# 第1章

# 绪　论

## 1.1　数据管理问题与需求

数据已成为企业中最宝贵的资产之一，也是企业数字化和智能化转型的一个关键元素。但是，如何使数据的价值变得清晰，数据管理仍然是一个挑战。企业不断储存数据，对于大多数企业来说，数据都是跨孤岛收集的，并且通常以几乎无法共享使用的方式存储。

在企业寻求利用其数据价值的过程中，会面临来自不同数据源、类型、结构、环境和平台的挑战。当企业采用混合和多云架构时，这种多维数据困境会变得更加复杂。对于许多企业而言，运营数据在很大程度上仍然是孤立的和隐蔽的，企业收集到的大量数据可能包括客户反馈、机器设备的传感器数据及应用程序的日志文件，通常是非结构化的、不完整的，或以不易搜索或访问的格式存储，并未被用于任何有意义的洞察或商业价值的分析。尽管企业的数据具有潜在价值，但许多企业缺乏资源或专业知识来分析这些数据，因为数据是不可见且未被开发的，仍然未被探索和加以开发利用。

随着数据源的多样化，集成和整合数据变得越来越重要，企业需要将来自事务数据存储、数据仓库、数据湖、机器日志、非结构化数据源、应用程序存储、社交媒体存储和云存储的数据汇集在一起，但这些数据通常保存在离散的孤岛中，特别是随着云存储和物联网（IoT）设备的增加，数据孤岛问题

与日俱增。

任何希望利用企业内外所有数据的人都面临着在现在数据世界更加严峻的数据挑战：首先是来自多种结构化和非结构化来源的数据比以往任何时候都多，原始形式的数据在数据质量方面变化很大，有时是成型的和干净的，有时是稀疏和不均匀的；其次是数据有许多不同（且不兼容）的格式，通常是稀疏的，缺少值，需要不同级别的、按需深入到详细信息的整合；再次就是必须支持强大的查询，以便根据需要将数据提供给应用程序，同时必须能够根据需要经常近乎实时地刷新；最后，许多不同的非结构化和灵活结构化数据中，除了带来概念上的复杂性之外，还带来了实体类型和实体之间关系的数量激增。面对纷繁复杂的多源异构大数据，传统的商业智能应用一般包含如图 1.1 所示的数据集成和建模过程，在这个过程中，以下四个技术挑战和数据问题特别突出：

（1）数据可访问问题，即访问所有这些数据，并将其转换为可以管理和集成的格式或表示；

（2）数据的结构、整合和转型问题，即为数据带来结构，并将其集成和转换为新的形式的问题；

（3）整合视图，即跟踪更广泛的存储库及为使数据有用而创建的所有模型和转换问题；

（4）灵活实施问题，即在最适合应用程序的平台上集成和转换数据，无论是在本地、一个云上还是在多个云上。


图 1.1　商业智能应用实现中使用的集成和建模过程

现代数据管理需要新兴设计理念来应对一直存在的数据管理挑战，例如，高成本和低价值的数据集成周期、早期集成的频繁维护、对实时性不断增长的需求和事件驱动的数据共享等。在日益多样化、分布式和复杂的环境中，数据管理敏捷性已成为企业的任务关键优先事项。为了减少人为错

误和总体成本，数据和分析需要超越传统的数据管理实践，转向现代解决方案，例如，支持人工智能的数据集成。

当我们想到数据管理的新方法如何改变数据环境时，现代数据管理新范式提出了以下需求主题。

**1. 业务概念方面的规范和统一描述**

数据建模师可以对域中的数据资源进行规范建模和统一描述，并为它们提供与业务中熟悉的概念相对应的名称和结构。将业务概念和流程与数据结构相匹配，为业务分析提供一个正式的数据结构来表达业务模型、数据模型及它们之间的联系。

（1）任何格式的任何数据：必须集成来自所有源平台的数据，无论源的结构变化、格式差异、各自的数据模型或原始时的其他区别如何。这包括常见的结构化源（如关系数据库和平面文件）、半结构化数据（如 JSON 或 XML 文件）和非结构化源（如 PDF 或文档）。

