AVALIAÇÃO DO DESEMPENHO DE FERRAMENTAS PARA PERSISTÊNCIA DE DADOS EM FORMATO DE TRIPLAS

FELIPE LUZZARDI DA ROSA[1]; ROGER DA SILVA MACHADO[2], ADENAUER CORRÊA YAMIN[2]; ANA MARILZA PERNAS[3]

[1]*Universidade Federal de Pelotas – fldrosa@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas – {rdsmachado, adenauer}@inf.ufpel.edu.br*
[3]*Universidade Federal de Pelotas – marilza@inf.ufpel.edu.br*

1. INTRODUÇÃO

Aplicações cientes de contexto desenvolvidas em áreas como Big Data, Internet das Coisas (IoT) e Web Semântica, utilizam fontes de informação para tomadas de decisões, sendo estas informações disponibilizadas por ontologias. Embora diversas propostas apresentadas neste âmbito façam uso de tais informações, poucas versam sobre a maneira de garantir a persistência das mesmas, ou então, quando fazem, utilizam repositórios relacionais para tal fim (VEIGA 2016, SENA 2016). Uma forma mais eficiente de manter a persistência de dados ontológicos é utilizando um modelo de armazenamento em triplas RDF (*Resource Description Framework*).

O modelo de armazenamento de ontologias na forma de triplas pode ser dividido em três grandes categorias com base na arquitetura de sua implementação: em memória; nativo; e externo. No armazenamento em memória, todo o conjunto de triplas é mantido na memória principal do dispositivo para manipulação, o que o torna ineficiente quando o volume de dados é grande (MAHARAJAN 2012). O modelo de armazenamento nativo fornece persistência pelo uso de uma base de dados. As ferramentas que implementam este modelo oferecem seus próprios recursos de manipulação da base de dados, utilizando linguagens como a SPARQL para acessá-lo. Na categoria de armazenamento externo, as triplas são persistidas em bancos de dados de terceiros, podendo, neste caso, serem utilizadas bases de dados não apropriadas para triplas, como uma base de dados relacional (CAN et al. 2017).

Nota-se que o armazenamento nativo de triplas está ganhando força e popularidade, em grande parte devido ao seu desempenho por contar com uma base de dados apropriada para o modelo de triplas (RDF) (BioOntology 2011). Apesar disso, poucos trabalhos presentes na literatura tratam efetivamente sobre a persistência e a manipulação dos dados provenientes de ontologias, sendo estas tarefas fundamentais para o provimento da ciência de contexto. A manipulação eficiente desses dados tem impacto nos serviços oferecidos, podendo tornar o processo de tomada de decisão mais eficiente.

Visando contribuir para a pesquisa na área, neste artigo é analisado o desempenho de cinco ferramentas populares para armazenamento nativo de triplas: AllegroGraph, GraphDB, MarkLogic, Stardog e Virtuoso, para identificação daquela mais apta a compor parte da solução proposta para provimento de ciência de contexto.

