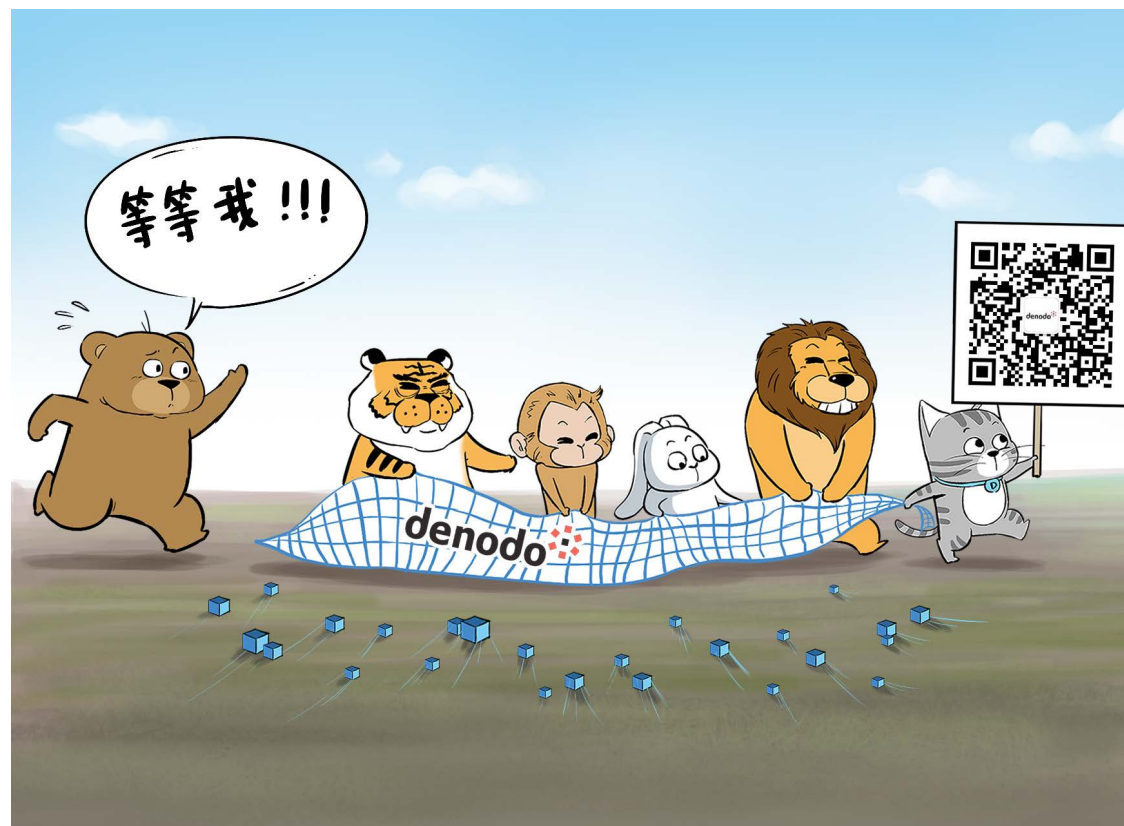
以Denodo为代表的逻辑数据编织平台，正在引领数据集成和数据管理技术的新方向。

