

Unificação de banco de dados para alimentar modelos ontológicos

Database unification to feed ontological models

RESUMO

Com os recorrentes avanços na Internet e na Internet das Coisas, está cada vez mais rápido adquirir zettabytes de dados de fontes e tipos diversos. A heterogeneidade de dados se torna um problema quando se deseja integrar tal montante de informação. A integração virtual de banco de dados é uma técnica versátil para solucionar este problema, focada no acesso dos dados, permitindo que eles estejam sempre atualizados sem a necessidade de realizar cópias. Este trabalho apresenta uma aplicação prática dessa técnica ao usar dois bancos de dados relacionais para alimentar uma ontologia com instâncias. Foram utilizados, principalmente, o software Stardog e a ontologia `function_cost_18#`, uma ontologia que trata da estimativa de custos nas fases iniciais do desenvolvimento do produto ponte rolante, levando em consideração as peças e componentes das pontes rolantes assim como as funções a serem realizadas. A ferramenta de mapeamento de banco de dados Stardog Mapping Syntax 2 apresentou grande rapidez e flexibilidade para unificar os dados com a ontologia, tendo a unificação validada com sucesso ao executar corretamente as questões de competência que nortearam o desenvolvimento da ontologia.

PALAVRAS-CHAVE: União de dados. Ontologia. Integração virtual. Stardog. SMS2.

ABSTRACT

WITH RECURRING ADVANCES IN THE INTERNET AND THE IOT, IT IS BECOMING FASTER AND FASTER TO ACQUIRE ZETTABYTES OF DATA FROM DIVERSE SOURCES AND TYPES. DATA HETEROGENEITY BECOMES A PROBLEM WHEN IT IS DESIRED TO INTEGRATE SUCH AN AMOUNT OF INFORMATION. VIRTUAL DATABASE INTEGRATION IS A VERSATILE TECHNIQUE FOR SOLVING THIS PROBLEM, FOCUSED ON ACCESSING DATA, ALLOWING IT TO BE ALWAYS UP TO DATE WITHOUT HAVING TO MAKE COPIES. THIS PAPER PRESENTS A PRACTICAL APPLICATION OF THIS TECHNIQUE WHEN USING TWO RELATIONAL DATABASES TO FEED AN ONTOLOGY WITH INSTANCES. MOSTLY, STARDOG SOFTWARE AND `FUNCTION_COST_18 #` ONTOLOGY WERE USED, AN ONTOLOGY THAT DEALS WITH COST ESTIMATION IN THE EARLY STAGES OF THE CRANE PRODUCT DEVELOPMENT, TAKING INTO ACCOUNT THE PARTS AND COMPONENTS OF THE CRANE AS WELL AS THE FUNCTIONS TO BE PERFORMED. THE STARDOG MAPPING SYNTAX 2 DATABASE MAPPING TOOL PROVIDED GREAT SPEED AND FLEXIBILITY TO UNIFY DATA WITH ONTOLOGY, AND THE UNIFICATION SUCCESSFULLY VALIDATED BY CORRECTLY EXECUTING THE COMPETENCY ISSUES THAT GUIDED THE DEVELOPMENT OF ONTOLOGY.

KEYWORDS: Data union. Ontology. Virtual integration. Stardog SMS2.

Caio Alexandre Soares Fabrão
caiofabrao@gmail.com
Universidade Tecnológica Federal
do Paraná, Curitiba, Paraná, Brasil

Milton Borsato
borsato@utfpr.edu.br
Universidade Tecnológica Federal
do Paraná, Curitiba, Paraná, Brasil

Rafael Voltolini
voltolini@alunos.utfpr.edu.br
Universidade Tecnológica Federal
do Paraná, Curitiba, Paraná, Brasil

Recebido: 19 ago. 2019.

Aprovado: 01 out. 2019.

Direito autorial: Este trabalho está licenciado sob os termos da Licença Creative Commons-Atribuição 4.0 Internacional.



1 INTRODUÇÃO

O mundo encontra-se na era dos dados digitais. Com os recorrentes avanços na Internet, na análise de dados e na IoT, está cada vez mais fácil adquirir zettabytes de dados em ritmo acelerado (AGRAWAL et al., 2007). Essa facilidade em aquisição de dados trouxe consigo uma grande heterogeneidade nos dados.

Os dados proveem de diversos meios, possuindo naturezas distintas umas das outras. Por conta disso, há diversas opções de soluções de armazenamento de dados, cada um com suas vantagens e desvantagens. Buscando melhor suprir as diferentes características dos dados, há, por exemplo, os métodos de sistemas de arquivos, armazenamentos de objetos, armazenamento de *blobs*, banco de dados NoSQL, entre tantos outros (LI, 2018). Em um primeiro momento, essa diversificação não apresenta grandes problemas. Porém, isso muda quando é preciso relacionar dados de fontes distintas.

