

数据空间基础设施数据连接规则研究

殷致云 龙宇希 詹淑欣

广东联合电子服务股份有限公司, 中国·广东 广州 510620

摘要: 本文从数据空间的发展趋势出发, 介绍了其参考架构, 并重点分析了数据空间的核心组件—数据连接器。文中详细分析了数据连接器的功能需求、分类、架构及其各组件的功能和交互流程, 并探讨了其应用现状及未来发展趋势。作为数据空间基础设施的关键部分, 数据连接器不仅能够有效控制数据的使用, 还因其开源和灵活部署的特性, 为数据要素的发展提供了重要的参考和借鉴。

关键词: 数据空间; 数据连接器; 国际数据空间协会; 国际数据空间

Research on data connection rules of data space infrastructure

Yin Zhiyun, Long Yuxi, Zhan Shuxin

Guangdong United Electronic Service Co., Ltd, China Guangdong Guangzhou 510620

Abstract: Based on the development trend of data space, this paper introduces its reference architecture, and emphatically analyzes the core component of data space-data connector. In this paper, the functional requirements, classification, architecture, functions and interaction flow of each component of data connector are analyzed in detail, and its application status and future development trend are discussed. As a key part of data space infrastructure, data connector can not only effectively control the use of data, but also provide important reference for the development of data elements because of its open source and flexible deployment characteristics.

Keywords: Data space; Data connector; International Data Space Association; International data space

0 引言

近年来, ChatGPT、Midjourney 和 Sora 大模型的飞速发展, 凸显了数据在现代科技领域的核心作用, 不仅推动了人工智能技术的进步, 也展示了数据在塑造未来世界中的关键作用, 数据成了与计算资源同等甚至更重要的战略资源, “以数据为中心”的技术体系[1]在数据快速发现、数据可信流通等方面提出了新的需求。数据空间以数据为一阶实体, 定义数据生态系统中各方之间的可信赖数据关系, 基于互操作性与安全可溯的技术架构, 是基于共同商定原则的去中心化的数据生态系统基础设施。目前, 全球主要国家正积极研究和推动数据空间的部署和应用, 以期构建一个安全、可信的数据流通环境, 促进数据资源的高效利用。随着国家数据局向社会公开征求可信数据空间发展行动计划(2024—2028年)的意见, 越来越多的企业将启动以数据为牵引的多种数据空间类型的建设, 包括企业、行业、城市、个人、跨境等, 形成促进数据要素合规、高效流通使用的解决方案和最佳实践。数据连接器在数据空间内的数据共享和交换中起着至关重要的作用, 它们旨在提供数据主权, 确保信任和互操作性, 是数据空间的中的

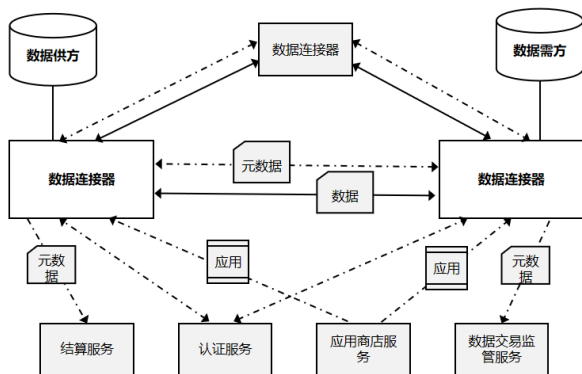
核心部分。

1 数据连接器简介

数据空间的概念最早由富兰克林(Franklin)、哈勒维(Halvey)和麦尔(Maier)于2005年在数据管理社区首次提出, 数据空间可以模拟不同组织的任何数据源(如数据库、CSV、网络服务)之间的关系, 目标是允许成员共享数据提供并传输数据, 同时保持对数据使用的控制。2015年, 弗劳恩霍夫协会联合多家德国企业和组织成立了“工业数据空间协会”, 2016年该协会发展成为国际数据空间协会(IDSA)。由此, 数据空间的概念得以扩展和推广, 涵盖了更广泛的行业和应用领域。国际数据空间协会旨在通过建立一个标准化的框架, 确保数据在不同组织和行业之间的安全共享, 同时保持数据所有者对其数据的控制权。国际数据空间协会通过其“国际数据空间(IDS)参考架构”推动了数据空间的理论创新和实践发展。

国际数据空间参考架构, 包括核心组件数据连接器、认证服务、应用商店服务、结算服务、数据交易监管服务等, 其中, 结算服务作为一个记录服务, 它记录了与清算和计费以及使用控制相关的信息, 包括交易的详细情况、

