Departamento de Ciência da COMPUTAÇÃO

UDESC UNIVERSIDADE DO ESTADO DE SANTA CATARINA

FragRDF: Um Fragmentador de dados RDF baseado em Esquemas

Laís Pisetta Van Vossen
lais.vossen@gmail.com

Vinicius Gasparini
v.gasparini@edu.udesc.br

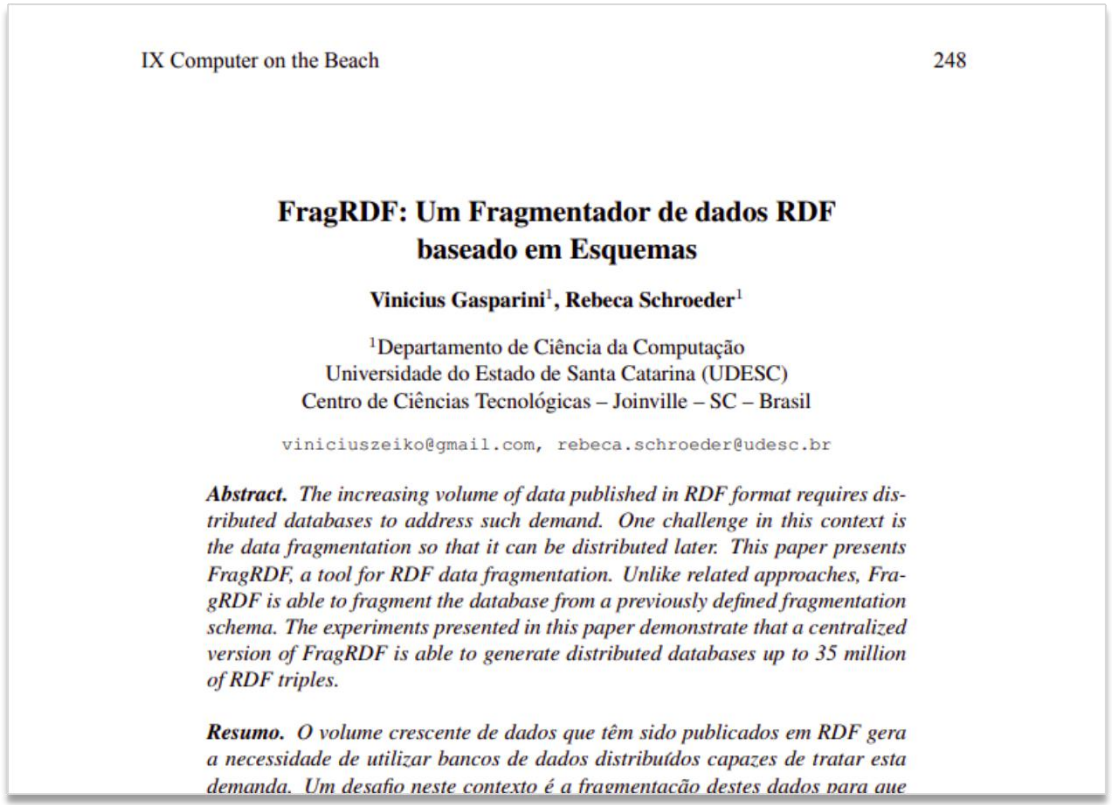
Departamento de Ciência da Computação
Centro de Ciências Tecnológicas
Universidade do Estado de Santa Catarina

2020

Departamento de Ciência da COMPUTAÇÃO

UDESC UNIVERSIDADE DO ESTADO DE SANTA CATARINA

Introdução



IX Computer on the Beach 248

FragRDF: Um Fragmentador de dados RDF baseado em Esquemas

Vinicius Gasparini[1], Rebeca Schroeder[1]

