QualiDados: Jogo Didático para Consolidação de Conceitos de Qualidade de Dados*

Geovane F. Piccinin, Salatiel R. Santos, Suelen L. Romano, Mirella M. Moro

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais, Belo Horizonte, MG

{geovanepcc,salatiel.ribeiro,suelen.loiola,mirella}@dcc.ufmg.br

Abstract. *Data from real systems are often dirty, inconsistent, duplicated, inaccurate, incomplete or outdated. These problems generate incorrect results and analyzes, putting companies at risk of losing money, reputation and customers. The solution is then to count on data quality management concepts. This paper overviews such concepts and introduces a prototype for an educational game that simulates a data quality management environment.*

Resumo. *Sistemas reais frequentemente apresentam dados sujos, inconsistentes, duplicados, imprecisos, incompletos ou desatualizados. Tais problemas geram resultados e análises incorretos, criando risco de perda de receitas, credibilidade e clientes. Uma solução é utilizar conceitos de gerência de qualidade de dados. Este trabalho resume tais conceitos e introduz o protótipo de um jogo didático que simula um ambiente de gerência de qualidade de dados.*

1. Introdução

Sistemas reais frequentemente apresentam dados sujos, o que pode comprometer a confiabilidade dos dados e levar a erros. Estima-se que, em média, empresas possuem entre 1% e 5% de erro em seus dados e, em algumas delas, essa taxa pode chegar a 30%. Tais taxas elevadas podem prejudicar os sistemas de maneiras diferentes. Por exemplo, na maioria dos projetos de Data Warehouse, o processo de limpeza dos dados é responsável por 30% a 80% do tempo de desenvolvimento e do orçamento [Fan and Geerts 2012].

Dados sujos, especialmente os duplicados, custam à economia dos EUA aproximadamente três trilhões a cada ano. De acordo com *Larry English*, entre 15% e 20% do orçamento operacional de uma empresa pode ser desperdiçado devido a dados sujos. Isso demonstra a necessidade de melhorar métodos que visam detectar e tratar dados inconsistentes, incompletos, imprecisos, desatualizados e duplicados. Em resumo, dados sujos são responsáveis por prejuízos financeiros e redução da produtividade, devido ao tempo gasto na adequação desses dados.

Aprender a lidar com tais dados é fundamental para assegurar a qualidade dos mesmos. Com diversos conceitos ligados à qualidade de dados, o aprendizado dos mesmos bem como a correta utilização das ferramentas relacionadas são desafios para estudantes e profissionais. Para auxiliar em tal tarefa, este artigo apresenta o *QualiDados*, um jogo didático com situações de baixa qualidade de dados que podem ser corrigidas com a utilização de certas ferramentas. A ideia de utilizar jogos didáticos como auxílio

*Um vídeo de demonstração da ferramenta está disponível em http://sbbd2013.cin.ufpe.br/screencasts.

educacional não é nova. De fato, nós pensamos e entendemos melhor quando podemos imaginar um problema e nos preparar para respondê-lo [Gee 2003]. Nesse contexto, jogos são perfeitos para apresentar situações similares através de simulação, fornecendo então a oportunidade para pensar, entender, preparar e executar as devidas ações necessárias.

A detecção e o tratamento de dados sujos não são tarefas triviais devido ao caráter subjetivo. Neste artigo, são apresentados alguns conceitos básicos sobre qualidade de dados na Seção 2, e o caráter subjetivo de tais conceitos é discutido na Seção 3. Para facilitar o aprendizado e simular situações reais de controle de qualidade de dados, a Seção 4 introduz o *QualiDados*, protótipo de um jogo didático.

2. Conceitos Básicos sobre Qualidade de Dados

Esta seção resume alguns dos principais conceitos sobre qualidade de dados. Tais conceitos são explorados nas diversas características do jogo *QualiDados*.