数据交换的量、时间戳、参与方的身份等,以便于进行审计、计费 and 确保合规性;应用商店为连接器提供应用程序,这些应用程序在国际数据空间生态系统中执行数据转换、清洗或分析等各种任务,这些应用程序可以被连接器用来扩展其功能,以适应不同的数据处理需求;认证服务负责处理识别、认证和授权的任务,从而确保了参与者的身份可以被验证,并且只有授权的用户才能访问特定的数据或服务;数据交易监管服务是一个中心组件,它注册并分发关于数据提供的信息,包括哪些参与者提供了哪些数据,以及使用这些数据的条件,它本质上充当了数据目录的角色,通过确保关于提供和相应使用政策的透明度,帮助参与者发现和理解可用的数据资源。数据供需双方配置相应的数据控制策略后,由数据连接器将数据的控制策略转化为 ODRL (Open Digital Right Language) 描述的形式化语言,与需要进行交换共享的数据资产文件一起流转至对端上层应用系统,并通过第三方权威机构,比如国际数据空间协会、及其授权的认证评估机构,对上层应用系统进行详细的认证评估,确保数据使用策略能得到严格的执行。

图1 IDS参考架构^[2]

在国际数据空间（IDS）参考架构中,最重要的就是数据连接器,数据连接器集成在数据供方、数据需方等各个参与者的基础设施中,在数据空间中担任逻辑守门人的角色,提供自动化连接、合同谈判以及数据空间中数据传输的政策执行能力,具体体现在以下四个方面的控制:(1) 参与者保持对其身份的主权控制;(2) 参与者决定信任谁;(3) 参与者决定数据共享的使用策略;(4) 参与者保持对其部署的控制。

由于数据的类型及其共享交换的场景不同,对于连接器的需求也会不同。基于不同的维度,可以将数据连接器分为以下几种类别:(1) 作为基础框架的数据连接器:此类连接器基于开源代码提供,旨在为数据连接器解决方案的

开发者提供基础框架,不能直接用于数据的共享交换;(2) 开源的或通用的数据连接器解决方案:此类连接器可以通过组件配置和添加自定义扩展之后直接集成到使用者的IT环境中,可以满足特定场景的数据共享交换需求;(3) 专有数据连接器解决方案:此类连接器由专门的解决方案提供商提供,使用者通常无需再进行任何开发即可部署使用。

数据空间中的数据连接器遵循国际数据空间标准,是为了确保不同数据空间之间能够实现标准化、安全且可信的数据交换。国际数据空间标准提供了一个共同的框架,允许不同组织和行业在保持数据主权的同时,安全地交换和链接数据。通过遵循国际数据空间标准,数据连接器可以实现以下目标:

数据主权: 确保数据提供者能够完全自主地控制其数据的使用方式,包括决定谁可以访问数据以及数据的使用条件。

信任: 建立一个可信的商业生态系统,其中所有参与者和软件在获得访问权限之前都经过认证,确保了数据交换的可靠性。

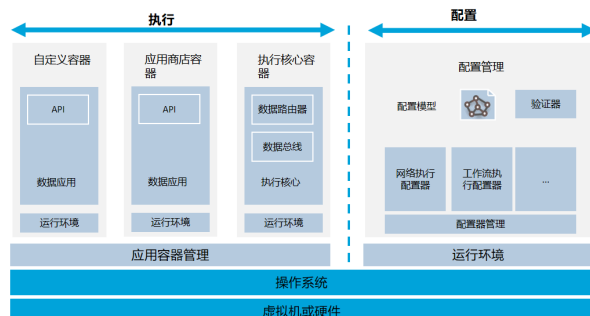
安全性: 采用最先进的安全措施来保护数据交换过程,并作为认证标准的一部分,确保数据的安全。

标准化互操作性: 确保不同的数据空间能够使用统一的规则 and 标准进行通信和数据交换,从而实现不同系统间的互操作性。

生态系统构建: 支持构建数据生态系统,使数据能够跨企业、跨行业进行应用,促进新的商业模式和服务的创新。

认证: IDS 标准包括了对参与者和组件的认证过程,确保了数据空间的安全性和信任度。

2 数据连接器架构

图2 数据连接器架构^[3]