（2）高性能加载和高效存储：自动化和快速加载源数据，并提供高效存储数据的选项，提供有关如何将源数据连接到存储的选项，从完整的数据复制及载入到按需数据和虚拟化。

（3）大规模交互式查询：非常快速地生成大量分析就绪的混合数据集，以满足整个企业数据使用者的独特和紧急需求。快速响应时间是关键，处理已知问题查询及意外探索性查询的能力也是关键需求之一。

**2. 面对复杂的数据时灵活的显式知识表示**

显式知识表示的优点之一是很容易扩展模型以适应新的概念和资产，同时提供元数据灵活性。

（1）通过可靠、快速地将数据输送到存储来缩短洞察时间并做出更明智的决策；

（2）获得实时、360°视角，即任何业务实体（如客户、索赔、订单、设备或零售店）的 360°视图，以实现微细分、减少客户流失、提醒运营风险或提供个性化的客户服务；

（3）将总拥有成本降低到通过增量和快速的方式对遗留系统进行现代

化操作、扩展、维护和更改。

显式知识表示允许用户在提出问题时及在未预料到的后续问题时理解这些关系。当需求急剧变化时，语义模型还允许动态重塑数据，并快速添加新的数据源，以支持特定受众的新类型问题、分析汇总或上下文简化。

**3. 以数据为中心（而不是以应用程序为中心）**

大多数数据表示（尤其是关系数据库）都将数据封装在某种应用程序中，从一个平台到另一个平台，没有标准的方式来交换数据和大规模描述数据的模型。提取、转换和加载（ETL）项目既昂贵又脆弱。

（1）数据准备自动化，使数据科学家、数据工程师和其他 IT 资源免于执行繁琐的重复数据转换、清理和丰富任务；

（2）获得访问任何数据交付方法中的企业数据——包括批量数据移动（ETL）、数据虚拟化、数据流、更改数据捕获和数据交付 API；

（3）集成并增强了公司当前使用的数据管理工具，以提高成本效益。

**4. 数据即产品（包括服务级别协议 SLA、用户满意度等）**

提供数据是企业的一个部门支持另一个部门的一种服务方式，这种强调数据的转变有时被称为数据即产品（data as product）。提供数据的人承担的责任与我们期望的任何其他提供产品的人相同，包括担保、文档、服务协议、对客户请求的响应等。当人们将数据视为产品时，企业的其他部分不太可能希望接管对其版本数据的维护。

（1）数据工程师和数据消费者之间共享的通用语言改善数据和数据之间的协作业务团队；

（2）自助数据服务访问功能使数据消费者可以随时随地获取所需数据，从而提高业务敏捷性和速度。

**5. 处理意外问题的能力**

静态数据表示的一个经常令人遗憾的缺点是，虽然构建的数据结构对回答特定问题的过程是很好理解的，但重用这种结构来回答新问题的过程很困难，通常相当于重新开始。增量建模工作不会带来增量收益。

（1）企业级安全性和治理：就绪可用数据和知识的很大一部分与安全

性、数据治理和法规遵从性的基础知识有关。需要精细的访问控制来指定哪些用户可以访问、查询和更新的哪些部分及他们如何使用授权的数据。这种精细的安全性具有重大的数据治理影响，加强了基于角色的数据隐私访问控制，同时其对个人身份识别(personal identification information，PII)的适用性非常适合法规遵从性。

(2) 灵活部署：重要的是，可以选择部署到任何位置，包括本地、云或混合模型。最经济实惠的操作环境通常不需要额外的投资，例如，云部署选项包括本地混合、公共或私有选项或公私混合。

(3) 轻松与结构的其他组件集成：一个包含许多组件的架构，其中一些组件对企业来说是新的，另一些则是长期存在的。为了在这种背景下工作，需要支持开放标准，使构成的所有数据、模型和元数据都能够轻松地与其他应用程序同步或导出到其他应用程序。