Os silos de dados é a forma mais simples de solucionar esse problema. Por mais que possua vantagens, esse método demanda um alto custo de implementação, pois é preciso replicar os dados originais, centralizando-os em um novo banco de dados (WIDOM, 1995). A cópia dos dados gera outros problemas, como espaço de armazenamento, cópias desatualizadas dos dados, etc.

Em contrapartida com os silos de dados, a integração virtual busca solucionar os problemas com a duplicação dos dados. A grande diferença está na utilização de um mediador para buscar os dados necessários de cada fonte, convertendo-os em tempo de execução da consulta (HAN; JIAO, 2018). Isso significa que quanto mais fontes de dados distintos são necessários e quanto mais volátil for a natureza desses dados, isto é, o quão suscetível a alterações eles estão, mais a integração virtual se torna a solução ideal.

Este documento visa apresentar uma aplicação prática da integração virtual de dados para alimentar uma ontologia. As instâncias formam parte do alicerce de uma ontologia. Poder ter acesso fácil a dados atualizados e de diversas fontes, torna a ontologia ainda mais efetiva na tomada de decisões.

As seções seguintes desse artigo estão estruturadas da seguinte forma: a Seção 2 descreve a metodologia utilizada para realizar as atividades propostas; a análise dos resultados obtidos com o artigo são demonstrados na Seção 3; e a Seção 4 conclui com uma breve revisão e discussão sobre as atividades executadas.

2 MATERIAL E MÉTODOS

Para a realização deste projeto, foi utilizado o *software* Stardog com o intuito de integrar dois bancos de dados e alimentar a ontologia `function_cost_18#` com esses dados. A ontologia trata da estimativa de custos nas fases iniciais do produto ponte rolante, considerando as partes que compõe o produto (viga, motor, gancho, etc) e suas funções de operação (içar, deslocar em x, deslocar em y, etc) (VOLTOLINI, 2019).

Também foram utilizadas a VOWL e a IDE MySQL, para obter uma visualização gráfica da ontologia, assim realizando as primeiras análises, e para desenvolver os bancos de dados em SQL, respectivamente.

Por fim, a validação desta unificação foi realizada ao implementar as questões de competência, desenvolvidas para validar a ontologia, sobre o mapeamento realizado no Stardog.

2.1 CONSTRUÇÃO DOS BANCOS DE DADOS

O primeiro passo para a construção dos bancos de dados foi a identificação das classes cuja as instâncias deveriam ser armazenadas externamente. O processo de identificação foi realizado através de uma análise sobre a ontologia utilizando o visualizador gráfico VOWL, o que agilizou o processo de identificação. As classes que caracterizam a estrutura física de uma ponte rolante e as que caracterizam os custos de montagem e operação, deveriam ter suas instâncias armazenadas em bancos de dados. Essa decisão foi tomada uma vez que a natureza desses dados é massiva e/ou sujeitas a constantes atualizações.

As classes da estrutura física se agrupam em torno de três montagens: a ponte, o carro e o guindaste. As classes referentes ao custo se agruparam em três categorias: peças compradas, peças produzidas na própria empresa e depreciação por desgaste, sendo cada uma representada em uma tabela.

No MySQL, foram implementados dois bancos de dados, *overheadcrane* e *financial*. O primeiro ficou responsável por armazenar os dados referentes as peças e montagens, enquanto que o segundo ficou responsável pelos dados de custo, assim caracterizando dois setores em uma empresa: montagem e financeiro.

A população de dados do setor de montagem foi retirada dos catálogos em CIMAF (2014) e em CSM (2019). Para o setor financeiro, os custos dos componentes foram adquiridos realizando os cálculos das médias dos preços praticados no mercado brasileiro (MERCADO LIVRE, 2019).

2.2 MAPEAMENTO DOS DADOS RELACIONAIS

O mapeamento dos dados relacionais em SQL para RDF foi realizado utilizando a *Stardog Mapping Syntax 2* (SMS2). A estruturação do mapeamento em SMS2 começa com a definição do cabeçalho de prefixos e da consulta que extrairá os dados (linguagem alvo, colunas a serem selecionadas e a tabela de origem). A Figura 1 apresenta o início do mapeamento da tabela motor.

Figura 1 – Prefixos e início do mapeamento em SMS2.

```
1 PREFIX : <ontologies/function_cost_18#>  
2 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>  
3  
4 MAPPING  
5 FROM SQL {  
6   SELECT *  
7   FROM 'motor'  
8 }
```

Fonte: Do autor (2019).

Em seguida foi criado o RDF dos dados selecionados, associando os *Internationalized Resource Identifiers* (IRIs) e os valores das colunas pertinentes por meio das propriedades definidas na ontologia (Figura 2).

