

开放信息模型 SQL 的 XML Schema 设计

The designs of XML Schema of an Opening Information Model SQL

隆益民 陈立仪 (广东女子职业技术学院 广东 广州 511450)

摘要:信息交换平台是一个基于点对点(Peer to Peer)的信息发布系统。在这个系统里,从信息的产生、采集、加工、存储、发布、消费到监管,形成了一个完整的信息生命体系。基于 XML 技术,经过对信息交换平台的信息描述深入的研究,提出了开放信息模型(OIM),对信息进行统一的描述,使信息可以跨平台发布。OIM 研究目标有两个:一个是信息模型的设计(信息模型的设计重点在于对数据库查询语句的 XML Schema 的设计和数据库清洗模型的设计);另一个是信息模型在信息交换平台里的实现。限于篇幅,本文介绍数据库查询语句的 XML Schema 的设计。

关键词:信息交换平台 信息模型 信息描述 数据采集 XML

1 前言

信息模型是描述某个工具、应用程序、数据结构或

与信息交换平台一起使用的信息模型采用可扩展的标记语言(XML)进行描述,所以称它为开放的信息模型

(OIM: Opening Information Model)。开放的信息模型使信息的交换和发布变得更加便利和规范。

在信息交换平台里,信息包含两层的意义,一层是信息的定义(描述),叫元数据,另一层是信息的内容,是数据本身。信息交换平台交换的是信息的定义,而不是信息的实际内容,所以发布数据表的数据记录时,实际上是发布了查询这些数据的一条 SQL 语句。因 SQL 语句本身的复杂性,导致其模型设计的

复杂性。在注重 SQL 语句的功能完整性和模型容易实现的基础上,设计了 SQL 语句的 XML Schema。

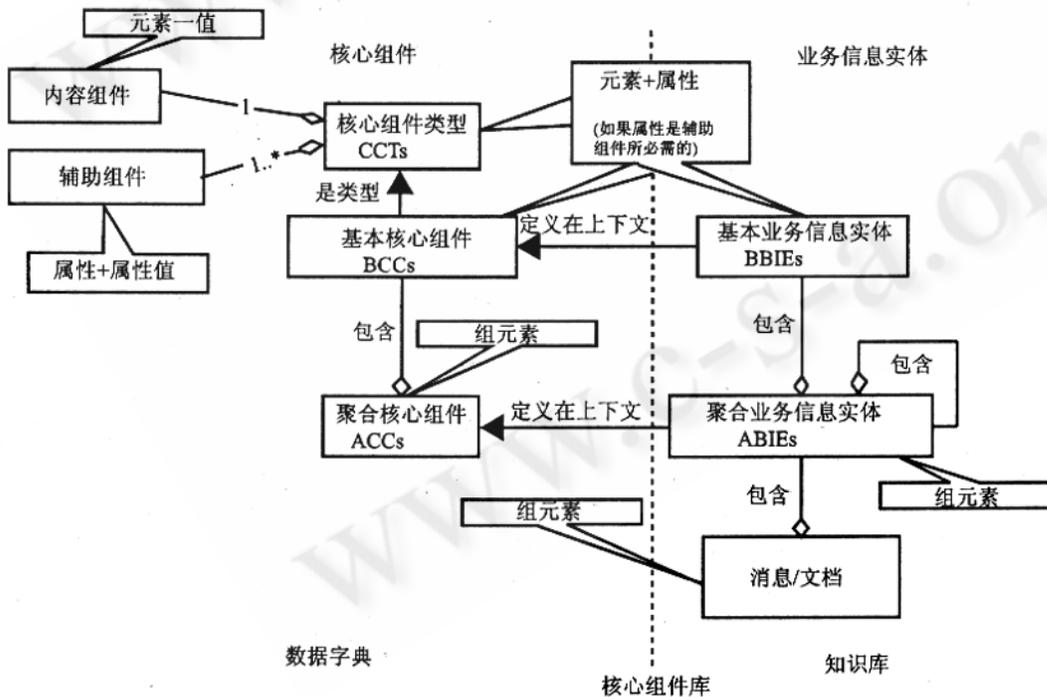


图 1 核心组件库

信息系统的元数据类型集合。在信息交换平台里,信息模型定义存储在计算机系统的信息(文件、文档、数据表、短消息等)里,并由应用程序使用元数据类型。

2 数据库查询语言的 XML Schema 的设计

2.1 核心组件与业务信息实体模型分析

在核心组件 (CCs: Core Components) 技术规范里,核心组件类型用作创建核心组件。核心组件类型由一个内容组件和一个或多个辅助组件构成。核心组件类型用于创建基本核心组件 (BCCs: Basic Core Components) 和聚合核心组件 (ACCs: Aggregate Core Components),分别用于基本业务信息实体 (BBIEs: Basic Business Information Entities) 或聚合业务信息实体 (ABIEs: Aggregate Business Information Entities)。