(4) 所有新系统的开发都必须要考虑数据隐私。企业需要能够全面了解自己的所有数据，还需要通过一定方式，通过单点对整个基础架构实施安全控制。数据虚拟化技术提供了这种能力，让企业能够快速、方便地满足数据保护法规的要求，同时又不必投资于新的硬件，也不必从零开始重建现有系统。

**6. 可查找、可访问、可互操作和可重用(FAIR)**

可查找、可访问、可互操作和可重用(F：findability，A：accessibility，I：interoperability，R：repeatation)，简称 FAIR 原则，对数据和元数据表示提出了各种要求。显式知识表示可以找到适合特定任务的数据。全局可引用术语允许互操作性，因为一个数据或元数据集可以引用任何其他数据或元数据。

1) 可查找(findability)

FAIR 原则的首要原则是 F(findability)原则，即数据的可查找性。如果无法识别和查找数据，则无从谈论数据的访问、互操作和重用。数据要符合 findability 原则需满足四个子原则，以下分别用 F1、F2、F3、F4 表示。

F1：(元)数据被分配有一个全球唯一且持久的标识符。F1 原则是所有原则的基础。如果没有一个全球唯一且持久的标识符，FAIR 的其他方面便

很难实现。全球唯一且持久的标识符消除了数据的歧义。许多数据存储库自动为已存储的数据生成全球唯一且持久的标识符。标识符可以帮助人们准确理解数据的意思，帮助计算机以一种有意义的方式解释数据。标识符对人机交互至关重要，而人机交互正是开放科学的前景所在。标识符可以帮助他人在重用数据时正确引用该数据。标识符需满足两个特征：①全球唯一。人们可以通过注册表服务获得数据的全球唯一标识符，该注册表服务使用的算法可以保证标识符的唯一性。不存在有两个不同的数据拥有同样的标识符。②持久存在。标识符对应的网络链接应一直存在。维护网络链接需要成本，随着时间的推移，很多网络链接往往会失效。而人们通过注册表服务获得的标识符可以（在某种程度上）保证网络链接在未来一直存在。

目前对标识符来说最大的挑战是确保它的寿命，尤其是确保由不同项目或社区创建的标识符在该项目结束或者社区结束后仍能存在。因此需要保证标识符与这些项目或社区相独立。

F2：数据使用了丰富的元数据进行描述。描述数据的元数据应当非常丰富，应当包括数据的背景、质量、状况或特征等情况。丰富的元数据可以让计算机自动完成日常且繁琐的分类和排序任务，这些任务目前耗费了研究人员大量的精力。F2 原则背后的基本原理是，即使没有数据标识符，人们也应该能够根据元数据提供的信息找到数据。遵守 F2 原则能够帮助人们定位数据，并增加该数据的重用和引用。

F3：元数据清晰且明示地包括了它们所描述数据的标识符。元数据和它们描述的数据集通常处于不同的文件夹中，元数据文件和数据集文件之间通过在元数据中提到数据集的全球唯一且恒久标识符相联系。F2 要求数据使用元数据进行描述，F3 表明元数据除了包含用以描述数据的元数据，还应包含被描述数据的标识符，用以确定数据的位置。

F4：（元）数据已在可检索的资源中注册或者建立了索引。标识符和丰富的元数据并不能确保数据在互联网上“可查找”。如果数据不可查找，那么再完美的数据也将失去价值。使得数据资源可查找的方法很多，比如建立索引。百度通过爬虫“读取”网页并自动将它们建立索引，便可以让人们

通过百度搜索查找到网页。对于大多数普通搜索者而言，百度搜索已是足够，但对于学术研究数据的检索，人们仍需要建立更明确的索引。F1～F3原则为这类索引的建立提供了核心要素。

2）可访问(accessibility)

FAIR原则中的第二个原则为A(accessibility)原则，即数据的可访问性。用户在查找到所需的数据后的下一步需访问该数据，访问可能要进行身份验证并获得授权。数据要符合accessibility原则也需满足四个子原则，以下分别用A1、A2、A3、A4表示。

