Arquitetura Híbrida de Integração entre Banco de Dados Relacional e de Grafos: Uma Abordagem Aplicada à Biodiversidade

Patrícia Raia Nogueira Cavoto[1], Orientador: André Santanchè[1]

[1]Instituto de Computação – Universidade Estadual de Campinas (UNICAMP)
Nível: Mestrado – Pós-graduação em Ciência da Computação
patricia.cavoto@gmail.com, santanche@ic.unicamp.br

Ingresso no programa: Junho de 2013
Defesa da Proposta: 25 de Abril de 2014
Previsão de conclusão: Junho de 2015

Etapas já concluídas:
- Entrevista com o usuário e levantamento de requisitos
- Revisão da literatura
- Proposta da arquitetura
- Qualificação

Etapas futuras:
- Implementação da arquitetura de integração
- Testes com o usuário
- Escrita da dissertação
- Escrita de artigos

Resumo. *A complexidade e o volume dos relacionamentos entre as informações, bem como a necessidade de manter e integrar dados de estruturas heterogêneas aumentam exponencialmente a cada dia. Isto é particularmente importante no contexto de eScience, especialmente biodiversidade, área de interesse deste projeto – em que as relações são fundamentais nas análises. Neste contexto, o modelo de banco de dados de grafos pode apresentar-se como uma abordagem mais apropriada e eficiente no gerenciamento e recuperação destas informações. Em contrapartida, há um grande legado de sistemas que utilizam bancos de dados relacionais, que cumprem um papel fundamental em diversas tarefas. Apresentamos então neste trabalho uma proposta de arquitetura híbrida de integração que permite a convivência dos modelos relacional e de grafos em sua forma nativa, reduzindo o impacto de adaptações em bases relacionais preexistentes e explorando as vantagens de cada modelo nativo nas operações de gerenciamento e recuperação.*

Palavras-chave: *integração de bases, banco de dados híbrido, modelo relacional, modelo de grafos.*

1. Introdução

Nos últimos anos, com o advento da Web Semântica e a popularização do movimento NoSQL (*Not Only Structured Query Language*), a adoção de modelos de bancos de dados de grafos está se tornando cada vez mais comum. Estes se mostram particularmente apropriados quando o modelo de dados não possui uma estrutura rígida ou apresenta grandes volumes de relações, pois têm uma infraestrutura otimizada para responder a operações com ênfase nas relações [8], por exemplo, percorrer um caminho, realizar uma busca em largura ou em profundidade. Mesmo com esse crescimento na utilização de bancos de dados de grafos, temos um grande legado em todas as áreas de sistemas que utilizam o modelo relacional para armazenamento dos dados, e este modelo é a melhor opção para alguns tipos de problema. Para aproveitar as vantagens que cada modelo de dados proporciona, pode-se integra-los em uma arquitetura híbrida unificada.

O problema tratado nesta pesquisa se apresentou a partir de um projeto de colaboração com o MNHN – *Muséum national d'Histoire naturelle*[1] – envolvendo o Xper, um *software* de criação, gestão, armazenamento, análise, edição e distribuição de dados descritivos de seres vivos [13]. Nesse *software*, sempre que um novo projeto é registrado, uma nova base de dados relacional é criada para armazenar as informações. Essas bases não são compartilhadas e esta abordagem dificulta o progresso das pesquisas, uma vez que a análise dos dados em biologia é altamente interdependente.

Os dados disponibilizados nas bases estanques do Xper possuem um grande volume de relações latentes que, se pudessem ser exploradas, possibilitariam aos biólogos a descoberta de novas informações a partir dos dados já existentes. Uma abordagem relacional para este problema torna certas consultas SQL complexas e/ou ineficientes, uma vez que este modelo apresenta, entre outras coisas, restrições na realização de consultas com relacionamentos transitivos [9], por exemplo, uma consulta para identificar o impacto ambiental sofrido com a extinção de uma determinada espécie. Bancos de dados de grafos são muito eficientes neste tipo de consulta, possibilitando a utilização de algoritmos conceituados de grafos, além de permitir a criação de regras de inferência, que facilitarão ainda mais as pesquisas dos biólogos.

