**南京信息工程大学 滨江学院 实验（实习）报告**

实验（实习）名称 数据集成与架构建模 实验（实习）日期 2020.10.23 指导教师 王玉祥

专业 软件 年级 2018 班次 2 班 姓名 毛济洲 学号 20182344050 得分

2.案例分析:数据集成方案本题主要考查数据集成中的数据仓库方案和联邦数据库方案的内容。

[问题l]

数据仓库集成是把多种来源的数据集中在一起,建立数据仓库,所有数据都驻留在单个数据库眼务器上,配置大型处理器和存储容量.数据仓库主要用于决策支持,在数据处理过程中强调分析。其特点是:1集成的数据;2面向主题;3数据相对稳定;4包含历史信息。但是此种集成方法需要将数据源的数据转换为数据仓库中的语义,而且需要定期的数据复制和数据更新.数据源往往指的是数据库系统。

[问题2]

Web信息系统中的数据往往是非结构化或半结构化的,同一数据源往往有统一的页面模式,但其数据现是嵌人在页面中的。困此需要采用Web内容提取(挖掘/文本挖掘)的集成办法来获取对应供应商的产品信息。

基本步骤如下。

(1)分析页面,确定页面中的数据模式。制定数据挖掘的特征和提取规则,编写特定的页面分析和数据提取程序(2)抓取页面,通过爬虫技术获取对应的网页。将远端的页面下载到本地存储,为后续数据提取作准备。

(3)特征提取与处理,获取相应数据,依据的原则是分析页面阶段所形成的特征和提取规则。

(4)数据清洗,根据规则进行判断,抛弃异常数据。

(5)数据转换,根据预先定义好的语义映射关系,将数据转换为统一格式。

Web内容提取的算法程序本身是近似的,因此在获取数据的内容上存在不确定性。

[问题3]

(1)数据源与中介模式之间的语义映射可能是近似的实际系统中往往很难有的非常精确、完整的语义映射。

(2)用户不熟悉模式或系统的域太宽,不能提供基于表单式的査洵接口,需要使用关键字查询,但将关键字查询转换成一组候选的结构化查询会带来不确定性(3)数据常常是使用信息获取技术,从非结构化数据源获取的,而这些获取技术一般只是“大致可用”,所取得的数据可能是不确定的因此,在本试题中,无论采用数据仓库或联邦数据库集成方案,都存在数据源之间的语义映射和转换问题,都会带来数据集成的不确定性。

3.

[问题1] 集中式数据架构是由一个处理器、与它相关联的数据存储设备及其他外围设备组成, 被物理地定义到单个位置。在该架构中,一个或多个局域网中的客户共享一个单独计算机系统中的单个数据库。系统提供数据处理能力,用户可以在同样的站点上操作,也可以在地理位置陽开的其他站点上通过远程终端操作。系统及其数据管理披某个站点或中心站点集中控制。单个数据库服务器结构的主要优点就是简单1易维护开发及运行成本低; 但由于所有客户直接请求服务器,容易发生性能瓶硕,如果服务失败,单个服务器不能提供备份和恢复,所有依赖的应用程序都将不能工作.

分布式数据架构由多个计算机系统上的多个局部数据库系统构成。在该架构中,用户能够访问远程系统数据,数振可以在多个不同的数据库中传送,由不同的数据库管理系统软件管理,运行在多种不同的计算机上,支持多种不同的操作系统。这些机器位〒(或分布在)不同的地理位置,并通过多种通信网络连接在一起.企业数据可以分布在不同的计算机上,一个应用程序可以操作位于不同地理位置的机器上的数据。多个数据库服务器结构的主要优点就是系统的容错能力和对广域网容量的需求有所降低,可以采用多种策略提升整个系统的服务质量;由于多个数据库系统分布在不同的网络节点上,位于不同位置的数据间需要同步和协作,系统结构复杂、运行成本高且维护困难.

在实际应用系统的数据架构建模中,应根据不同的需求选择集中式或分布式数据架构。本题中由于RMO要扩展其销售区域,潜在客户数量也会因此大幅度增加,所以良好的可扩展性是CRSS系统所必需的质量属性;并且由于销售区域扩大,系统中的数据会存储于不同的地理位置,所以采用分布式数据架构最合理。

[问题2] 读写分离架构利用了数据库的复制技术,将数据的读和写分布在不同的处理节点上, 从而达到提高'q了用性和扩展性的目的。读写分离架构的应用非常广泛,很多网站采用数据库+缓存的方式实现。通过缓存层承载大量的读访问,如广泛采用的Mencached,自身往往不具备持久层存储的功能,通常和数据库一起组成分布式数据架构,由数据库负责数据持久化存储和写人功能,缓存负责承载大量的并发访问,从而提高系统的数据处理效率。要避免数据访问的单点故障,通常采用主数据库热备份的方式来实现。所以,要实现题目要求的CRSS分布式数裾架构,需要多个局部数据库系统、多个热备份数据库系统利多个数据缓存组成。局部数据库负责数据的写人,多个热备份数裾库系统解决单点故障的问题,数据缓存负贵为应用提供所读取的数据。

读写分离结构中,应用读取数据时访问缓存,如果没有命中所需数据,则从主数据库中读取数据并写人缓存;对于新增、修改和删除操作,需要采用延迟加载的策略,新增时只修改主数据库,修改和删除时,除了修改主数据库中的内容,还需要将缓存中的数据标记为失效.

(1)读取数据:应用访问缓存,如果命中则返回,否则从局部数据库系统中读取数据,并将数裾加载到缓存后返回。

(2)添加数据:采用延迟加载策略,应用将数据直接写人局部数据库。

(3)更改数据:应用更改局部数据库中的数据,将缓存中的数据标记为失效。

(4)删除数据:应用删除局部数据库中的数据,将缓存中的数据标记为失效。

[问题3] 传统的集中式数据架构只有单个数据库系统,所以要满足可扩展性的要求,只能通过硬件的方式实现,通过向上扩展(scale UP)提升系统的可扩展性。具体的实现方式包括硬件扩容(增加cpu、内存容量和磁盘数量)和硬件升级(更换高端主机或高速磁盘等)。

基于网络的分布式数据架构由多个数据库系统共同组成,可以通过向外扩展(scale out)提升系统的可扩展性,更改和优化数据分布,满足系统可扩展性的要求。具体的实现方式包括数据复制、数振垂直切分或/和水平切分、缓存和全文搜索。