A1：(元)数据可通过标识符使用标准化的通信协议进行检索。A1原则指出，FAIR数据的检索不需要专门或专有的工具或通信方法，使用标准化的通信协议即可。标准化的通信协议有TCP、HTTPs、HTTP等。大多数网络用户通过点击链接来检索数据。链接是一个名为TCP协议的高级接口，计算机执行该协议进而在用户的Web浏览器中加载数据。HTTPs、HTTP则是构成现代互联网主干的协议，它们建立在TCP协议基础之上，但请求和提供数字资源比其他通信协议更容易。

A1.1：协议开放、免费、普遍可实现。为最大限度地实现数据重用，FAIR数据使用的通信协议应当免费、开放、可在全球范围内实现。任何人只要有一台电脑与互联网连接，就至少可以访问元数据。这一原则将影响人们对共享数据的存储库的选择。

A1.2：协议在必要时允许认证和授权程序。A1.2原则是FAIR原则中关键但经常被误解的一个原则。FAIR原则中的“A”并不必然意味着“开放”或“自由”。即使受到严格保护的私有数据也可以是符合FAIR原则的。“A”意味着应当提供数据可访问的确切要求。理想状况下，机器可以自动理解访问数据的要求然后自动执行该要求或提醒用户注意该要求。有些数据存储库要求用户在存储库中创建用户账户，这可以让存储库得以验证每个数据集的所有者(或贡献者)的身份，并可以根据用户的不同创设不同的用户权利。A1.2原则也将影响人们对共享数据存储库的选择。

A2：即使数据不再可用，元数据仍然可以被访问。维护数据资源的在线需要成本，随着时间的推移，网上的数据常常会减损，链接会失效。而存

储元数据往往比存储数据更方便、成本更低。因此，A2 原则要求保证元数据持续存在，即使数据本身不再存在。A2 原则与 F4 原则中描述的注册和索引问题有关。

3）可互操作(interoperability)

数据通常需要与其他数据进行集成。此外，数据还需要与应用程序或工作流进行互操作，以进行分析、存储和处理。数据的互操作指通过结合相互独立的数据以获得整体的分析结果。数据要符合 interoperability 原则需满足三个子原则，以下分别用 I1、I2、I3 表示。

I1：(元)数据使用一种正式、可访问、共享和广泛适用的语言来表示知识。正如人类之间需要能够交换和理解彼此的信息，计算机之间也需要能够互相交换和理解彼此的数据。因此数据应当是机器可读的，并且不需要借用专门或特别的算法、翻译器或映射来进行数据的转换。每个计算机至少需要了解其他计算机的数据交换格式。为实现这一点，以及为确保数据的自动可查找和互操作，需要：①使用常见、受控的词汇、本体和主题词表(具有可解析的全球唯一且恒久标识符)；②使用良好的数据模型。

I2：(元)数据使用的词汇表符合 FAIR 原则。用于描述数据集的受控词汇表需适用全球唯一且恒久标识符进行记录和解析，并且能够轻松地被任何使用该数据集的人查找和访问。

I3：(元)数据包括对其他(元)数据的限定引用。限定引用是一个解释了其意图的交叉引用。例如，$X$ 是 $Y$ 的监管者是比 $X$ 与 $Y$ 有关系或者 $X$ 也能看到 $Y$ 更恰当的引用。限定引用可以在元数据之间创建有意义的链接，丰富人们对数据背景的了解，可以让人们明确一个数据集是否建立在另一个数据集之上，是否需要额外的数据集来完成目前的数据集，或者互补信息是否存储在不同的数据集中。I3 原则需要注意两点：①根本上而言，实现数据的互操作性不是为了连接不同的数据，而是为了实现数据用户的互操作；②为实现数据的互操作，描述它的元数据也应当可以互操作。

4）可重用(reuse)