Grande parte dos trabalhos analisados integrando modelos distintos de dados, [5], [14] e [15], não trata a integração específica dos modelos relacional e de grafos, e outros, mais recentes, [2], [3], [6] e [7], produzem uma *view* de grafo sobre os dados relacionais, sem materializá-la ou materializando-a somente para leitura. Dessa forma, o desafio desta pesquisa se apresenta na definição de uma arquitetura híbrida que permita a integração de uma ou mais bases de dados relacionais com um banco de dados de grafos, mantendo suas representações nativas, de forma a combinar a otimização de consultas com relacionamentos complexos e a possibilidade de gravação na base de grafos.

A principal contribuição da arquitetura proposta é a combinação das seguintes estratégias: mapeamento automático ou manual de uma ou mais bases relacionais para

[1] http://www.mnhn.fr/fr/

Agradecimentos: Trabalho parcialmente financiado pelos projetos NavScales (FAPESP-MSR Virtual), MuZOO (CNPq), Complex Descriptions on the Web (CNPq), Computational Engineering and Sciences (FAPESP/Cepid), eScience (FAPESP-PRONEX) e Web Science (INCT).

grafos; materialização deste mapeamento com replicação total ou parcial dos dados; criação de novos nós e arestas no modelo de grafos; e garantia de consistência de dados, nas operações de inserção, atualização e exclusão. Uma generalização da arquitetura proposta está ilustrada na Figura 1.

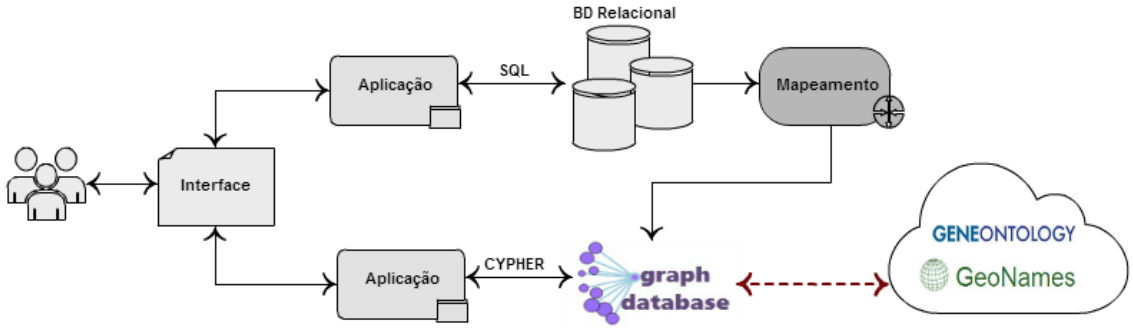

Figura 1 – Arquitetura Híbrida de Integração entre BD Relacional e de Grafos

Este trabalho está organizado da seguinte maneira: a Seção 2 apresenta a revisão da literatura e uma análise comparativa dos trabalhos relacionados; a Seção 3 detalha os desafios e contribuições; a Seção 4, trata da validação da arquitetura proposta; e, por fim, a Seção 5 apresenta as considerações finais.

2. Revisão da Literatura: Integrando Diferentes Modelos de Dados

Apresentamos a seguir uma classificação das arquiteturas propostas por trabalhos relacionados, cujo foco é a integração de dados relacionais e não estruturados, com destaque para o modelo de grafos. Essa classificação é uma contribuição deste trabalho baseada em [11] e aplicada ao cenário de integração com grafos.

2.1. Grafos em Bancos Relacionais

A base desta arquitetura é o modelo relacional, em sua forma nativa ou através da extensão de suas funcionalidades para suportar outros modelos de dados. Classificamos essa arquitetura em três principais abordagens:

(i) Mapeamento dos dados não estruturados no banco relacional: Nesta abordagem, a representação nativa não sofre modificações, pois os dados são representados no próprio modelo relacional pela criação de novas tabelas, colunas ou relações. Operações sobre os dados são realizadas utilizando operadores nativos da linguagem SQL.

(ii) Extensão do modelo relacional: O foco desta abordagem está na criação de novas estruturas na representação nativa relacional, extensões para a linguagem SQL e na aplicação de estratégias de otimização na execução das consultas. Os autores da patente descrita em [6] propõem um modelo relacional melhorado, que permite armazenar, recuperar e manipular grafos dirigidos, através da implementação de novos tipos de dados e de duas novas diretivas ao comando *SELECT*: *expand* e *depth*, utilizadas para melhorar a performance em consultas com relacionamentos transitivos. O *Database Graph Views* [7] propõe uma camada de abstração dos dados como um mecanismo para criação, visualização e manipulação de grafos, independente da organização física onde os dados originais estão armazenados. O *Virtuoso RDF View* [3] propõe o mapeamento de dados relacionais para grafos RDF sem materialização. O acesso aos dados é

realizado de forma independente, utilizando SQL ou SPARQL, ou unificada, através de um mecanismo que combina os modelos.