Consistência de Dados. A consistência de dados se refere à validade e integridade da representação do dado no mundo real. Bancos de dados relacionais costumam apresentar inconsistência dentro de uma mesma tupla, entre diferentes tuplas em uma mesma relação e até mesmo entre diferentes tuplas de relações diferentes. Portanto não se trata apenas de restrição de integridade, é preciso levar em conta a semântica dos dados. Conforme o exemplo mostrado na Tabela 1, existem conflitos dentro de uma mesma tupla e entre tuplas diferentes. Na série de livros "Game of Thrones", homens da Patrulha da Noite abdicam da família, não podendo ser casados como informado na tupla $t1$. Igualmente, o símbolo da família Lannister é um leão, e não um lobo como informado na tupla $t2$, que também entra em atrito com o valor correto armazenado na tupla $t3$.

Tabela 1. Exemplo sobre consistência de dados

	nome	**sobren**	**símbolo**	**posição**	**civil**
t1	John	Snow	–	patrulha	casado
t2	Jaime	Lannister	lobo	escudeiro	solteiro
t3	Jaime	Lannister	leão	cavaleiro	solteiro
t4	Aria	Stark	lobo	–	solteira
t5	Aria	Stark	lobo	herdeira	solteira
t6	Aria	Bolton	lobo	herdeira	casada

Eliminação de Duplicatas. A eliminação de duplicatas identifica tuplas em uma ou mais relações que se referem à mesma entidade do mundo real. Por exemplo, considere as tuplas $t4$, $t5$ e $t6$ da Tabela 1: uma consulta para descobrir os dados da Aria retorna três tuplas. Porém, é possível que todas essas tuplas se refiram à mesma pessoa. Para descobrir se é o caso, uma alternativa é verificar se existe outro dado indicando que *Aria Stark* e *Aria Bolton* têm, por exemplo, os mesmos genitores, o que indicaria que são a mesma pessoa. Tal averiguação é fundamental para realizar não apenas a limpeza de dados mas também a integração de fontes de dados diferentes [Carvalho et al 2011]. Nesse caso, os dados são coletados a partir de diferentes bases com o intuito de reunir informações sobre uma mesma entidade. Como essa entidade pode estar representada de maneiras diferentes, ao integrar é necessário eliminar duplicatas também.

Precisão. A precisão dos dados se refere a quão próxima a representação no banco de dados está do valor da entidade no mundo real. Por exemplo, considere a Tabela 2 cujas tuplas especificam o nome, sobrenome, idade, peso e estado civil de uma pessoa, e $t1$ apresenta os dados verdadeiros da personagem Daenerys Targaryen, ao iniciar a série de livros. Pode-se concluir que $t2(idade, altura)$ é mais preciso do que $t3(idade, altura)$, já que tais valores estão mais próximos aos valores verdadeiros; enquanto $t3(nome, civil)$ é mais preciso do que $t2(nome, civil)$. Entretanto é mais desafiador determinar a precisão relativa de $t2$ e $t3$ quando *não* existe $t1$ como referência. Porém, mesmo sob essas circunstâncias, é possível descobrir que em certos atributos o valor em uma tupla é mais preciso do que em outra através da análise semântica dos dados e outras informações.

Tabela 2. Exemplo sobre precisão de dados

	nome	**sobren**	**idade**	**altura**	**civil**
t1	Daenerys	Targaryen	13	1,50	solteira
t2	D.	Targaryen	13	1,49	casada
t3	Daenerys	Targaryen	31	1,40	solteira

Completude da Informação. Um banco deve apresentar dados completas em resposta a uma consulta. Para uma base de dados D e uma consulta C, é necessário identificar quando C retornará um resultado correto usando-se apenas os dados fornecidos por D. Se D possui informação incompleta, é provável que retorne um resultado impreciso ou até mesmo incorreto [Fan and Geerts 2012].