FAIR 原则的最终目的是实现数据的可重用。数据要符合可重用(reuse)原则需满足两个子原则，以下分别用 R1、R2 表示。

R1：(元)数据被多个准确且相关的属性所描述。添加了很多标签的数据将更易被发现和重用。R1 原则与 F2 原则相关，但 R1 关注的是用户(机器或人)判断数据在特定场景中是否真的有用的能力。数据发布者不仅应提供让数据能被发现的元数据，还应提供丰富的描述数据生成场景的元数据，如实验协议、生成数据的机器或传感器的制造商和品牌等。数据发布者不应试图预测数据消费者的身份和需求，而是应当尽可能多地提供元数据，即使提供的元数据看起来与数据不甚相关。

R1.1：(元)数据在发布时需提供清晰且可访问的数据使用许可。许可中应当清晰地描述数据使用的范围。重用数据的企业都在努力遵循数据使用的种种限制和规范，如果数据使用的范围描述不清，将会严重限制数据的重用。而随着涉及更多许可考虑的自动搜索技术的发展，许可状态的明确将变得更加重要。因此必须让机器和人都清楚数据可以使用的条件。上文提到的 I 原则描述的是数据在技术上的可互操作性，R1.1 则是关于数据在法律上的互操作性。

R1.2：(元)数据有详细的来源。重用数据的人应当清楚数据来自哪里，需要如何引用或作者希望如何被承认。数据应当包括生产它的完整工作流：谁生成或采集了这些数据、它们是如何处理的、它们以前是否发布过、它们是否包含其他人的数据。理想情况下，这个数据处理工作流应当是机器可读的。

R1.3：(元)数据符合相关领域的社区标准。如果数据集相似，它们将更容易重用。例如，相同类型的数据、以标准化方式组织的数据、完善和可持续的文件格式、遵循通用模板且使用通用词汇表的文档(元数据)。如果存在数据归档和共享的领域标准或最佳实践，则应该遵循这些标准或实践。例如，许多社区都有最低限度的信息标准(例如，MIAME、MIAPE)。FAIR 数据至少应符合这些标准。有些情况下，提交者提交的数据可能会偏离这一类型数据的标准，这时他们都会提供有效且明确的理由。FAIR 原则并不解决数据的可靠性问题。数据的可靠性取决于使用者，并且与数据的应用目的有关。

综上，当前在数据集成和从数据孤岛中生成深刻见解方面，企业面临着

重大挑战。当前数据领域最大的障碍之一是数据碎片化，即数据分布在各种系统和平台上，难以访问、分析和管理。随着混合云和多云环境中数据源数量的不断增加，这种数据孤岛现象日益严峻，企业需要努力集成来自多个异构源的数据，以创建统一的数据视图应对复杂的竞争环境。

## 1.2 数据管理架构综述

目前已经出现了多种技术方法来帮助企业处理相关的数据集成问题，包括数据联邦(data federation)、数据编织(data fabric)、数据网格(data mesh)、湖仓一体(data lakehouse)及传统数据栈的数据中台(data middleware)、数据仓库(data warehouse)、数据湖(data lake)等。

**1. 数据联邦(data federation)**

数据联邦是一种数据集成方法，将来自不同数据源的数据整合在一起，形成一个虚拟的数据集成视图。数据联邦的特点包括：

(1) 灵活性高：数据联邦可以快速地进行数据集成，而且不需要对数据源进行改动，因此可以适应快速变化的数据需求。

(2) 低运维成本：相比传统的数据集成方法，数据联邦的运维成本更低，因为它不需要将数据复制到中央数据存储中。

(3) 实时性好：数据联邦可以实现实时信息访问，因为它不需要将数据复制到中央数据存储中。

(4) 安全性高：数据联邦可以消除数据复制和备份的需要，从而提高数据的安全性。

(5) 跨异构数据源处理：数据联邦可以处理异构数据源，包括结构化、半结构化和非结构化数据。

数据联邦可以包括集成多个数据库系统、云存储、数据仓库及其他数据源，使企业能够更全面地了解其数据，并更容易进行跨数据源的分析和报告。数据联邦有助于解决数据分散、复杂性和多样性带来的挑战，使企业能够更灵活地访问和利用其数据资产。数据联邦的实现需要一个联邦计算引