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* espera como entrada um conjunto de dados que podem envolver um ou mais arquivos nos formatos NT ou RDF/XML bem como o esquema de fragmentação. Esse esquema é descrito por um arquivo XML que define a qual fragmento cada elemento deve ser colocado. A ferramenta então analisa toda informação do conjunto e determina a qual elemento do esquema ele se refere. Uma vez identificado, a tripla é encaminhada ao fragmento respectivo.

FragRDF foi desenvolvida em Python, com o suporte das bibliotecas *ElementTree* e *RDF Lib*. *ElementTree* é uma biblioteca *built-in* do Python que foi utilizada para processar o arquivo de entrada definido em XML. A biblioteca *RDFLib*[1] por sua vez foi utilizada para processar os dados de entrada contendo os arquivos de dados RDF em formato NT e RDF/XML ([Gasparini and Schroeder 2018]).

Um estudo experimental foi realizado para avaliar o desempenho de *FragRDF*. A avaliação foi executada em duas etapas em uma única máquina. O ambiente de execução conta com um processador Intel Core i7 3632QM de 4 núcleos e 2,2 gHz, 8Gb de RAM e sistema operacional Fedora 26. O conjunto de dados a ser fragmentado foi obtido a partir do gerador de bases RDF do Berlin SPARQL Benchmark (BSBM)[2]. O esquema RDF do BSBM foi fragmentado de forma que seus elementos foram distribuídos em 6 tipos de fragmentos.

Em uma primeira avaliação, as bases foram geradas no formato NT em um único arquivo. Os resultados apresentados pela Figura 1 correspondem ao tempo de resposta em segundos para a completa execução do processo *FragRDF*. Durante os testes foi atingido o limite de memória para bases de escala superior a 20.000 produtos. Uma das razões para a elevação do tempo de resposta para a base de 20.000 produtos é o tamanho do arquivo único da base que foi de 919,09 MB. Esta constatação foi comprovada pelo segundo experimento, em que

[1] https://github.com/RDFLib/rdflib

[2] Bizer,C.andSchultz,A.(2009). TheBerlinSPARQLBenchmark. InInternational Journal on Semantic Web & Information Systems.

optou-se por gerar a base em mais arquivos de no máximo 180 MB cada. Nesse experimento foram geradas bases de 20.000 a 100.000 produtos dispostos em *p/5.000* arquivos, onde *p* é a escala de produtos. Os resultados deste segundo experimento estão apresentados na Figura 2, onde atingiu-se a marca de 5.014,24 segundos para a base de 100.000 produtos, que corresponde a 35,4 milhões de triplas.

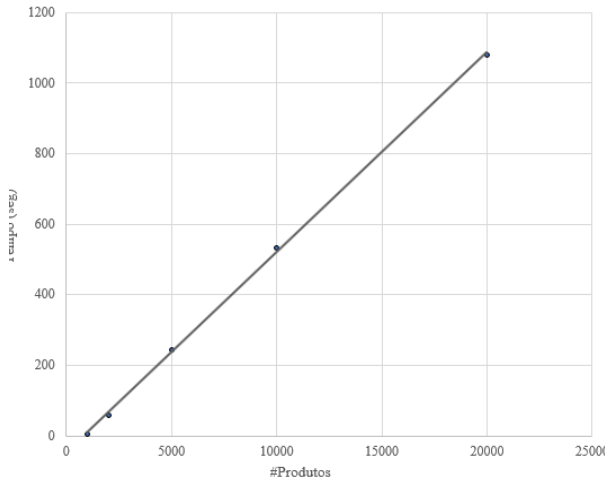

Fig. 1: Base em arquivo único

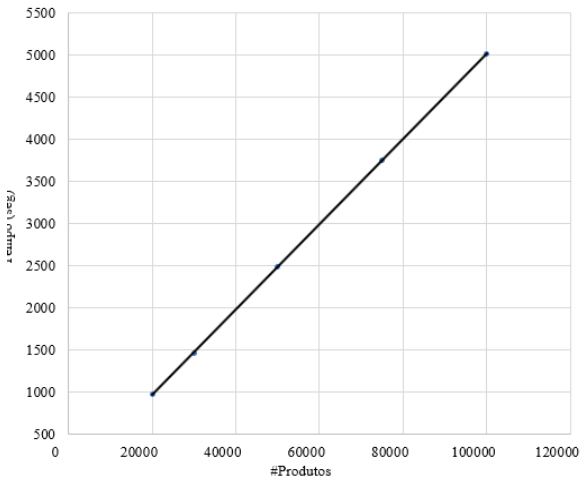

Fig. 2: Base em vários arquivos

O resultado experimental executado mostra que *FragRDF* conseguiu processar bases de até 35 milhões de triplas em uma única máquina. Por se tratar de uma instalação centralizada, não foi possível o teste com bases de tamanhos superiores pois o limite de memória da máquina foi atingido. Não foram encontrados outros limitantes com exceção do gargalho de memória. Quanto ao tempo de execução, o algoritmo apresenta comportamento linear em função do crescimento da escala de produtos.

Este trabalho teve como foco o desenvolvimento da ferramenta *FragRDF* que é capaz de fragmentar conjuntos de dados RDF a partir de um esquema de fragmentação pré-estabelecido. Entretanto, a ferramenta desenvolvida limita-se ao processamento em uma única máquina, o que impõe restrições à escalabilidade da ferramenta. Como algumas bases RDF podem atingir a casa de trilhões de triplas, uma proposta é a implementação futura de uma versão paralela da ferramenta. Espera-se que deste modo, ao usar *frameworks* de processamento paralelo, o processo *FragRDF* se torne escalável.

Referências

Curino, C., Jones, E., Zhang, Y., and Madden, S. (2010). Schism: A workload-driven approach to database replication and partitioning. Proc. VLDB Endow., 3(1-2):48–57.

Gasparini, V. and Schroeder R. (2018) FragRDF: Um Fragmentador de dados RDF baseado em Esquemas. In IX Computer on the Beach, 248-257.

Schroeder, R. and Hara, C. S. (2015). Partitioning templates for rdf. In Advances in Databases and Information Systems, pages 305–319. Springer International.