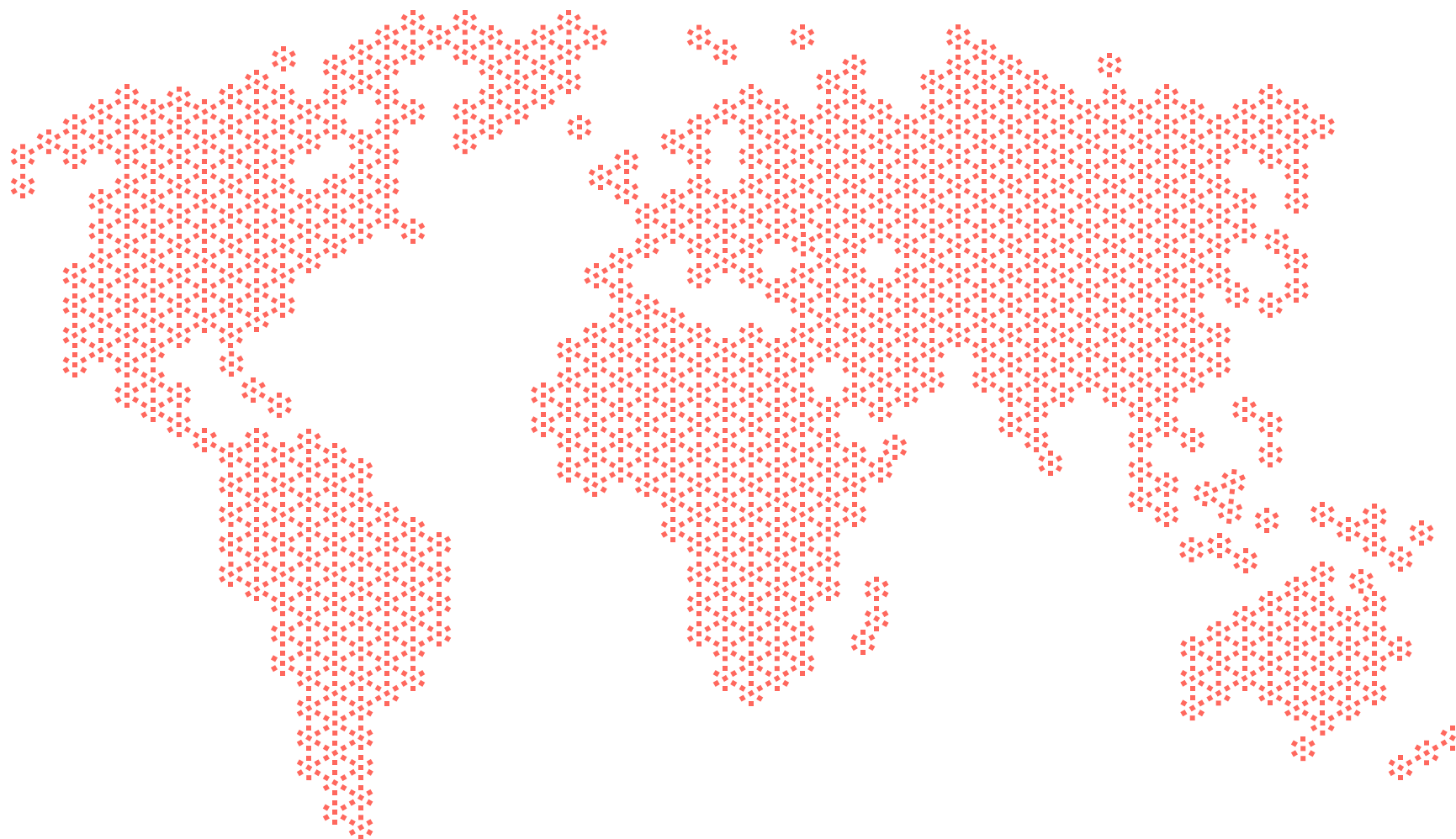
企业数据无缝集成、高效治理，原有的湖仓台的建设成果，也可以被继承下来，并实现价值最大化。

不同角色的用数人员，都能在平台上编织出自己的“幸福感”，老板满意，CIO称心，数据专家增效，数据工程师减负，业务人员舒坦...，这样的“香饽饽”，又有谁不爱呢？



想了解「数据编织」的更多细节么？可以扫码关注Denodo官方公众号，也可以下载一份干货满满的《[逻辑数据编织技术白皮书](#)》。





官方网站 www.denodo.com.cn | 电子邮件 info.cn@denodo.com | 社区网站 community.denodo.com