既然 BBIEs 派生于 BCCs 而且必须用人能读懂的业务术语定义,那么 BCCs 本身必须定义成元素。大多数情况下内容组件的内容是需要拼写检查的。因此,内容组件将被描述为元素 - 值。辅助组件会描述为属性。每一个辅助组件的大多数信息是受约束属性,并由程序使用和描述成无序的。此外,辅助组件可以用枚举类型来包含信息。在一个层里,所有聚合组件 (ACCs 和 ABIEs) 都是节点;并且层里面的节点将被定义成组元素。如图 1 所示,描述了核心组件与业务信息实体之间的关系,并给出 XML 语法的描述。

2.2 核心组件与业务信息实体到 XML Schema 的转换

(1) 在所有组件 (如 BCC, ACC, BBIE 和 ABIE) 里,全局唯一的标识 (GUID: Globle Unique Identifier) 必须用属性作描述。(2) 组件之间的关系应该用属性定义。在 XML 里,关系也可以用 ID - IDREF(S) 属性描述。使用这些属性,一个元素可以指向其它一个或多个元素 (包括那些元素的 ID 值域指向元素自身的 IDREF 或 IDREFS 域)。这与关系数据库的键机制有点相类似,但有一个重要的不同点是:大多数解析器处理这些指针时是单向的。换句话说,给出一个 IDREF 或 IDREFS 域,有可能迅速找到目标元素或用 ID 或 IDS 关联的元素。(3) 如果核心组件或业务信息实体的标记名用另外一种语言描述,应该用属性。(4) 属性应该只用于那些用作定义基本核心组件 (叶子元素) 的核心组件类型的内部。(5) 属性应该只用于定义辅助组件。辅助组件是固定的,定义在 XML 里的核心组件并不一定是不可扩展的。(6) 内容组件应该只定义为叶子元素的内容。(7) 如果定义的辅助组件存在默认值,那么属性并不是必需的。如图 2 所示,有两种类型的属性是必需的。

2.3 SQL 语句的 XML Schema 实现

通过上面的理论分析和实践,实现 SQL 语句的 XML Schema 的设计。

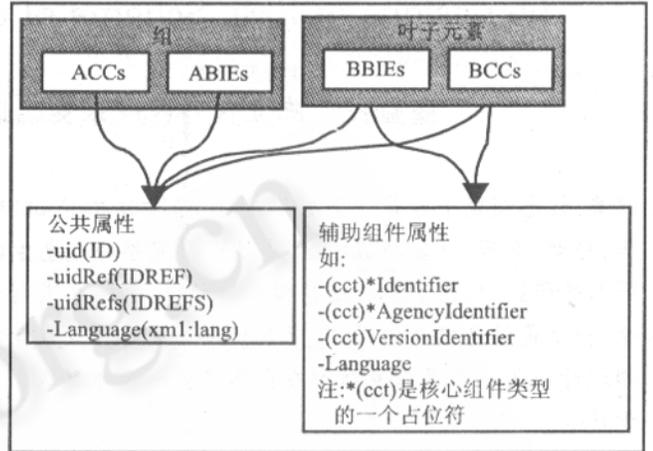


图 2 两种必须的属性类型

2.3.1 根模型组

首先,我们定义出模式的根元素,名称为 SQLSchema,根元素又叫文档元素,是全局性的,而且是唯一的。根元素下有两个子元素,一个叫 SelectStat,一个叫 DBCon。这两个元素也不包含具体信息,分别代表两个模型组。这两个组的基数是 1..∞,也就是说,一个 XML 文档可以存放多条 SQL 语句信息的定义,也就满足了一条复杂的 SQL 语句可以分解成多条简单的 SQL 语句的要求。SelectStat 组有三个子元素: Subquery、OrderBy 和 SuffixInfo。它们的连接是有序的 (Sequence),而且 OrderBy 子元素是用虚线框表示,因它是可选的。根模型组的结构图如图 3 所示。

2.3.2 SelectStat 模型组

(1) 这里 SelectStat 组只介绍 Subquery, OrderBy 和 SuffixInfo 略。Subquery 模型组包含了完整的 SQL 子查询语句的所有子句:有 Select 子句、From 子句、Where 子句、GroupBy 子句和 Having 子句。它们之间是顺序关系 (Sequence),前两项是必须的,后三项可选,如图 4 所示。

(2) Subquery 组只介绍 Select。Select 模型组包含有子元素:SELECT、Distinct 和 FieldFullName 等,之间也是顺序关系。如图 5 所示。

2.3.3 DBCon 模型组

这个组有两个子元素:DBConID 和 DBConString。具体如图 6 所示:

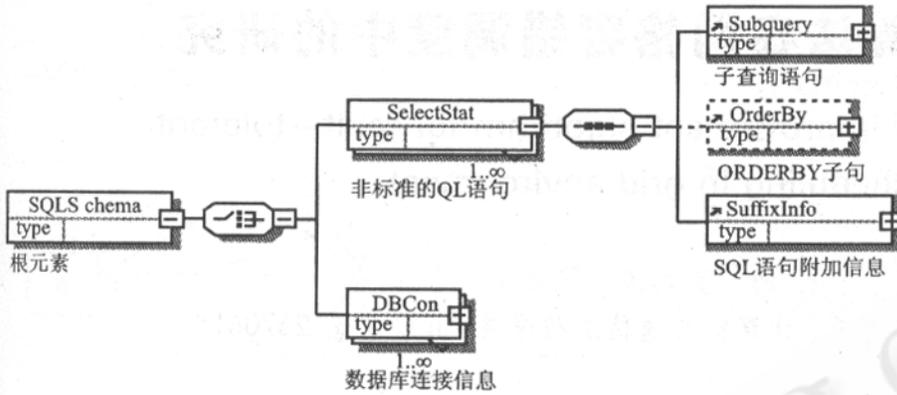


图 3 SQL 语句 XML Schema 主干

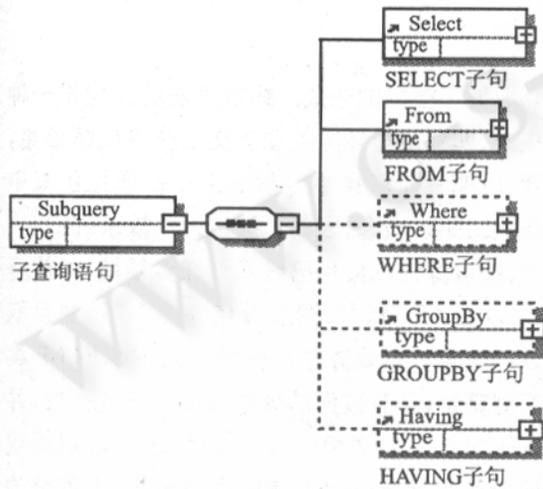


图 4 Subquery 模型组

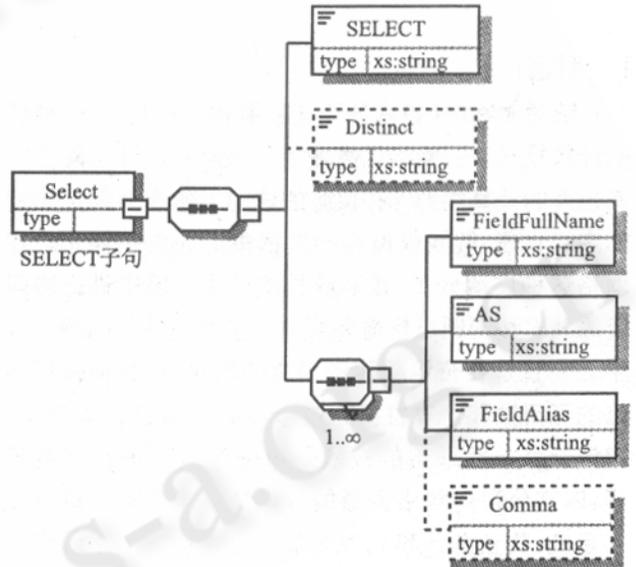


图 5 Select 子句模型图

3 结论

数据库查询语言是一个功能强大的数据处理语言,随着分布式的异构数据库的出现,数据的访问就显得尤为重要,数据是否结构良好,数据的处理效率是否高,直接影响了一个企业的运作和市场竞争能力。随着 XML 的问世,数据格式不统一的混乱现象得到了很大改善。为了更好地实现分布式的数据库访问,用 XML 格式来描述 SQL 语言有极大意义。这里把复杂的 SQL 语句转换成一组简单的 SQL 语句,然后实现 SQL 语句的 XML Schema 的设计。运用这些技术可以使数据交换平台的功能更加强大,适应范围更加广阔。

参考文献

- 1 XML Schema Part 0: Primer W3c Recommendation 2 May 2001. <http://www.w3.org/TR/2001/REC-xm1schema-0-20010502>.
- 2 Patrick O'Neil Elizabeth O'Neil Database Principles, Programming, and Performance (第二版).1998.
- 3 MSxml4.0 SDK Document Microsoft XML Code Services.

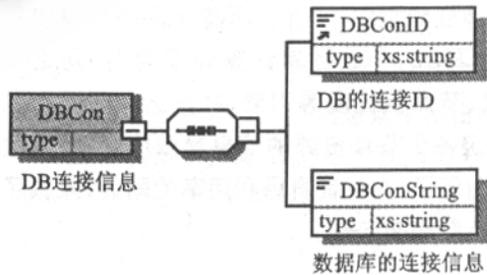


图 6 DBCon 组模型图

- 4 [美]Peter G. Aitken. 微软 XML 技术指南北京,中国电力出版社.2000:105 ~ 155.