数据连接器架构使用了应用容器管理技术,以确保各个数据服务有一个隔离且安全的环境。数据服务是指提供API来存储、访问或处理数据的系统。为了确保敏感数据

的隐私, 数据处理应该尽可能靠近数据源进行。任何数据预处理(例如, 过滤、匿名化或分析)应该由内部数据连接器执行。只有打算向其他参与者提供的数据, 才应该通过外部数据连接器公开。

数据连接器的内部结构如上图, 可以按需修改或添加可选组件。整个分为两个阶段: 执行和配置。

数据连接器的执行阶段涉及以下组件:

(1) 应用容器管理: 在大多数情况下, 执行核心容器和选定的数据服务的部署是基于应用容器的。为了防止意外的相互依赖, 数据服务通过容器相互隔离。使用应用容器管理, 可以加强对数据服务和容器的控制。在开发过程中, 以及在资源有限的系统中, 可以省略应用容器管理。容器部署的困难可以通过特殊的执行配置器来处理。

(2) 执行核心容器: 提供与数据服务交互的组件, 并支持通信(例如, 数据路由器或数据总线到连接器)。

数据路由器处理与数据服务的通信, 根据预定义的配置参数调用数据服务。在这方面, 它负责数据服务和数据总线之间的数据传输。参与者可以选择用不同供应商的替代实现替换数据路由器组件。配置差异可以通过专门的执行配置器插件来处理。如果嵌入式数据连接器有单一数据服务或固定连接配置组成, 数据路由器可以被硬编码软件替代, 或者数据服务可以直接使用。数据路由器可以调用执行使用策略的相关组件。

数据总线负责数据服务和其他数据连接器的数据总线组件交换数据。它也可以在数据连接器内存储数据。通常, 数据总线提供数据连接器之间的交换数据。

(3) 应用商店容器是从应用商店下载的经过认证的容器, 为连接器提供特定的数据服务。

(4) 自定义容器提供自行开发的数据服务。自定义容器通常不需要认证。

(5) 数据服务定义了一个公共 API, 由数据路由器调用。这个 API 在元数据描述中正式指定, 并导入到配置模型中。数据服务要执行的任务可能各不相同, 针对不同的运行时环境, 数据服务可以用任何编程语言实现。可以重用现有组件, 以简化从其他集成平台的迁移。

(6) 数据服务的运行环境取决于选定的技术和编程语言。运行环境和数据服务一起构成了容器的主体部分。不同的容器可能使用不同的运行环境。

数据连接器的配置阶段涉及以下组件:

(1) 配置管理器构成了连接器的行政部分。其主要任务是管理和验证配置模型, 其次是部署连接器。部署由配

置器管理委托给一组执行配置器。

(2) 配置模型是一个可扩展的领域模型, 用于描述连接器的配置。它由技术无关的、相互连接的配置方面组成。

(3) 配置器管理加载和管理一组可交换的执行配置器。当连接器部署时, 配置器管理将每项任务委托给特殊的执行配置器。

(4) 执行配置器是可交换的插件, 执行或将配置模型的单个方面翻译成特定技术。执行配置的过程取决于所使用的技术。常见的例子包括生成配置文件或使用配置 API。使用不同的执行配置器, 可以采用新技术或将它们集成到连接器中。因此, 每一种技术(操作系统、应用容器管理等)都有自己的执行配置器。

(5) 验证器检查配置模型是否符合自定义规则和 IDS 规定的一般规则。违反规则可以视为警告或错误。如果出现这样的警告或错误, 部署可能会失败或被拒绝。

由于配置阶段和执行阶段是分开的, 因此可以独立开发和随后独立运行这些组件。不同的数据连接器实现可能使用各种通信和加密技术, 具体取决于给定的要求。

3 数据连接器的交互

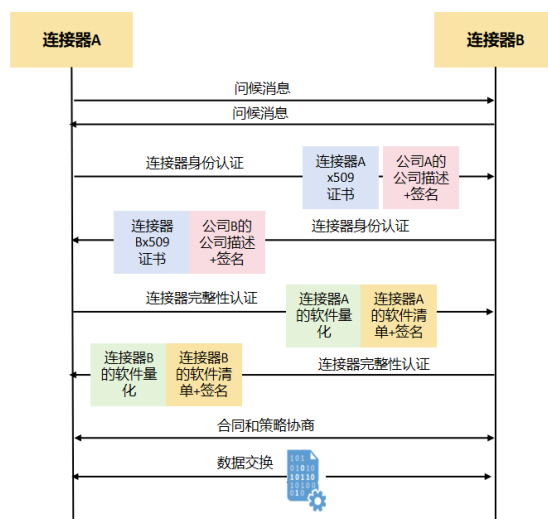


图3 数据连接器交互时序图^[4]