Atualidade. Esse princípio identifica se as entidades apresentam valores atuais e retornam valores atuais para as consultas. Esse é um problema que, em teoria, poderia ser resolvido facilmente através da inserção de *timestamps* em todos os dados. No entanto, em bases de dados reais, os dados constantemente possuem *timestamps* imprecisas ou simplesmente não as possuem. Além disso é comum que dados sejam copiados ou importados de outras fontes, que nem sempre suportam o mesmo tipo de representação. Isso torna a tarefa de identificar os valores mais recentes mais difícil [Fan and Geerts 2012]. Por exemplo, para a Tabela 1, deseja-se descobrir o sobrenome e a posição de Aria. Após um processo de detecção de dados duplicados, descobriu-se que todas as três tuplas cujo nome é Aria representam a mesma pessoa. Se os dados não possuem *timestamp* fica difícil saber se o sobrenome atual de Aria é Stark ou Bolton, e se sua posição é nenhuma ou herdeira. Porém, mesmo sem um indício direto da atualidade do dado é possível descobrir qual é a tupla mais recente. A partir do contexto desses dados, é possível inferir que a posição atual de Mary é herdeira, pois normalmente as personagens passam de nada à herdeiras. De forma similar, pode-se deduzir que o sobrenome de Aria é Bolton, pois para esse valor o campo *civil* está definido como *casada*, ou seja, Aria é casada. Tal dedução considera o fato de que uma pessoa solteira pode mudar seu estado civil para casada, mas uma casada não muda seu estado para solteira, e sim para divorciada.

3. Qualidade de Dados: Aspectos Subjetivos

A seção anterior apresentou alguns conceitos relevante para analisar qualidade de dados. É importante notar que existem diversas interações entre tais conceitos. Por exemplo, a atualidade dos dados pode ser melhorada se mais informações temporais puderem ser

obtidas em um processo de melhoria da completude dos dados. Outro exemplo é que às vezes, para determinar qual dado é o mais atual, é preciso recorrer a um processo de identificação de duplicatas para saber quando várias tuplas se referem à mesma entidade. Tais interações, enfatizam o caráter *subjetivo* da qualidade de dados. Em muitos casos, a correção de uma informação depende de outros campos, outras tabelas e em alguns casos até mesmo da correção prévia de outros problemas.

Em várias situações, os bancos de dados tentam representar entidades muito complexas, onde cada entidade possui seus requisitos específicos de qualidade (precisão, representação, mecanismos de obtenção dos dados, etc). A construção do banco de dados envolve várias etapas, as quais podem ser críticas para a determinação da qualidade final dos dados. Essas etapas vão desde a coleta (por exemplo em pesquisas de campo, medições com instrumentos mecânicos ou digitais) até a apresentação final dos dados.

Nesse contexto, problemas de qualidade dos dados podem surgir em qualquer etapa. Exemplos de etapas e seus problemas incluem os seguintes:

- *Coleta dos dados* (amostragem) com falha dos instrumentos, erros de operação, dados sem qualidade;
- *Processamento dos dados amostrados* com procedimentos imprecisos, erros de processamento, falhas de processamento;
- *Representação* com a representação proposta distorcendo a realidade;
- *Apresentação* com a visualização gerada compromentendo a avaliação do usuário.

Portanto, não há uma definição genérica sobre qualidade de dados que norteie como o tratamento deve ser realizado. Há a certeza de que cada aplicação deve receber uma avaliação particular, e os métodos devem ser adequados ao objetivo dos usuários, a cada etapa de constituição da base de dados.

4. *QualiDados*: Definição e Características

Cabe ao analista de banco de dados definir processos e ferramentas adequados para manter os dados com a precisão, consistência e atualidade necessárias para o bom funcionamento do sistema. Conforme discutido nas seções anteriores, o caráter subjetivo da qualidade de dados e da necessidade de tomar decisões de acordo com o banco de dados e as características do sistema, é muito difícil que estudantes e profissionais consigam praticar tais conceitos sem auxílio devido.

