DAMA-DMBOK

数据管理知识体系指南CDGA/CDGP认证

第8章 数据集成和互操作（完整课程视频请扫描二维码）

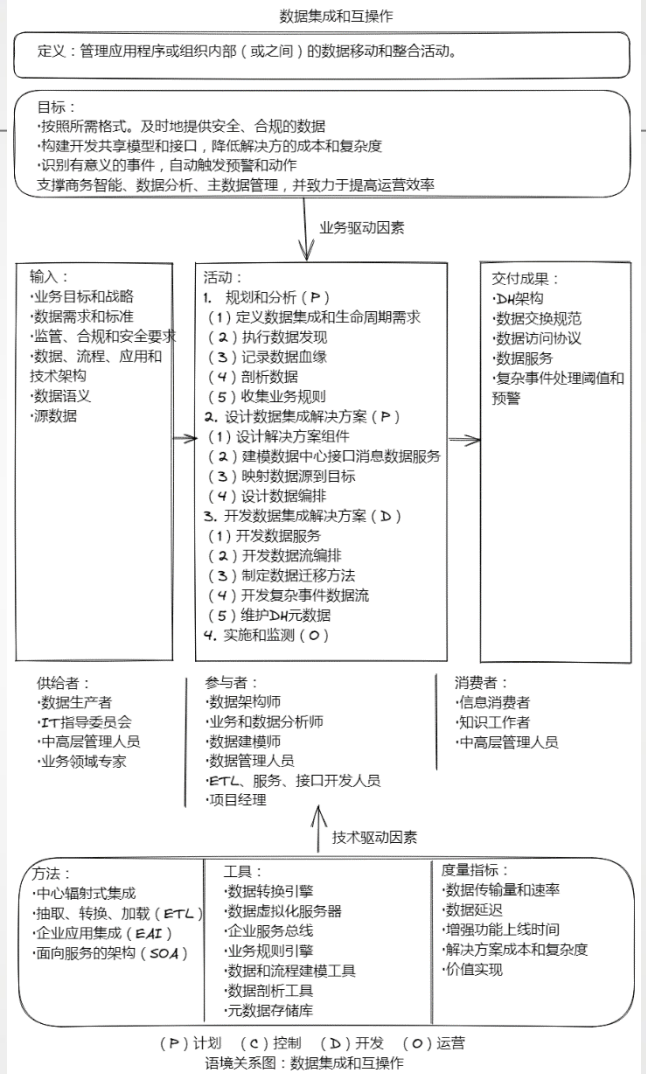
第8章 数据集成和互操作

01

引言

定义、业务驱动因素、目标和原则、基本概念

数据集成和互操作的语境图



语境关系图：数据集成和互操作

定义

数据集成和互操作（DII）描述了数据在不同数据存储、应用程序和组织这三者内部和之间进行移动和整合的相关过程。数据集成是将数据整合成物理的或虚拟的一致格式。数据互操作是多个系统之间进行通信的能力。数据集成和 互操作的解决方案提供了大多数组织所依赖的基本数据管理职能：

1）数据迁移和转换
2）数据整合到数据中心或数据集市
3）将供应商的软件包集成到组织的应用系统框架中
4）在不同应用程序或组织之间数据共享
5）跨数据存储库和数据中心分发数据
6）数据归档
7）数据接口管理
8）获取和接收外部数据
9）结构化和非结构化数据集成
10）提供运营智能化和管理决策支持