[1]Departamento de Ciência da Computação
Universidade do Estado de Santa Catarina (UDESC)
Centro de Ciências Tecnológicas – Joinville – SC – Brasil

viniciuszeiko@gmail.com, rebeca.schroeder@udesc.br

Abstract. *The increasing volume of data published in RDF format requires distributed databases to address such demand. One challenge in this context is the data fragmentation so that it can be distributed later. This paper presents FragRDF, a tool for RDF data fragmentation. Unlike related approaches, FragRDF is able to fragment the database from a previously defined fragmentation schema. The experiments presented in this paper demonstrate that a centralized version of FragRDF is able to generate distributed databases up to 35 million of RDF triples.*

Resumo. *O volume crescente de dados que têm sido publicados em RDF gera a necessidade de utilizar bancos de dados distribuídos capazes de tratar esta demanda. Um desafio neste contexto é a fragmentação destes dados para que*

[IX Computer on the Beach]

UDESC UNIVERSIDADE DO ESTADO DE SANTA CATARINA

Seminário de Iniciação Científica
Universidade do Estado de Santa Catarina
28º SIC UDESC

FRAGRDF: UM FRAGMENTADOR DE DADOS RDF BASEADO EM ESQUEMAS

Vinicius Gasparini [1], Rebeca Schroeder Freitas [2]

[1] Acadêmico(a) do Curso de Ciência da Computação CCT – bolsista PROIP/UDESC
[2] Orientador, Departamento de Ciência da Computação CCT – rebeca.schroeder@udesc.br

Palavras-chave: RDF. Fragmentador. Banco de Dados.

Atualmente, RDF (*Resource Description Framework*) é o padrão para a publicação de dados na Web. Fontes de dados RDF são definidas por conjuntos de triplas interligadas, e que podem ser generalizadas por uma estrutura de dados em grafos. Diante da disseminação deste padrão, o excessivo volume de dados neste formato gera desafios de gerenciamento, onde muitas fontes têm seu volume qualificado como *Big Data*.

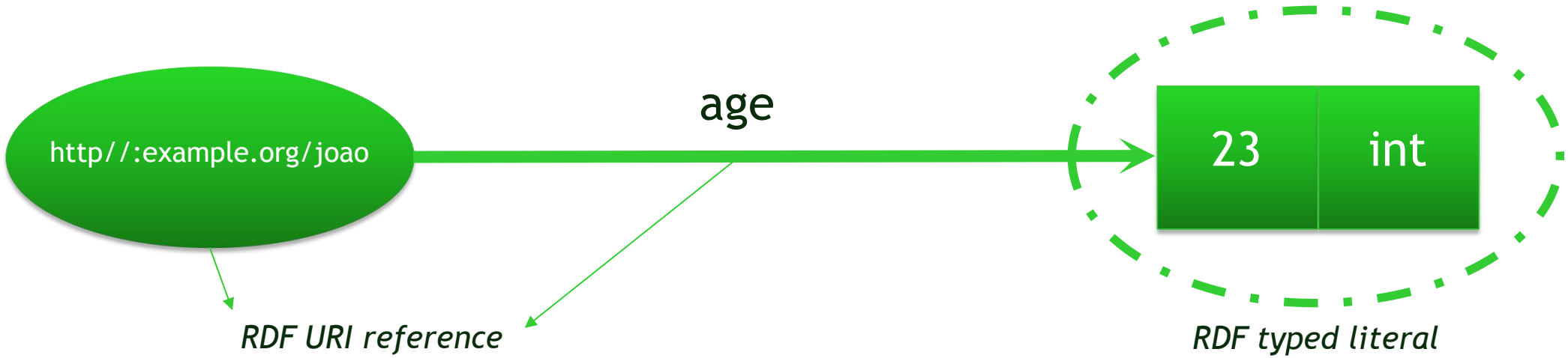
O problema relacionado a esse grande volume vem sendo tratado através da adoção de sistemas distribuídos ou paralelos. Entretanto, existem diversos métodos para fragmentação de dados neste contexto. Uma classe de trabalhos opera considerando conhecimentos da carga de trabalho dos bancos para determinar como os dados relacionados podem ser melhor agrupados. Nesta categoria, uma abordagem é a definição prévia da fragmentação sobre um esquema RDF conhecido. De posse do esquema, alguns trabalhos analisam as principais consultas e determinam como os elementos do esquema são acessados conjuntamente ([Curino et al. 2010], [Schroeder and Hara 2015]).