Figura 2 – Definição do RDF dos dados mapeados em SMS2.

```
1 to {
2   ?subject rdfs:label ?name .
3   ?subject :hasMotorPower ?hasMotorPower .
4   ?subject rdf:type :Motor
5 }
```

Fonte: Do autor (2019).

Por fim, tem-se a definição dos IRIs e as conversões de tipos de dados (Figura 3).

Figura 3 – Definição dos IRIs e transformação de tipos de dados em SMS2.

```
1 WHERE {
2   BIND(template(
3     "ontologies/function_cost_18#Motor{idMotor}"
4     ) AS ?subject)
5   BIND(xsd:double(?power) AS ?hasMotorPower)
6 }
```

Fonte: Do autor (2019).

O processo de mapeamento das tabelas foi repetido para todas as tabelas do *overheadcrane* e do *financial*, agrupando-as em dois blocos de mapeamento, um para cada banco.

2.3 VALIDAÇÃO DO MAPEAMENTO

Para finalizar, foi realizada a conversão das questões de competência propostas em Voltolini (2019) em consultas SPARQL. As questões de competência visam validar a ontologia *function_cost_18#* como um todo, implicando em elas serem bons parâmetros para validar também o mapeamento e integração dos bancos de dados.

Como o Stardog apresenta os mapeamentos na forma de grafos virtuais, as consultas em SPARQL podem ser facilmente escritas como se cada mapeamento representasse um grafo nomeado da ontologia. Assim, a Figura 4 apresenta a implementação da pergunta “qual o custo de um motor com potência maior que 50 cavalos” em SPARQL.

Figura 4 – Exemplo de consulta SPARQL utilizando grafos virtuais.

```
1 SELECT ?motorName ?cost WHERE{
2   graph <virtual://overheadcrane>{
3     ?motor a :Motor;
4     ?motor rdfs:label ?motorName;
5     ?motor :hasMotorPower ?power.
6   }
7   graph <virtual://financial>{
8     ?f rdfs:label ?motorName;
9     ?f :hasCost ?cost.
10  }
11  FILTER (?power > 50)
12 }
```

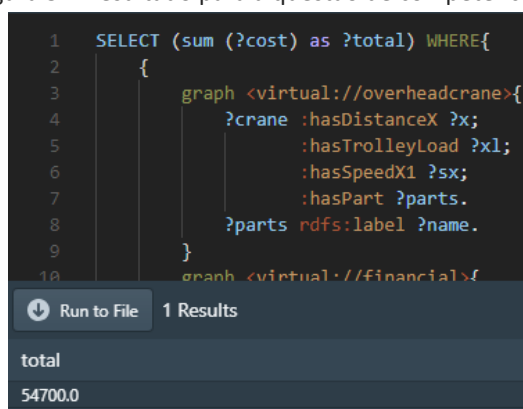
Fonte: Do autor (2019).

3 RESULTADOS E DISCUSSÃO

Foram obtidos com esse trabalho dois grafos virtuais que unem os bancos de dados com a ontologia, sendo um grafo para cada banco. Os grafos virtuais proporcionam duas facilidades ao usuário da ontologia: trazer os dados mais atuais dos bancos de dados e requerer do usuário apenas o conhecimento de SPARQL.

A validação da integração da ontologia com os bancos de dados foi assegurada ao receber corretamente a resposta das questões de competência. A Figura 5 apresenta a resposta para a questão de competência 1. Ela busca saber o custo total do produto que atende uma dada especificação de operação. Nela, a somatória dos custos das peças que melhor atendem as necessidades descritas foi no valor de 54.700,00 reais.

Figura 5 – Resultado para a questão de competência 1.



```

1 SELECT (sum (?cost) as ?total) WHERE{
2 {
3   graph <virtual://overheadcrane>{
4     ?crane :hasDistanceX ?x;
5           :hasTrolleyLoad ?xl;
6           :hasSpeedX1 ?sx;
7           :hasPart ?parts.
8     ?parts rdfs:label ?name.
9   }
10  graph <virtual://financial>{

```

Run to File 1 Results

total
54700.0

Fonte: Do autor (2019).

Ao alterar a consulta para retornar os valores das variáveis utilizadas, a Tabela 1 é obtida. As colunas *crane* e *parts* são nativas do banco *overheadcrane*; as colunas *cost*, *xp*, *yp* e *zp*, do banco *financial*; e a coluna *name* é comum a ambos. Os valores da Tabela 1 são os componentes da montagem desejada e, assim como descritos em Voltolini (2019), são apresentados 16 itens. Ao somar os valores da coluna *cost* para identificar o custo total do produto, obtém-se o valor de 54.700,00 reais, assim validando o processo de unificação dos dados.