2. METODOLOGIA

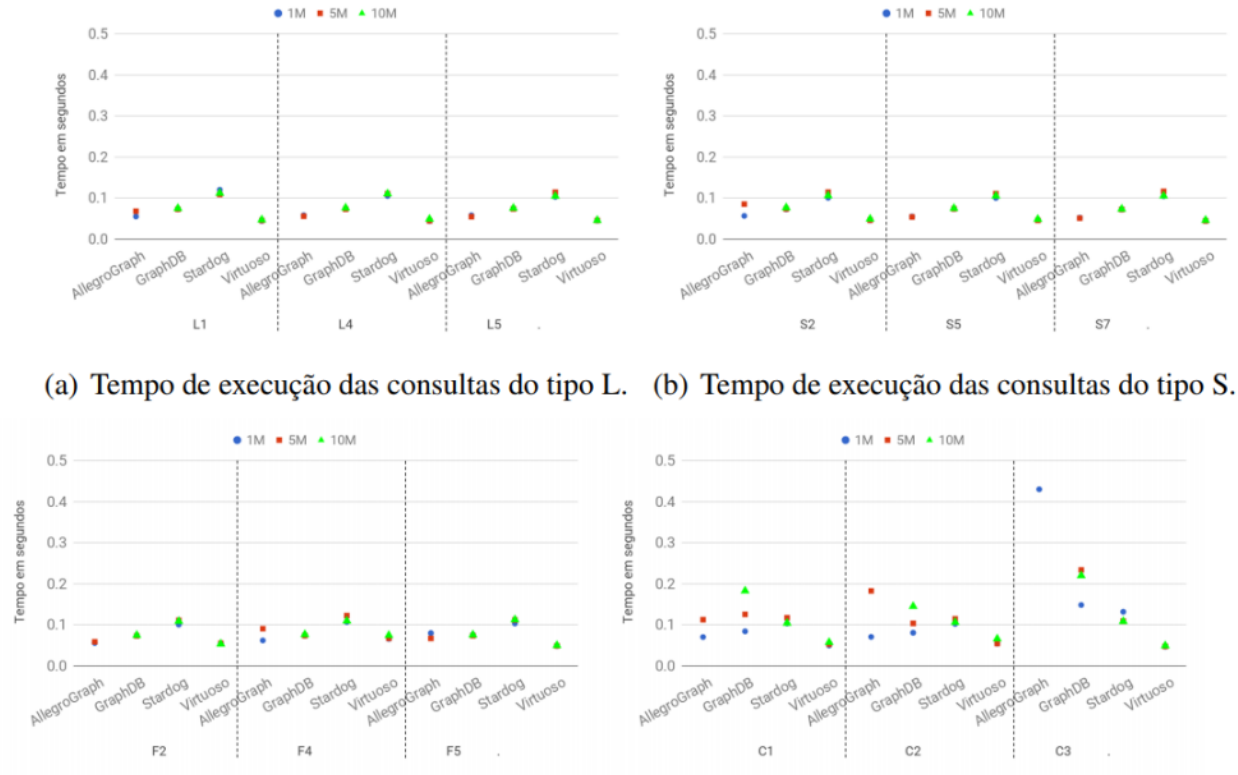
O estudo de caso foi realizado aplicando o *benchmark* WatDiv (ALUÇ 2014), o qual disponibiliza um número diverso de consultas para teste, possibilitando uma

análise mais criteriosa das ferramentas. Os experimentos foram realizados em duas dimensões, na primeira foi analisada a escalabilidade de cada ferramenta, considerando diferentes tamanhos da base de dados. Na segunda os desempenhos fornecidos pelas ferramentas foram comparados entre si. O objetivo do experimento é o de identificar a ferramenta que apresente a melhor relação entre escalabilidade e desempenho.

Foram realizados testes com 30 execuções, sendo estas focadas no tempo de execução de doze consultas SPARQL, utilizando *datasets* com 1, 5 e 10 milhões de triplas nas cinco ferramentas analisadas. Os testes com 10 milhões de triplas não foram realizados na ferramenta AllegroGraph devido a limitações de sua licença gratuita.

3. RESULTADOS E DISCUSSÃO

Os tempos médios de execução em segundos para cada tipo de consulta oferecida pelo *benchmark*, “Lineares" (L), “Estrela" (S) ,”Floco de Neve" (F) e “Complexas" (C), são apresentados na Figura 1. Os desempenhos com a ferramenta MarkLogic não são apresentados pois, além de ter um desempenho muito inferior às demais ferramentas, o desvio padrão se apresentou elevado.



(a) Tempo de execução das consultas do tipo L. (b) Tempo de execução das consultas do tipo S.

(c) Tempo de execução das consultas do tipo F. (d) Tempo de execução das consultas do tipo C.

Figura 1. Tempo de execução, em segundos, nos diferentes tipos de consultas.