Conhecido esse cenário, uma ferramenta chamada *FragRDF* foi desenvolvida para fragmentar dados RDF utilizando um esquema de fragmentação pré-definido. *FragRDF*

[28º Seminário de Iniciação Científica]

RDF - *Resource Description Framework*

Sujeito *(RDF subject)* <http://example.org/joao>	Predicado *(RDF predicate)* <http://example.org/ns#age>	Objeto *(RDF object)* “23”<http://example.org/int>

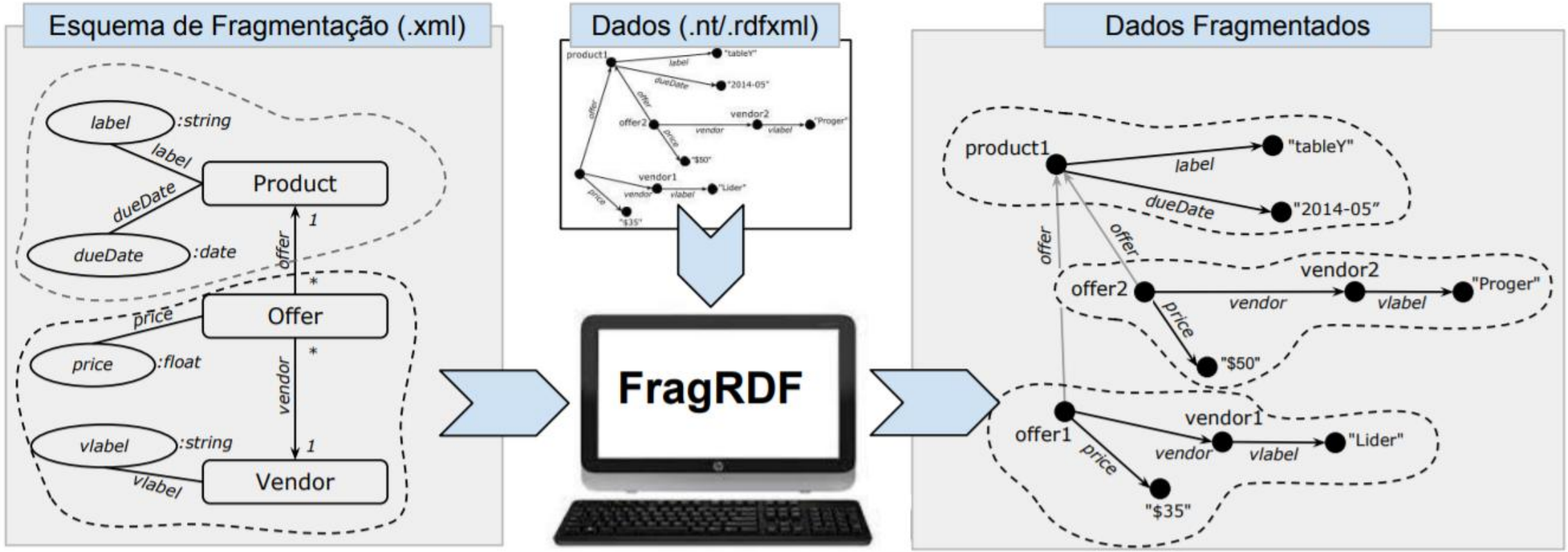


RDF no formato *N-Triples e* Esquema de Fragmentação no formato XML

```
<http://www4.wiwiss.de/Product1> <http://www4.wiwiss.de/#type> <http://www4.wiwiss.de/Product> .
<http://www4.wiwiss.de/Product1> <http://www4.wiwiss.de/#label> "tableY" .
<http://www4.wiwiss.de/Product1> <http://wiwiss.de/#dueDate> "2014-05"<http://wiwiss.de/#date> .
<http://www4.wiwiss.de/Offer1> <http://www4.wiwiss.de/#type> <http://www4.wiwiss.de/Offer> .
<http://www4.wiwiss.de/Offer1> <http://www4.wiwiss.de/product> <http://www4.wiwiss.de/Product1> .
<http://www4.wiwiss.de/Offer1> <http://www4.wiwiss.de/vendor> <http://www4.wiwiss.de/Vendor1>
<http://www4.wiwiss.de/Offer1> <http://www4.wiwiss.de/#price> "35.00"<http://www4.wiwiss.de/USD> .
```

