Universidade Federal de Pernambuco
CIn - Centro de Informática
Graduação em Engenharia da Computação

Data profiling: Uma análise funcional de ferramentas gratuitas

Proposta de Trabalho de Graduação

Aluna: Sylvia Marcella Dubeux Ratis
Orientador: Robson do Nascimento Fidalgo
Área: Banco de Dados

Recife, abril de 2017

Sumário

Resumo

Sistemas de informação são cada vez mais críticos para dar suporte a decisões táticas, estratégicas e operacionais dentro de uma organização. A qualidade das decisões, no entanto, está diretamente relacionada à qualidade da informação fornecida para o gestor; esta, por sua vez, depende intrinsecamente da qualidade dos dados dentro do sistema. A geração de perfis de dados é uma maneira de demonstrar a existência de problemas nos dados que podem afetar o desempenho da organização. Este trabalho visa a realização de um estudo a respeito de *data profiling* e conceitos relacionados, bem como a identificação das operações mais comuns a serem realizadas sobre os dados na geração de um perfil de dados. Os conhecimentos adquiridos serão então utilizados na elaboração de uma análise comparativa de ferramentas comerciais gratuitas de geração de perfil de dados.

Contextualização

Atualmente, sistemas de informação são cada vez mais críticos para dar suporte a decisões táticas, estratégicas e operacionais dentro de uma organização. A qualidade das decisões, no entanto, está diretamente relacionada à qualidade da informação fornecida para o gestor; esta, por sua vez, depende intrinsecamente da qualidade dos dados dentro do sistema. Idealmente, os dados utilizados para tomada de decisão devem ser legíveis, compreensíveis, consistentes, relevantes e temporalmente oportunos. O gerenciamento da qualidade de dados torna-se mais importante à medida que cresce a dependência das organizações de tecnologias orientadas por dados. [1]

O processo de geração de perfil de dados, ou *data profiling*, consiste em examinar os dados disponíveis em uma determinada fonte e coletar informações a respeito deles, produzindo metadados cuja análise é um passo importante para gerenciar a qualidade dos dados da fonte. Um cenário típico seria a varredura das tabelas de um banco de dados relacional para obter informações como tipos de dados, padrões de valores, completude e unicidade de colunas, ou até mesmo dependências funcionais e regras de associação. Vale salientar, no entanto, que a geração de dados além das bases de dados relacionais tradicionais cresce em volumes nunca vistos antes, e esses dados também precisarão ser submetidos a *profiling*. [2]

Apesar da importância de manter a qualidade de dados de um sistema e do fato de existirem produtos voltados para isso há cerca de 20 anos, muitas empresas possuem pouca ou nenhuma iniciativa de governança ou gerenciamento de qualidade de dados, devido à visão de gestores que acreditam que se trata de um problema de TI e não de negócios, e portanto não estão dispostos a realizar investimentos nesse sentido. Nesses casos, a realização de *profiling* gratuito é uma maneira de demonstrar a existência de problemas nos dados que podem afetar o desempenho da organização. [3]

Objetivo

Este trabalho visa a realização de um estudo a respeito de *data profiling* e conceitos relacionados, bem como a identificação das operações mais comuns a serem realizadas sobre os dados na geração de um perfil de dados. Os conhecimentos adquiridos serão então utilizados na elaboração de uma análise comparativa de ferramentas comerciais gratuitas de *data profiling*. Tais ferramentas serão definidas de acordo com suas posições no Quadrante Mágico de Qualidade de Dados da Gartner para 2016 [4] e no relatório da Bloor de Ferramentas Gratuitas de *Profiling* [3].

Cronograma

	Março	Abril	Maio
Elaboração da proposta			
Definição do escopo			
Revisão da literatura			
Estudo e análise de ferramentas			
Elaboração do relatório			
Apresentação final			

Possíveis Avaliadores

- Fernando da Fonseca de Souza
- Ana Carolina Salgado
- Bernadette Faria Lóscio

Referências

[1] Singh, R., & Singh, K. (2010). A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing. IJCSI International Journal of Computer Science Issues, 7(3), 41–50.
[2] Naumann, F. (2014). Data profiling revisited. ACM SIGMOD Record, 42(4), 40–49.
[3] Howard, Philip (2014). “Free” Data Profiling Tools.
[4] Judah, Saul; Selvage, Mei Yang; Jain, Ankush (2016). Magic Quadrant for Data Quality Tools. Disponível em: <https://www.gartner.com/doc/reprints?id=1-3MOFN4I&ct=161128&st=sb>. Acesso em: 12 abr 2017.

_____

Sylvia Marcella Dubeux Ratis
(Orientanda)

_____

Robson do Nascimento Fidalgo
(Orientador)