定义

数据集成和互操作依赖于数据管理的其他领域，如：

1）数据治理：治理转换规则和消息结构

2）数据架构：用于解决方案设计

3）数据安全：无论是数据持久化、虚拟化还是在应用程序和组织之间流动，都要确保解决方案对数据的安全性进行适当的保护

4）元数据：用于知晓数据的技术清单（持久的、虚拟的和动态的）、数据的业务含义、数据转换的业务规则、数据操作历史和数据血缘

5）数据存储和操作：管理解决方案的物理实例化

6）数据建模和设计：用于设计数据结构，包括数据库中的物理持久化的结构、虚拟的数据结构以及应用程序和组织之间传送的消息结构。

业务驱动因素

主要目的是为了对数据移动进行有效管理，另一个驱动因素是维护管理成本。

管理数据集成的复杂性以及相关成本是建立数据集成架构的原因

目标和原则

目标：

1）及时以数据消费者所需的格式提供数据

2）将数据物理地或虚拟地合并到数据中心

3）通过开发共享模型和接口来降低管理解决方案的成本和复杂度

4）识别有意义的事件（机会和威胁），自动地出发警报并采取相应行动

5）支持商务智能、数据分析、主数据管理以及运营效率的提升

原则：

1）采用企业视角确保未来的可扩展性设计，通过迭代和增量交付实现

2）平衡本地数据需求与企业数据需求，包括支撑与维护

3）确保数据集成和互操作设计和活动的可靠性。业务专家应参与数据转换规则的设计和修改，包括持久性和虚拟性。

1、抽取、转换、加载

（1）抽取

（2）转换：是让选定的数据与目标数据库的结构相兼容

1）格式变化　　2）结构变化　　3）语义转换

4）消除重复　　5）重新排序

（3）加载：加载过程实在目标系统中物理存储或呈现转换结果。

（4）抽取、加载、转换（ELT）

如果目标系统比源系统或中间应用系统具有更强的转换能力，那么数据处理的顺序可以切换为ELT

ETL 和ELT的区别要掌握：数据湖会采用哪种？

（5）映射

是转换的同义词，它既是从源结构到目标结构建立查找矩阵的过程，也是该过程的结果。映射定义了要抽取的源数据与抽取数据的识别规则、要加载的目标与要更新的目标行的识别规则以及要应用的任何转换或计算规则。

2、时延

（1）批处理

（2）变更数据捕获

1）源系统填入特定的数据元素

2）源系统进程在更改数据时被添加到一个简单的对象和标识符列表，然后用于控制抽取数据的选择

3）源系统复制已经变化的数据

（3）准实时和事件驱动

（4）异步：提供数据的系统在继续处理之前不会等待接收系统确认更新。不会阻塞源应用程序继续执行，也不会在任何目标应用程序不可用时导致源应用程序不可用。

（5）实时，同步：执行下一个活动或事务之前需等待接收来自其他应用程序或进程的确认。

（6）低延迟或流处理：低延迟旨在减少事件的响应时间。可能包括使用像固态硬盘的硬件解决方案或使用内存数据库的软件解决方案。

3、复制

监视数据集的更改日志。如果数据更改动作发生在多个副本站点时，那么数据复制解决方案不是最佳的选择。

4、归档

5、企业消息格式/规范格式

规范化的数据模型时组织或数据交换团队使用的通用模型，用于标准化数据共享的格式

6、交互模型

（1）点到点

1）影响处理：如果源系统是操作型的，那么提供数据的工作量可能会影响交易处理。

2）管理接口：点对点交互模型所需的接口数量接近系统数量的平方数。

3）潜在的不一致：当多个系统需要不同的版本或数据格式时，就会出现设计问题。

（2）中心辐射型

企业服务总线（ESB）是用于在多个系统之间接近实时共享数据的数据集成解决方案，其数据中心是一个虚拟概念，代表组织中数据共享的标准和规范格式。

（3）发布与订阅

发布和订阅模型涉及推送（发布）数据的系统和其他接受（订阅）数据的系统。

7、数据集成和互操作架构概念

（1）应用耦合

松耦合是一种优选的接口设计，其中在系统之间传送数据不需要等待响应。基于企业服务总线EBS的面向服务架构是松散耦合数据交互设计模式的一个示例。

（2）编排和流程控制：

数据传送架构中经常被忽略的方面：

1）数据库活动日志

2）批量作业日志

3）警报

4）异常日志

5）作业依赖图，包含补救方案、标准回复

6）作业的时钟信息，如依赖作业的定时、期望的作业长度、计算（可用）的窗口时间

（3）企业应用集成：在企业应用集成模型（EAI）中，软件模块之间仅通过定义良好的接口调用（应用程序编程接口-API）进行交互。数据存储只能通过自己的软件模块更新，其他软件不能直接访问应用程序中的数据，只能通过定义的API访问