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <schemafrag> <!-- definitions for fragmentation schema -->
3      <items> <!-- map elements/attributes to fragments -->
4          <item name="label" entity="Product" frag="1" />
5          <item name="dueDate" entity="Product" frag="1" />
6          <item name="offer" entity="Offer" frag="2" />
7          <item name="price" entity="Offer" frag="2" />
8          <item name="vendor" entity="Offer" frag="2" />
9          <item name="vlabel" entity="Vendor" frag="2" />
10     </items>
11 </schemafrag>
```

Funcionamento do FragRDF



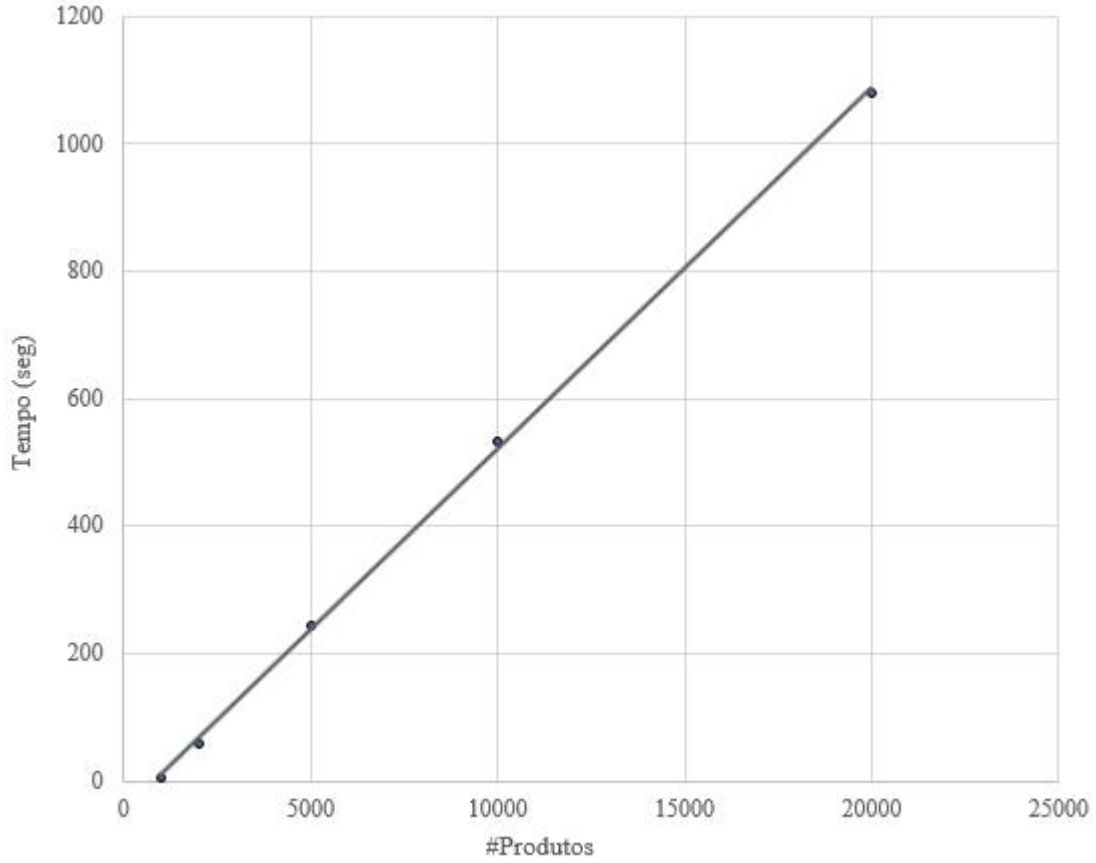
Experimentos

- Fragmentação em uma única máquina:
  - Diversos tamanhos de BDs
  - Quantidade/Tamanho arquivos BD