Por outro lado, as teorias que unem educação ao entretenimento têm ganhado muita força nos últimos anos, apresentando uma tendência para um modelo de educação que emerge: a educação apoiada no uso de jogos digitais. De fato, projetistas de jogos e educadores argumentam que jogos capturam a atenção de seus jogadores, engajando-os em raciocínio e resolução de problemas complexos [Barab and Dede 2007]. Nesse contexto, o nosso objetivo é apresentar o protótipo de um jogo didático que simula diversas situações para as quais o jogador terá de avaliar a qualidade dos dados.

O propósito do jogo é fornecer um ambiente didático de competição onde os estudantes (e profissionais) tenham a oportunidade de praticar conceitos aprendidos na sala de aula ou na prática. O ambiente de competição é estimulante e incentiva os alunos a solidificar seu aprendizado e a buscar soluções para os problemas apresentados [Reis et al. 2012]. Além disso, jogos educacionais, incluindo os digitais estimulam o

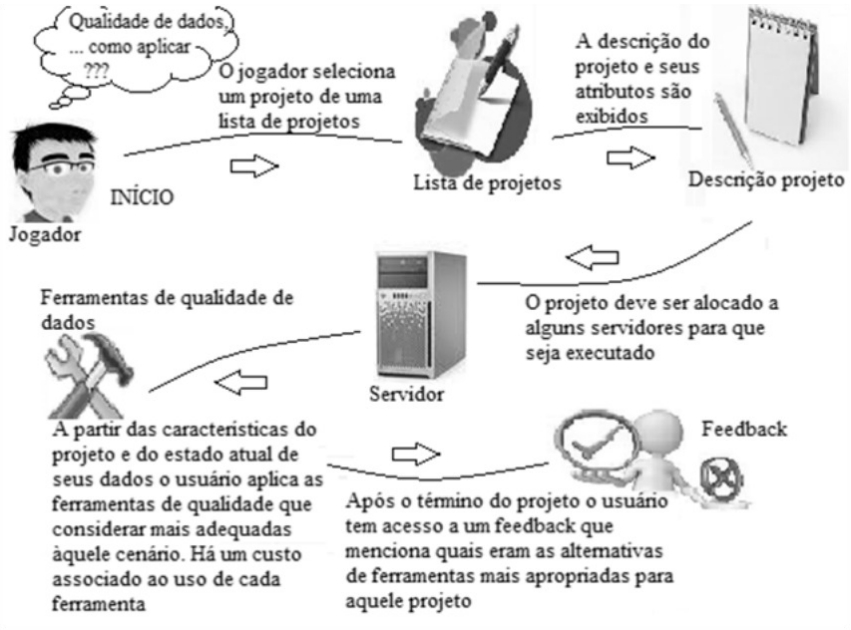

Figura 1. Fluxo do jogo didático QualiDados

desenvolvimento cognitivo, auxiliando na criação de estratégias para a solução de problemas [Hopf et al. 2007].

O jogo *QualiDados* simula uma empresa de TI na qual o jogador é o "analista de banco de dados". O jogador tem à sua disposição: *servidores*, os quais são alocados a uma lista de *projetos*, e *ferramentas* de qualidade de dados. Cada Projeto possui uma descrição com seus objetivos e atributos que revelam a qualidade de dados atual e a desejada. O jogador deve, então, selecionar um projeto, alocar servidores para a sua realização e aplicar as ferramentas de qualidade de dados para que atinja os objetivos pretendidos e ganhe os pontos. Dentre a lista de Ferramentas de qualidade de dados estão: backup, comparar com base externa, completar dados, criptografar, eliminar dados duplicados, inserir meta-dados, limpar dados sujos, notificar fonte de dados. Nessa perspectiva, o jogador é desafiado a escolher e aplicar as ferramentas de qualidade de dados mais adequadas ao seu projeto. Além de gerenciar os recursos para que otimize a quantidade de projetos realizados.