(iii) Incorporação de outros modelos ao relacional: Os principais fabricantes de sistemas gerenciadores de banco de dados comerciais também estão investindo esforços para contemplar a integração dos modelos, introduzindo estruturas nativas de grafos dentro de esquemas relacionais: o *Oracle Spatial and Graph* inclui um conjunto avançado de funcionalidades para tratar dados espaciais e analíticos e aplicações sociais e semânticas com grafos [10]. O DB2 RDF [4] é uma proposta para armazenamento e recuperação de informações em grafos RDF no banco relacional DB2 da IBM.

2.2. Arquitetura com Módulo de Integração

Esta arquitetura integra dois ou mais modelos de bancos de dados, mantidos em sua forma nativa, através de um módulo integrador independente, que provê uma interface de acesso unificado aos modelos, utilizando uma terceira estrutura de dados e de consulta como suporte. Cada banco de dados de origem possui um *wrapper*, com regras de mapeamento e de tradução da linguagem nativa para a unificada. Embora bastante custosa para ser implementada, essa arquitetura possibilita a interação com as diversas fontes de dados através de interface e linguagem únicas.

Seguindo esta arquitetura, o projeto *Garlic* [5] define a *Garlic's object query language*, uma extensão orientada a objetos da linguagem SQL. As consultas são distribuídas nas diversas bases e os resultados são posteriormente recuperados e agrupados. O *Garlic* possui um repositório de metadados com as informações do esquema, regras de conversão e semântica entre as diversas fontes, permitindo que novas fontes de dados sejam integradas dinamicamente. O projeto D2RQ [2] propõe a integração de ontologias RDF com bases relacionais, através de um módulo de integração que cria um grafo RDF virtual somente para consulta. A linguagem de mapeamento do D2RQ realiza as traduções necessárias e gera as sentenças SQL ou SPARQL correspondentes.

2.3. Arquitetura em Camadas

Nesse tipo de arquitetura, um modelo de banco de dados é implementado como uma aplicação que opera sobre outro modelo, tornando o modelo da camada inferior dependente do modelo implementado na camada superior. O acesso aos dados de ambos os modelos passa a ser exclusivamente realizado pelo modelo da camada superior. Em [14], um *Information Retrieval System* (IRS) é integrado com um banco de dados orientado a objetos, o VODAR. Os documentos são armazenados e indexados no IRS e o VODAR é implementado no topo da arquitetura, encapsulando as funcionalidades do IRS e mantendo uma representação abstrata da estrutura interna dos documentos. Todo o acesso ao IRS é realizado indiretamente a partir de consultas realizadas no VODAR.

Em [15], encontramos uma arquitetura que apresenta um IRS integrado a um banco de dados orientado a colunas, o MonetDB. A arquitetura é organizada da seguinte forma: o IRS como interface do usuário com o sistema, representando documentos, formulando consultas e classificando processos; a álgebra de objetos, denominada MOA, que provê a funcionalidade de gerenciar a estrutura lógica dos documentos; e, por fim, na base da arquitetura, o MonetDB, acessado pelo MOA.

2.4. Análise Comparativa das Arquiteturas

Inspirados no *framework* criado pelo W3C RDB2RDF Incubator Group [12], classificamos as arquiteturas e trabalhos apresentados anteriormente na Tabela 1, considerando os seguintes critérios: (a) Tipo de mapeamento; (b) Materializa o modelo que se integra com o relacional?; (c) Linguagem de acesso aos dados; (d) Volume de modificações nos sistemas, esquemas e dados atuais ao adotar a abordagem; (e) Esforço de implementação; (f) Coexistência de modelos nativos; e (g) Possibilidade de integrar mais de uma base de dados relacional.