7、数据集成和互操作架构概念

（4）企业服务总线：是一个系统，充当系统之间的中介，在它们之间传送消息。应用程序可以通过ESB现有的功能封装发送和接收的消息或文件。

（5）面向服务的架构：SOA，通过在应用程序之间定义良好的服务调用，可以提供推送数据或更新数据的功能

（6）复杂事件处理：是一种跟踪和分析（处理）有关发生事件的信息流（数据流），并从中得出结论的方法。复杂事件（Complex Event Processing，CEP）将多个来源的数据进行合并，通过识别出有意义的事件（如机会或威胁），为这些事件设置规则来指导事件处理及路由，进而预测行为或活动，并根据预测的结果自动触发实时相应，如推荐消费者购买产品。

（7）数据联邦和虚拟化

数据联邦提供访问各个独立数据存储库组合的权限

数据虚拟化使分布式数据库以及多个异构数据存储能够作为单个数据库来访问和查看，

（8）数据即服务

软件即服务SaaS是一种交付和许可模式。数据即服务DAAS的一个定义是从供应商获得许可并按需由供应商提供数据，而不是存储和维护在被许可组织数据中心的数据。

（9）云化集成

云化集成，也称为集成平台即服务或IPaaS，是作为云服务交付的一种系统集成形式。

8、数据交换标准

交换模式定义了任何系统或组织交换数据所需的数据转换结构。数据需要映射到交换规范中。

国家信息交换模型（NIEM）是为美国政府之间交换文件和交易而开发的数据交换标准。使用XML来定义模式和元素的表述。

02

活动

规划、设计、开发、实施

1、规划和分析

1、定义数据集成和生命周期需求
2、执行数据探索：数据探索应该在设计之前进行，目标是为数据集成工作确定潜在的数据来源。数据探索还包括针对数据质量的高级别评估工作，以确定数据是否适合集成计划的目标。
3、记录数据血缘：数据是如何被组织获取或创建的，它在组织中是如何移动和变化以及如何被组织用于分析、决策或事件触发的。详细记录的数据血缘可以包括根据哪些规则改变数据及其改变的频率。
4、剖析数据：数据剖析有助于理解数据内容和结构。基本剖析包括：
    1）数据结构中定义的数据格式和从实际数据中推断出来的格式
    2）数据的数量，包括null值、空或默认数据的级别
    3）数据值以及它们与定义的有效值集合的紧密联系
    4）数据集内部的模式和关系，如相关字段和基数规则
    5）与其他数据集的关系
5、收集业务规则
    1）评估潜在的源数据集和目标数据集的数据
    2）管理组织中的数据流
    3）监控组织中的操作数据
    4）指示何时自动触发事件和警报

2、设计数据集成解决方案

1、设计数据集成解决方案

（1）选择交互模型

（2）设计数据服务或交换模式

包括所涉及数据结构的清单（持久和可传递、现有和必需）、数据流的编排和频率指示、法规、安全问题和补救措施以及有关备份和恢复、可用性和数据存档和保留。

2、建模数据中心、接口、消息、数据服务

持久化的数据结构：主数据管理中心、数据仓库和数据集市、操作型数据存储库

临时数据结构：接口、消息布局、规范模型

3、映射数据源到目标

对于映射关系中的每个属性，映射规范如下：

1）指明源数据和目标数据的技术格式

2）指定源数据和目标数据之间所有中间暂存点所需的转换

3）描述最终或中间目标数据存储区中每个属性的填充方式

4）描述是否需要对数据值进行转换，如通过在表示适当目标值的表中查找源值

5）描述需要进行哪些计算

4、设计数据编排

数据集成解决方案中的数据流必须做好设计和记录。数据流程编排是从开始到结束的数据流模式，包括完成转换和事务所需的所有中间步骤。

3、开发数据集成解决方案

1、开发数据服务

开发服务来获取、转换和交付指定的数据，并且匹配所选的交互模型。

2、开发数据流编排

对集成ETL数据流通常会采用专用工具以特有的方式进行开发。对批量数据流将在一个调度器中开发（如CTRL-M）。互操作性需求可能包括开发数据存储之间的映射或协调点。

