Avaliação de Eficiência de Ferramentas de Perfilamento de Dados

Levy Souza[1], Dimas Cassimiro Nascimento[1]

[1]Unidade Acadêmica de Garanhuns (UAG) – Universidade Federal Rural de Pernambuco (UFRPE)
Garanhuns – PE – Brasil

{levybatera,dimascnf}@gmail.com, dimascnf@uag.ufrpe.br

Abstract. *Data profiling tools provide important features in order to analyze and evaluate data sets. One important characteristic of these tools must be the efficient execution of the features. This article describes an experimental performance evaluation of the functionalities provided by the two popular data profiling open source tools, evaluating the features in the context of large data bases (Big Data).*

1. Introdução

Com o avanço da computação, uma imensa quantidade de dados é produzida diariamente pela humanidade. Analisar e resumir estes dados, assim como identificar padrões e informações genéricas pode representar, em muitos casos, uma tarefa não trivial. Nesse contexto, o perfilamento de dados consiste em examinar os dados de uma fonte de dados (arquivos, bancos de dados, etc.) e gerar metadados sobre os mesmos [Naumann 2014]. Em geral, é possível descobrir a estrutura dos dados, um resumo de seu conteúdo, possíveis relacionamentos e derivar regras de negócio sobre os mesmos. No contexto atual, o perfilamento de dados é desafiado por fatores como a grande quantidade de dados e o formato não estruturado dos mesmos. Ao lidar com bancos de dados extensos, variando de gigabytes até petabytes, é muito mais conveniente analisar dados por meio de sumários [Dorr e Murnane 2011]. Na prática, a geração destes sumários pode representar uma tarefa computacionalmente custosa. Desse modo, o tempo de execução de diferentes ferramentas de perfilamento de dados pode ser bastante influenciado pela quantidade de dados processada. Neste contexto, este artigo visa realizar uma avaliação experimental de duas ferramentas de perfilamento de dados de código aberto no contexto de grandes bases de dados.

Os resultados da avaliação de desempenho experimental reportados neste artigo podem ser bastante úteis para ajudar cientistas, analistas de dados e/ou usuários interessados em perfilar dados a escolherem ferramentas apropriadas para seus objetivos.

2. Planejamento dos Experimentos

De acordo com características como a abrangência das funcionalidades, a popularidade e o tempo de mercado, as seguintes ferramentas de perfilamento de dados foram selecionadas: *Talend Open Studio for Data Quality* (versão 5.4.2) e *Data Cleaner*

Community Edition (versão 3.6.2). O experimento consistiu na comparação do tempo de execução das funcionalidades presentes em ambas as ferramentas.

As funcionalidades em comum disponíveis em ambas as ferramentas selecionadas são: dispersão, cardinalidade, distribuição dos dados, tamanho dos dados, avaliação de formato e identificação de valores extremos. Desse modo, o design experimental consistiu em calcular o tempo médio de 30 (trinta) execuções de cada funcionalidade, utilizando para tal uma base de dados real (disponível em http://cdiac.ornl.gov/ftp/ndp026b/) contendo 10 milhões de registros, a qual contém informações sobre condições climáticas providas por navios e pontos de coleta terrestre.

3. Resultados

Na Figura 1 são apresentados os tempos médios de execução de cada uma das funcionalidades avaliadas nas ferramentas.

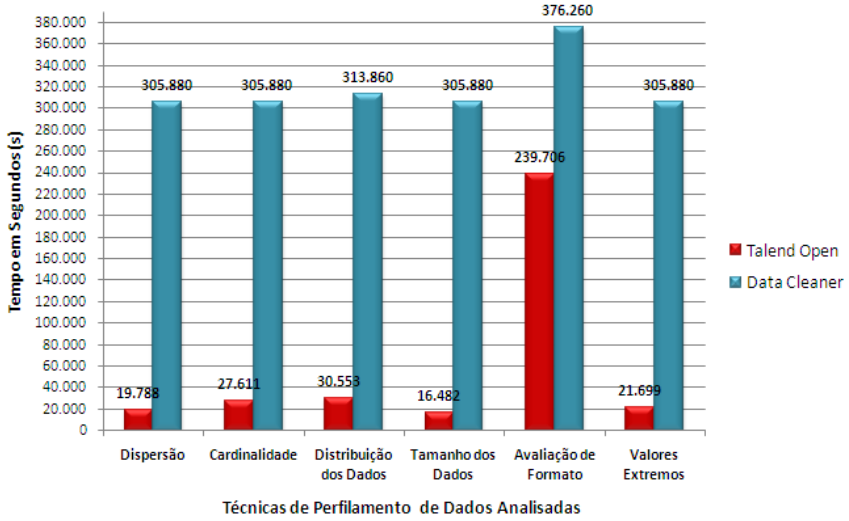

Figura 1. Tempo médio de 30 (trinta) execuções.

4. Conclusões e Trabalhos Futuros

Por meio da análise gráfica, percebe-se que a ferramenta *Talend Open Studio for Data Quality*, em relação às técnicas de perfilamento dados de dispersão, cardinalidade, distribuição dos dados, tamanho dos dados e identificação de valores extremos, gerou tempos de execução significativamente inferiores quando comparada à outra ferramenta. Por outro lado, observa-se que na técnica de perfilamento de avaliação de formato, ambas as ferramentas apresentaram tempos de execução elevados e similares.

Como trabalho futuro, pretende-se utilizar outras bases de dados na avaliação experimental, assim como realizar análises estatísticas sobre os dados dos experimentos e a inclusão de outras funcionalidades de perfilamento de dados no design experimental.

Referências

Naumann, F. (2014) “Data profiling revisited”, ACM SIGMOD Record, 42(4), p. 40-49.

Cormode, G., Garofalakis, M., Haas, P. J., and Jermaine, C. (2012) “Synopses for massive data: Samples, histograms, wavelets, sketches”, Foundations and Trends in Databases, 4(1–3), p. 1-294.

Dorr, B., and Murnane, R. (2011). “Using Data Profiling, Data Quality, and Data Monitoring to Improve Enterprise” Information. Software Quality Professional Magazine, 13(4).