Tabela 1: Análise Comparativa das Arquiteturas e Trabalhos

Arquitetura: Trabalhos	**(a)**	**(b)**	**(c)**	**(d)**	**(e)**	**(f)**	**(g)**
Mapeamento: -	Manual	N	SQL	Baixo	Baixo	N	N
Extensão: [6], [7]	Manual	N	SQL	Baixo	Baixo	N	N
Extensão: [3]	Manual e Automático	N	SQL / SPARQL	Baixo	Médio	S	S
Incorporação: [10], [4]	Semiautomático	S	SQL / SPARQL	Alto	Médio	S	N
Mód. Integração: [5]	Manual	S	Própria	Baixo	Alto	S	S
Mód. Integração: [2]	Manual e Automático	N	SQL / SPARQL	Baixo	Médio	S	N
Camadas: [14], [15]	Manual	S	Camada Superior	Alto	Alto	S	S
Nossa proposta	Manual e Automático	S	SQL / Cypher	Baixo	Médio	S	S

3. Arquitetura Híbrida de Integração entre Banco Relacional e de Grafos

O objetivo da proposta apresentada neste trabalho é definir uma arquitetura híbrida de mapeamento e integração entre bases de dados relacionais e de grafos, mantendo a representação nativa dos modelos. Será adotado o modelo de banco de dados de grafos baseado em propriedades. A principal contribuição desta proposta é uma arquitetura híbrida capaz de combinar as estratégias a seguir:

3.1. Mapeamento: Nossa arquitetura irá trabalhar com replicação total ou parcial dos dados do modelo relacional, materializando-os fisicamente na base de dados de grafos. O mapeamento dos dados poderá ser realizado de duas formas: (i) Automática: como estratégia inicial, aplicaremos as regras definidas em [1] no contexto RDF; ou (ii) Manual: permitirá a configuração de quais tabelas e/ou campos deverão ser migrados para o grafo, e qual o mapeamento específico de cada um: nós, propriedades ou arestas (relacionamentos). A configuração do mapeamento manual permite também definir regras para relacionar dados de fontes distintas.

3.2. Implementação do Mapeamento: A implementação do mapeamento será realizada por um serviço que poderá ser agendado em intervalos definidos pelo usuário. O serviço utilizará dados registrados em campos e tabelas específicos, produzidos por gatilhos nas bases relacionais, que terão a função de notificar quaisquer mudanças nos dados. Neste cenário, teremos a convivência integrada de ambos os modelos, mantendo sua representação nativa: os dados dos bancos de dados relacionais continuam a ser acessados por seus sistemas via SQL e os dados do banco de grafos serão acessados através da linguagem *Cypher*, específica do Neo4J, que será adotado nesta arquitetura.

3.3. Consistência dos Dados: Modificações na base de grafos: será possível criar novos nós, propriedades e arestas no banco de dados de grafos, cuja existência não será replicada para a base relacional; Modificações na base relacional: (i) inclusão: dados inseridos no modelo relacional serão migrados para o banco de grafos com acesso somente para consulta, limitado pela atribuição de um tipo específico "*read-only*" dentro

do banco de grafos; (ii) alteração: primeiramente realizada no modelo relacional e, posteriormente, refletida no grafo, seguindo a estratégia definida em 3.2; (iii) exclusão: o usuário poderá configurar qual política de exclusão será adotada no banco de grafos após a exclusão de um registro no banco relacional: "excluir" em que todas as propriedades, nós e arestas referentes ao registro excluído também serão excluídas, ou "manter" em que os dados serão mantidos no grafo desvinculados do modelo relacional, removendo o tipo "*read-only*" dos dados no grafo.

4. Validação da Arquitetura Proposta

Um protótipo desta arquitetura será implementado e validado na integração das diversas bases relacionais do Xper [13] (apresentado na introdução) a um banco de grafos. O protótipo será utilizado e testado pelo departamento de biologia do MNHN. O projeto pretende também resolver os problemas de compartilhamento das informações vivenciados pelos biólogos, além da exploração dos novos relacionamentos e possibilidade de realização de consultas e inferências, atualmente inviáveis ou ineficientes na arquitetura utilizada.

Etapas a serem validadas: (i) mapeamento: definição de mapeamentos automáticos e manuais entre bases de dados existentes do Xper; (ii) implementação do mapeamento: validação da base de dados de grafos mapeada de acordo com as definições de mapeamento; (iii) consistência dos dados: projeto e execução de lotes de inserção, alteração e exclusão nas bases relacionais e de grafos e verificação de consistência de acordo com as políticas implementadas; e, por fim, (iv) implementação e testes de consultas com grandes volumes de relacionamentos transitivos, executadas nas bases de dados relacionais e na base de dados de grafos, com o objetivo de avaliar os ganhos na utilização dos grafos.