3、制定数据迁移方法

4、制定发布方式

5、开发复杂事件处理流

1）准备有关预测模型的个人、组织、产品或市场和迁移前的历史数据

2）处理实时数据流，充分填充预测模型、识别有意义的事件（机会或威胁）

3）根据预测执行触发的动作

6、维护数据集成和互操作的元数据

SOA注册中心提供了一个不断发展变化的受控信息目录：即访问和使用应用程序中数据和功能的可用服务。

实施和监测

应建立表示潜在问题的度量指标以及直接反馈问题的机制，尤其是当触发响应的复杂性和风险增加时，应建立对反馈问题的自动化处理和人工监控流程。

必须采用与最苛刻的目标应用程序或数据使用者相同的服务级别进行监视和服务。

03 工具和方法

工具

1、数据转换引擎/ETL工具：基本考虑应该包括是否需要运用批处理和实时功能，以及是否包括非结构化和结构化数据。目前最成熟的是用于结构化数据的批量处理工具。

2、数据虚拟化服务器：数据虚拟化服务器对数据进行虚拟抽取、转换和集成。数据虚拟化服务器可以将结构化数据和非结构化数据进行合并。数据仓库经常是数据虚拟化服务器的输入，但数据虚拟化服务器不会替代企业信息架构中的数据仓库。

3、企业服务总线

ESB既指软件体系结构模型，又指一种面向消息的中间件，用于在同一组织中的异构数据存储、应用程序和服务器之间实现近乎实时的消息传递。

ESB以异步格式使用，以实现数据的自由流动。

企业服务总线在各个环境中安装适配器或代理软件，在参与消息交换的各个系统上实现数据传入和传出的消息队列。

4、业务规则引擎：业务规则引擎中允许非技术用户管理软件的业务规则，因为业务规则引擎可以在不改变技术代码的情况下支持对预测模型的更改。

5、数据和流程建模工具：不仅用来设计目标结构，而且用来设计数据集成解决方案所需的中间数据结构。

6、数据剖析工具：包括对数据集的内容统计分析，以了解数据的格式、完整性、一致性、有效性和结构。

7、元数据存储库：元数据存储库包含有关组织中数据的信息，包括数据结构、内容以及用于管理数据的业务规则。

基本目标是保持应用程序松散耦合，限制开发和管理接口的数量，使用中心辐射方法并创建标准规范的接口等。

04 实施指南

GRUD、安全补丁部署、数据安全属性、安全要求、加密搜索、文件清理

1、就绪评估/风险评估

基于多个系统之间实现集成的成本合理性

应保持在关注业务目标和需求上，包括确保每个项目中的参与者都有面向业务或应用程序的人员，而不仅仅是数据集成工具专家。

2、组织和文化变革

卓越中心团队，实现共享数据的一致标准

05

数据治理

集成治理和度量指标

1、数据共享协议

开发接口或以电子方式提供数据之前，应制定一份数据共享协议或谅解备忘录（MOU）。协议规定了交换数据的责任和可接受的使用用途，并由相关数据的业务数据主管批准。数据共享协议应指定预期的数据使用和访问、使用的限制以及预期的服务级别，包括所需的系统启动时间和响应时间。

2、数据集成和互操作与数据血缘

治理需要确保记录数据来源和数据移动的信息。数据共享协议可能规定了数据使用的限制。为了遵守这些限制，有必要知道数据在哪里移动和保留。

对数据流进行更改时需要数据血缘信息，必须将此信息作为元数据解决方案的关键部分进行管理。

3、度量指标

要衡量实现数据集成解决方案的规模和收益：包括可用性、数量、速度、成本和使用方面的指标。

1）数据可用性

请求数据的可获得性。

2）数据量和速度。包括：传送和转换的数据量，分析数据量，传送速度，数据更新与可用性之间的时延，事件与触发动作之间的时延，新数据源的可用时间。

3）解决方案成本和复杂度。包括：解决方案开发和管理成本，获取新数据的便利性，解决方案和运营的复杂度，使用数据集成解决方案的系统数量。

本章完结 感谢观看

完整课程视频请扫描二维码咨询