Com os resultados apresentados é possível concluir que a ferramenta Virtuoso apresentou os melhores resultados de desempenho na grande maioria das consultas. Também é possível observar que as ferramentas Virtuoso e Stardog apresentaram boa escalabilidade, com baixa variação de tempo de

execução entre os diferentes tamanhos de base de dados, enquanto que as ferramentas AllegroGraph e GraphDB apresentaram uma escalabilidade também boa, porém inferior as duas ferramentas anteriormente mencionadas. Esta conclusão foi validada aplicando o teste t de Student com intervalo de confiança de 95\% entre as cinco ferramentas, em cada um dos três tamanhos e em cada uma das 12 consultas. Como resultados destes testes foram detectados que os resultados da consulta C1 com o tamanho 5M entre Stardog e AllegroGraph e entre Stardog e GraphDB não obtiveram diferenças estatísticas significativas. Ainda, a consulta C3 para o tamanho 1M entre Stardog e GraphDB também não obteve diferenças estatísticas significativas. Os demais resultados apresentaram diferenças estatísticas significativas, assegurando os resultados obtidos e apresentados na Figura 1.

Também pode ser notado que o Stardog obteve resultados muito similares em todos os testes realizados, tendo sido a ferramenta com menor alteração de desempenho entre tamanhos diferentes de base de dados. Além dos testes entre as ferramentas, foi analisado se o tamanho da base impactava no desempenho dos sistemas. Para isso, foi realizado o teste t entre as médias de execução de cada ferramenta comparando o tamanho 1M com 5M e 5M com 10M. Analisando os resultados obtidos, pode-se observar que a ferramenta AllegroGraph foi a que apresentou menor número de consultas com diferenças não significativas. No entanto, relembra-se que nesta ferramenta não foi possível realizar consultas em uma base de tamanho de 10M. A ferramenta que apresentou maior número de diferenças não significativas foi a MarkLogic, pois foi a ferramenta que apresentou altos valores para os desvios padrões anotados.

4. CONCLUSÕES

Este trabalho realizou uma comparação de desempenho relacionada ao tempo de execução de consultas SPARQL entre cinco ferramentas de armazenamento em formato de triplas: AllegroGraph, GraphDB, MarkLogic, Stardog e Virtuoso. Os testes consideraram o tempo de execução, em segundos, de doze diferentes consultas em três tamanhos de bases de dados. Ao fim dos testes foi possível concluir que a ferramenta Virtuoso foi a que mostrou o melhor desempenho geral, apresentando o tempo de execução mais baixo na grande maioria das consultas, apesar de apresentar uma flutuação no desempenho um pouco maior que Stardog quando varia-se a quantidade de dados tratada.

Estes resultados foram importantes nas decisões de pesquisa a serem tomadas pelo grupo, pois forneceram embasamento para a escolha da ferramenta mais adequada para gerenciar o modelo de triplas implementado no *middleware* para Computação Ubíqua desenvolvido pelo grupo. Esta escolha impacta diretamente nos serviços cientes de contexto oferecidos, pois as aplicações enviam continuamente requisições de consulta aos contextos gerenciados pelo *middleware* para sua tomada de decisão, sendo importante se ter o menor tempo possível de acesso aos dados.

Como trabalhos futuros destaca-se a continuidade da pesquisa ligada a provimento de serviços cientes de contexto, aplicando a ferramenta Virtuoso para acesso aos dados do modelo de triplas.

5. REFERÊNCIAS BIBLIOGRÁFICAS

ALUÇ¸, G.. **Diversified Stress Testing of RDF Data Management Systems.** 2014. Dissertação de mestrado - David R. Cheriton School of Computer Science, Waterloo, ON, Canada.

BioOntology. **Comparison of Triple Stores.** 2011. Acessado em novembro 2016. Online. Disponível em: https://www.bioontology.org/wiki/images/6/6a/Triple_Stores.pdf.

CAN, O.; SEIZER, E.; BURSA, O.; UNALIR, M. O.. **Comparing relational and ontological triple stores in healthcare domain**. 2017. In Entropy, 19(1):30.

MAHARAJAN, S.. **Performance of native SPARQL query processors**. 2012. Dissertação de mestrado - Department of Information Technology, Uppsala University.

SENA, M. V. O.; BULCAO NETO, R. F.. A solution to discard context information using metrics, ontology and fuzzy logic. In **WebMedia - Brazilian Symposium on Multimedia and the Web**, Teresina, 2016.

VEIGA, E. F; NETO, R. F. B. Um serviço de representação ontológica de contexto baseada no padrão de projeto estímulo-sensor-observação. In **SBCUP – 8º Simpósio Brasileiro de Computação Ubíqua e Pervasiva,** Porto Alegre, 2016.