Tabela 1 – Resultado da questão de competência 1 adaptada.

crane	parts	name	cost	xp	yp	zp
:Trolley3	:Accessory1	anticollisionDetector	2.000,00	:Purchase1		
:Trolley3	:Motor2	motor02	6.000,00	:Purchase10		
:Trolley3	:Reducer2	reducer02	2.200,00	:Purchase12		
:Trolley3	:Accessory2	severeUse	2.500,00	:Extra1		
:Trolley3	:Drum2	drum02	2.000,00	:Production2		
:Trolley3	:PulleyCompensationHoist2	puleyCompensation02	1.000,00	:Production8		
:Bridge3	:Accessory1	anticollisionDetector	2.000,00		:Purchase1	
:Bridge3	:Accessory2	severeUse	2.500,00		:Extra1	

(continuação)

crane	parts	name	cost	xp	yp	zp
:Bridge3	:Girder2	beam02	8.200,00		:Production5	
:Hoist3	:Accessory1	anticollisionDetec tor	2.000,00			:Purchase1
:Hoist3	:Accessory2	severeUse	2.500,00			:Extra1
:Hoist3	:Accessory5	handleTube	7.000,00			:Purchase3
:Hoist3	:Control3	control03	5.000,00			:Purchase8
:Hoist3	:Hook1	hook01	6.000,00			:Production6
:Hoist3	:PulleyHoist2	pulleyHoist02	1.300,00			:Production10
:Hoist3	:Rope2	rope02	2.500,00			:Purchase14

Fonte: Do autor (2019).

A plataforma Stardog é a responsável por integrar os dados provenientes dos bancos de dados implementados e deixá-los aptos a serem utilizados pela ontologia. A ontologia é a responsável por analisar as relações entre os dados e inferir novas informações.

4 CONCLUSÃO

Esta pesquisa demonstrou a praticabilidade da utilização da integração virtual para unificar banco de dados e ontologias. A técnica de integração virtual possibilita que ontologistas levem menos tempo para adicionar grandes quantidades de instâncias na ontologia, podendo focar no desenvolvimento dela.

A virtualização permite também que a responsabilidade na manutenção da consistência das instâncias fique centralizada no administrador dos bancos de dados, descartando a necessidade de retrabalho ao verificar se as cópias também estão concisas.

Foi demonstrado, de forma empírica, a versatilidade e facilidade de implementação da unificação de dados. A integração virtual dos bancos foi rapidamente executada devido à flexibilidade no mapeamento proporcionada pela SMS2 e sua validação atestada ao executar corretamente as questões de competências que nortearam a concepção da ontologia.

Para pesquisas futuras, fica aberta a oportunidade de verificar os ganhos e perdas ao se empregar a integração virtual para alimentar ontologias. Pode-se determinar os impactos no tempo de processamento das consultas SPARQL, ao considerar o tempo de acesso aos bancos. É possível também avaliar o esforço de implementação do mapeamento em relação a adicionar manualmente as instâncias na ontologia.

AGRADECIMENTOS

Agradeço a todas as pessoas e instituições que contribuíram para a realização deste trabalho, através da dedicação na orientação, da oferta de infraestrutura física e, em especial, ao CNPq (Conselho Nacional de

Desenvolvimento Científico e Tecnológico), pela bolsa de estudos e auxílio financeiro que possibilitou a dedicação ao programa de iniciação científica.

REFERÊNCIAS

AGRAWAL, N. et al. A five-year study of file-system metadata. ACM Transactions on Storage, [S. l.], v. 3, p. 9-41, out. 2007.

CIMAF. Manual Técnico de Cabos. 2014. Disponível em:
<https://www.aecweb.com.br/cls/catalogos/aricabos/catalogocimaf2014completo.pdf>. Acesso em: 12 abr. 2019.

CSM. Catálogo de Produtos. 2019. Disponível em:
http://www.csm.ind.br/catalogo_engenharia.pdf. Acesso em: 12 abr. 2019.

HAN, J.; JIAO, Y. Conception of QAR Data Application Mode Based on Virtual Integration. 2018 IEEE International Conference of Safety Produce Informatization (IICSPI). p.392–395, 2018. IEEE.

LI, Haoyuan. Alluxio: A Virtual Distributed File System. 2018. 93 f. Tese (Doutorado em Ciência da Computação) - Universidade da Califórnia, Berkeley, 2018.

Mercado Livre. Mercado Livre Brasil. 2019. Disponível em:
<https://www.mercadolivre.com.br/>. Acesso em: 13 abr. 2019.

VOLTOLINI, Rafael. Cost Estimation In Initial Development Stages Of Products - An Ontological Approach. 2019. 109 f. Dissertação (Mestrado em Engenharia Mecânica) - Universidade Tecnológica Federal do Paraná, Curitiba, 2019.

WIDOM, Jennifer. Research problems in data warehousing. Conference on Information and Knowledge Management, p. 25–30, nov. 1995.