A Figura 1 ilustra o funcionamento do jogo, conforme explicado a seguir.

1. Início do jogo. Elementos disponibilizados ao jogador: lista de Servidores, lista de Projetos, Ferramentas de qualidade de dados, Saldo inicial;
2. Andamento do jogo. O jogador seleciona um Projeto para executar e aloca Servidores para realizarem o processamento do Projeto. Nessa etapa, o jogador verifica quais são as características e demandas do Projeto e aplica as Ferramentas de qualidade de dados que considerar adequadas.
3. Processamento dos pontos (ao final de uma rodada). Após alocar os Servidores e aplicar as Ferramentas de qualidade de dados sobre o Projeto é feita a contabilização dos pontos obtidos avaliando-se quais as Ferramentas aplicadas frente às características do Projeto.
4. Feedback. Após a conclusão de um Projeto torna-se disponível ao usuário um

relatório explicando quais eram as Ferramentas adequadas àquele Projeto. Esse relatório é fundamental no jogo porque ele será o gabarito do jogador para medir e qualificar sua atuação.

5. Ações que podem ser realizadas a qualquer momento do jogo. O jogador pode comprar novos Servidores para aumentar sua capacidade de processamento. Acessar o tutorial.
6. Tutorial. Acompanha o jogo com o objetivo de elucidar os princípios, fundamentos, regras e objetivos.
7. Editor de Projetos. A lista de Projetos disponibilizada no jogo é criada com um programa auxiliar e armazenada em um arquivo, o qual é carregado no início do jogo. O Programa auxiliar consiste em um formulário onde o usuário define todos os atributos da entidade Projeto: Título, Descrição, Valor do Pagamento, peso que será aplicado a cada quesito de qualidade de dados, Quantidade de dados a serem processados, etc. Com essa ferramenta pode-se criar qualquer instância de Projeto.

5. Conclusões

O objetivo do *QualiDados* é ser uma opção didática para interessados no tema qualidade de dados. Sua base está na simulação de um ambiente para praticar e desenvolver o senso crítico a respeito dos principais conceitos e ferramentas de qualidade de dados. O protótipo pode ser estendido para que permita a interação entre os jogadores e que ofereça mais desafios de gerenciamento, típicos de um ambiente informatizado, como problemas de segurança dos dados e problemas de hardware. Adquirindo essas novas características e funcionalidades, ele pode se tornar um recurso de aprendizagem mais robusto e que estimule, através da competitividade, o aperfeiçoamento teórico dos jogadores.

Agradecimentos. Projeto parcialmente financiado por CNPq.

Referências

Barab, S. and Dede, C. (2007). Games and Immersive Participatory Simulations for Science Education: An Emerging Type of Curricula. *Journal of Science Education and Technology*, 16(1):1–3.

Carvalho et al, A. P. (2011). Incremental Unsupervised Name Disambiguation in Cleaned Digital Libraries. *JIDM*, 2(3):289–304.

Fan, W. and Geerts, F. (2012). *Foundations of Data Quality Management*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers.

Gee, J. P. (2003). What video games have to teach us about learning and literacy. *Computers in Entertainment*, 1(1):20.

Hopf, T., Falkembach, G. A. M., and Araújo, F. V. (2007). Incremental Unsupervised Name Disambiguation in Cleaned Digital Libraries. *Novas Tecnologias na Educação*, 5(2):289–304.

Reis, G. L., Souza, L. F. F., Carvalho, F. C. T., Abdalla Jr., M. A., Nepomuceno, E. G., Barroso, M. F. S., and Pereira, E. B. (2012). As Competições Universitárias e a Carreira Profissional do Aluno de Graduação: Um Estudo de Caso Sobre a Equipe UAIrobots-SEK. In *Workshop de Robótica Educacional*.