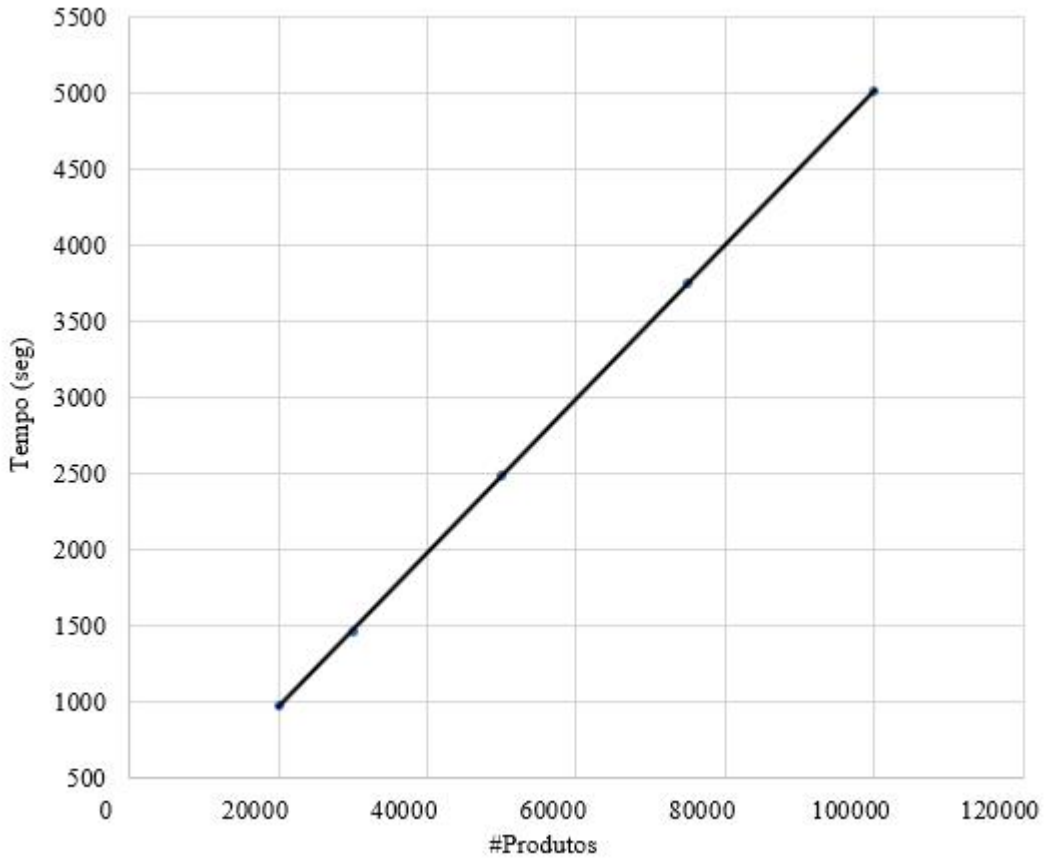
- Berlin SPARQL Benchmark (BSBM)
  - Caso de uso de um sistema de *e-commerce*
  - Fator de escala baseado no número de produtos a gerar.

Resultados

Experimento 1



Experimento 2



Resultados e Conclusões

- FragRDF consegue gerar bases de até 35,27 milhões de triplas em uma única máquina.
- 20 mil produtos para um único arquivo de entrada.
- Usando 180 MB por arquivo de entrada alcançamos 100 mil produtos.
- Não foi possível o teste com bases de tamanhos superiores em virtude do limite de memória da máquina utilizada.
- Não foram encontrados outros limitantes com exceção do gargalo de memória.
- O algoritmo apresenta comportamento linear em função do crescimento da escala de produtos.