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

Data Profiling: uma revisão exploratória

Proposta de Trabalho de Graduação

Aluna: Sylvia Marcella Dubeux Ratis

Orientador: Robson do Nascimento Fidalgo

Recife, setembro de 2016

Sumário

Contextualização

Objetivo

Cronograma

Possíveis Avaliadores

Referências

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 criação 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]

Objetivo

Este trabalho visa a elaboração de uma revisão exploratória do processo de data profiling, comparando o estado da arte na academia e as aplicações no mercado. Adicionalmente, será realizado um estudo comparativo de algumas ferramentas comerciais que realizam data profiling, a serem definidas de acordo com suas posições no quadrante mágico de ferramentas de qualidade de dados da Gartner para 2015. [3]

Cronograma

	Setembro	Outubro	Novembro	Dezembro
Elaboração da proposta				
Definição do escopo				
Revisão da literatura				
Estudo e análise de ferramentas				
Elaboração do relatório				
Apresentação final				

Possíveis Avaliadores

- Ana Carolina Salgado
- Bernadette Farias 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] Judah, S. & Friedman, T. (2015). Magic Quadrant for Data Quality Tools. Disponível em: <<https://www.gartner.com/doc/reprints?id=1-2T6O9CP&ct=151202&st=sb>>. Acesso em 16 set. 2016.

Sylvia Marcella Dubeux Ratis
(Orientanda)

Robson do Nascimento Fidalgo
(Orientador)