上图显示了两个数据连接器之间从连接建立到第一次数据交换的简化时序图。包含以下步骤:

(1) 问候消息(包括随机数): 在第一步中, 发送一个包含随机数的问候消息。这个数字用于确保后续消息的新鲜感, 从而防止重放攻击。

(2) 连接器身份认证: 在下一步中, 连接器使用它们的 X.509 证书和公司描述来识别自己。它们的身份证书的私钥用于建立通信通道。连接器特定的 X.509 证书的签名和公司描述的签名都可以验证到根 CA。这一步之后, 就

知道了身份和操作环境认证级别。

(3) 连接器完整性认证: 在这一步中, 验证两个连接器的完整性。每个连接器实现了一个证明者来证明其完整性, 并有多验证者来验证各种连接器的完整性。证明者发送所有签名的清单, 描述其连接器的 TCB 和证明者特定的系统状态描述。这通常包括一个从信任根开始的哈希链, 后面跟着代表安全启动链中所有组件的测量哈希值。然后, 证明者以安全的方式用其私钥签名此信息。验证者随后验证签名、新鲜度和软件工件的哈希链。之后, 对于每个工件, 包括开发者、评估设施和认证机构的签名的清单将被验证到根 CA。这一步之后, 整个 TCB 的可信度和组件认证级别就已知了。

(4) 合同和策略协商: 在建立了安全和可信的通信通道的基础上, 两个通信伙伴可以协商数据交换的条款和所需的使用控制策略。

(5) 数据交换: 在最后一步中, 可以交换数据。

4 数据连接器的应用

根据 IDSA 最新的统计^[5], 共有 38 个数据连接器, 在开源方面, 有 17 个数据连接器是开源的, 3 个部分开源; 在开发流派方面, 11 个数据连接器是基于 Eclipse Dataspace Components 数据连接器的变体, 9 个是基于 Dataspace 数据连接器的变体; 在成熟度方面, 有 5 个数据连接器获得 IDS-Ready 认证; 在身份管理方面, 17 个数据连接器支持 X.509, 7 个支持 DiD:web, 6 个支持 SSI; 在部署方面, 多数数据连接器支持边缘计算、支持云部署; 在通信协议方面, 多个数据连接器支持 HTTPS 作为通信协议, 部分支持 IDSCP; 在用户界面方面, 多个数据连接器可以提供图形用户界面。

数据连接器领域的发展趋势显示出明显的多样性和成熟度, 其中开源解决方案占据主导地位, 反映出行业对协作和共享的重视。随着技术标准的逐步统一和国际认证的

推进, 市场正朝着更高层次的互操作性和信任度发展。身份管理和安全成为设计的核心, 而云部署、边缘计算和混合云的灵活性则满足了不同业务场景的需求。此外, 随着数据空间协议的推出和 IDS 认证的实施, 数据连接器正逐步成为推动数据经济和数字生态系统发展的关键技术。

5 结语

展望数据空间的未来, 其发展蓝图将围绕构建一个以数据为核心的高度互联互通的生态系统。在这一生态系统中, 数据连接器将成为实现数据流动性和可访问性的关键组件, 它们不仅促进了数据的即时交换, 而且在确保数据传输的安全性、隐私保护和合规性方面发挥着至关重要的作用。随着技术的不断演进和国际标准的逐步统一, 数据连接器预计将显著降低数据交换的成本, 同时提升数据处理的效率和透明度。此外, 数据空间的基础设施建设将遵循严格的技术规范和认证流程, 确保数据主权得到尊重, 从而为数据经济的繁荣和创新提供坚实的基础。随着数据空间协议的广泛采纳和 IDS 认证的实施, 数据连接器将在全球范围内实现更高层次的互操作性, 为打造一个开放、协作、高效的数据生态系统提供支持。

参考文献:

- [1] LUO C R, MA Y, JING X, et al. Internet of data: a solution for dataspace infrastructure and its technical challenges[J]. Big Data Research, 2023, 9(2): 110-121.
- [2] A Survey of Dataspace Connector Implementations.
- [3] IDS-Reference-Architecture-Model-3.0-2019.
- [4] Designing Data Spaces 978-3-030-93975-5.
- [5] IDSA-Data-Connector-Report-84-No-16-September-2024-1.

作者简介: 殷致云 (1979.12-), 男, 汉族, 广东开平人, 博士学历, 研究方向: 大数据、数据治理、数据要素。