5. Considerações Finais e Resultados Esperados

Existem diversas arquiteturas e abordagens disponíveis para realização da integração de dados entre modelos distintos e, a partir da análise dessas abordagens e do problema apresentado pelo MNHN, definimos o desafio desta pesquisa: criar uma arquitetura híbrida de integração entre bancos de dados relacionais e de grafos, com baixo impacto na estrutura atual, explorando as vantagens de cada modelo nativo nas operações de gerenciamento e recuperação da informação. Em referência aos trabalhos relacionados, nossa arquitetura se difere por ser capaz de combinar as seguintes estratégias: mapeamento manual/automático; materialização de mapeamentos; possibilidade de atualização e consistência nas operações executadas em ambos os modelos.

Dentre os trabalhos futuros desta pesquisa, será investigada a conexão das bases com fontes externas na forma de grafos, como o sistema GeoNames[2], que identifica unicamente todas as localidades do planeta, e a ontologia Gene Ontology[3], uma base de bioinformática que possui informações sobre genes e proteínas (vide Figura 1). Outro aspecto bastante relevante para pesquisa será a criação de uma interface para o usuário, que facilite a realização de consultas complexas envolvendo algoritmos de grafos.

[2] http://www.geonames.org/
[3] http://www.geneontology.org/

Referências

[1] Berners-Lee, T. (1998). Relational Databases on the Semantic Web. Retrieved June 10, 2014, from http://www.w3.org/DesignIssues/RDB-RDF.html

[2] Bizer, C., & Cyganiak, R. (2007). D2RQ - Lessons Learned. *W3C Workshop on RDF Access to Relational Databases*, 1–5.

[3] Blakeley, C. (2007). Virtuoso RDF Views Getting Started Guide. Retrieved June 10, 2014, from http://www.openlinksw.co.uk/virtuoso/Whitepapers/pdf/ Virtuoso_SQL_ to_RDF_Mapping.pdf

[4] Briggs, M. (2012). DB2 NoSQL Graph Store What, Why & Overview. Retrieved February 27, 2014, from https://www.ibm.com/developerworks/community/blogs/ nlp/resource/ DB2_NoSQLGraphStore.pdf?lang=en

[5] Carey, M. J., Haas, L. M., Schwarz, P. M., Arya, M., Cody, W. F., Fagin, R., Flickner, M., Luniewski, A.W., Niblack, W., Petkovic, D., Thomas, J., Williams, J.H., Wimmers, E.L. (1995). Towards heterogeneous multimedia information systems: the Garlic approach. *Proceedings RIDE-DOM'95*, 124-131.

[6] Goldberg, R. N., & Jirak, G. A. (1993). Relational database management system and method for storing, retrieving and modifying directed graph data structures. *United States Patents*. Retrieved February, 25, 2014, from http://www.google.com /patents/US5201046

[7] Gutiérrez, A., Pucheral, P., Steffen, H., & Thévenin, J. (1994). Database Graph Views: A Practical Model to Manage Persistent Graphs. *VLDB*, *33*(1), 1–20.

[8] Hecht, R., & Jablonski, S. (2011). NoSQL evaluation: A use case oriented survey. *2011 International Conference on Cloud and Service Computing*, 336–341.

[9] Kleppmann, M. (2009). Should you go Beyond Relational Databases? Retrieved March 01, 2014, from http://blog.teamtreehouse.com/should-you-go-beyond-relational-databases

[10] Oracle Corportation. (2013). Oracle Spatial and Graph Developer's Guide. Retrieved February 26, 2014, from http://docs.oracle.com/cd/E16655_01/ appdev.121/e17896.pdf

[11] Raghavan, S., & Garcia-Molina, H. (2001). Integrating Diverse Information Management Systems: A Brief Survey. *Technical Report. Stanford.*

[12] Sahoo, S. S., Halb, W., Hellmann, S., Idehen, K., Jr, T. T., Auer, S., Sequeda, J. & Ezzat, A. (2009). A survey of current approaches for mapping of relational databases to RDF. *W3C RDB2RDF Incubator Group.*

[13] Ung, V., Dubus, G., Zaragüeta-Bagils, R., & Vignes-Lebbe, R. (2010). Xper2: introducing e-taxonomy. *Bioinformatics (Oxford, England)*, *26*(5), 703–4.

[14] Volz, M., Aberer, K., Böhm, K., & GMD-IPSI, D. (1996). An oodbms-irs coupling for structured documents. *IEEE Data Eng. Bull.*, 34–42.

[15] Vries, A. De, & Wilschut, A. (1999). On the integration of IR and databases. *Proceedings of the 8th IFIP 2.6Working Conferene on Database Semantics*.