3º CONTECSI May 31 to June 02, 2006 São Paulo, Brazil

3rd International Conference on Information Systems and Technology Management
3º Congresso Internacional de Gestão da Tecnologia e Sistemas de Informação
11th World Continuous Auditing Conference

De 31 de Maio a 02 de Junho de 2006 – São Paulo/SP – Brasil

QUALIDADE DE DADOS PARA DATA WAREHOUSE – ROTEIRO DE IMPLEMENTAÇÃO

Mozart da Silva Britto
(Escola Politécnica da USP, São Paulo, Brasil) mozartsbritto@yahoo.com.br

Jorge Rady de Almeida Júnior
(Escola Politécnica da USP, São Paulo, Brasil) jorge.almeida@poli.usp.br

ABSTRACT

Despite the general consensus of the importance of data quality, many organizations are not aware that the lack of data quality is one of the main sources of loss of time, money and opportunities in the business environment. Even in organizations with this level of conscience, in many of them there is not a consolidated and applicable methodology in order to guarantee a continuous improvement in the data quality. The failure in many business iniciatives has been caused for wrong strategical decisions, based in low quality organizational data.
Many organizations incorporate, on their daily routine, reactive activities like correction and order reprocessing, and the unavoidable result is the waste of resources.
The objective of this article is to present the several aspects involved in data quality and show an implementation methodology for continuous improvement of data quality, mainly for data warehouses. In this article, are showed the impacts suffered by organizations caused by low quality data. and an introduction to the concept of data quality. Subsequently, is presented an itinerary step-by-step of how to implement and maintain a continuous data quality improvement program for data warehouses. This itinerary contains technical and political issues to implement the change, and is also useful for data quality improvement programs in transactional databases. Afterwards, is presented an practical case of data quality program improvement in a large organization, showing is targets, challenges and observed results. Finally is presented a conclusion and recommendations for a data quality implementation program.

Key-Words: Data Quality, Data Warehouse

RESUMO

Apesar do consenso geral da importância da qualidade de dados, em muitas organizações ainda não se percebe que a falta de qualidade de dados é um dos principais causadores de perda de tempo, dinheiro e oportunidades dentro dos negócios. Mesmo dentre as organizações que têm essa cultura, em muitas delas ainda não existe uma metodologia consolidada e aplicada para garantir a melhoria contínua da qualidade de dados. O insucesso de muitas investidas empresariais tem sido causado por decisões estratégicas erradas, baseadas em dados organizacionais de baixa qualidade. Muitas organizações incorporam, em sua rotina diária, atividades reativas como correção e reprocessamento de pedidos, cujo resultado inevitável é o desperdício de recursos.
Este artigo tem por objetivo apresentar os vários aspectos envolvidos na qualidade de dados e mostrar uma metodologia de implementação para melhoria contínua da qualidade, sobretudo para data warehouses. Neste artigo, são mostrados os impactos sofridos pelas organizações devido à má qualidade de dados e uma introdução ao conceito de qualidade de dados. É apresentado um roteiro passo a passo de como implementar e manter um programa de melhoria contínua de qualidade de dados para data warehouses. Este roteiro contém questões técnicas e políticas para implementar a mudança, e também é útil para programas de melhoria de qualidade em bancos de dados transacionais. Após isso, é apresentado um caso prático de aplicação de programa de melhoria de qualidade de dados em uma grande organização, apresentando suas metas, desafios e resultados observados. Finalmente, é apresentada uma conclusão e recomendações para implementação do programa de qualidade de dados.

Palavras-Chave: Qualidade de Dados, Data Warehouse

Agradecimentos: Sobretudo a Deus pela vida, pelo amor incondicional e pela inspiração; ao meu orientador Jorge Rady, por acreditar em meu potencial e trabalho, guiando-me no caminho a ser seguido; à minha esposa Ellen por seu carinho, apoio e paciência, e aos meus pais Raymar e Roseli pelo incentivo constante.

1. Introdução

1.1. Ambiente corporativo atual

No mundo organizacional da atualidade, observa-se que o maior diferencial de competitividade está sendo a capacidade de gerar conhecimento e produzir inovação. Assim, a informação está recebendo cada vez mais foco, sendo considerada como o recurso estratégico mais valioso. Neste cenário, os dados armazenados em data warehouses estão ganhando cada vez mais espaço para auxílio à tomada de decisões estratégicas organizacionais.
Os dados armazenados devem apresentar visões do mundo real, e se essas visões não corresponderem substancialmente com esse mundo real por algum período de tempo razoável, então a organização pode começar a agir de forma desordenada e sem coordenação.

1.2. Casos clássicos de falta de qualidade de dados

Há dois casos clássicos recentes na história norte-americana sobre falta de qualidade de dados e seus impactos, sobretudo financeiros.
O primeiro caso ocorreu em setembro de 1999, quando a sonda Climate Orbiter enviada até Marte pela NASA explodiu depois que seus jatos propulsores se incendiaram na entrada da órbita do planeta. Após as investigações, concluiu-se que os engenheiros da NASA cometeram um erro de conversão de unidades de medidas de força, mas que foi suficiente para a sonda entrar incorretamente na órbita de Marte. O impacto foi a perda de 125 milhões de dólares americanos e um atraso considerável na capacidade da NASA na exploração de Marte (Loshin, 2001: 3).
O segundo caso ocorreu no ano seguinte, em 2000, durante as eleições americanas para presidente (Loshin, 2001: Prefácio). Faltava apenas apurar o resultado da Flórida, que apontaria o vencedor da eleição. Entretanto, houve atrasos consideráveis na divulgação de resultados, e muitos dos resultados apresentados tinham características contraditórias.
Após seis semanas de atrasos e muitos desencontros, o presidente George Bush foi eleito num clima de considerável incerteza sobre a legitimidade do processo. A falta de legitimidade de um presidente representa um impacto político enorme, uma vez que pode lhe faltar apoio para aprovação de seus principais projetos de governo, gerando impacto financeiro (sobretudo nas Bolsas de Valores) na maior economia do planeta. As principais causas apontadas do problema foram:

- Anomalias em várias urnas eletrônicas na Flórida
- Uso de mecanismo de recontagem de votos
- Apresentação confusa das informações na cédula manual
- Falta de mecanismo de confirmação da opção para o eleitor
- Fontes de informação de resultados sem sincronismo, atrasadas e conflitantes

Diante disso, fica claro como podem ser catastróficos os impactos (sobretudo financeiros) da falta de qualidade de dados. Em pesquisa recente realizada pela empresa de auditoria PriceWaterhouseCoopers com 599 companhias foi levantado que os Estados Unidos perdem anualmente por volta de 1,4 bilhão de dólares americanos devido a problemas de qualidade de dados (Olson, 2003: 9).

1.3. Conceito de qualidade de dados

Mas o que vem a ser realmente o conceito de qualidade de dados ? Por qualidade de dados entende-se o grau de aderência entre as visões apresentadas pelos dados armazenados e os mesmos dados no mundo real. Um nível de qualidade de dados de 100% indicaria perfeita aderência com o mundo real, enquanto que um nível de 0% de qualidade de dados indicaria total falta de aderência. Nenhum sistema de grande porte tem qualidade de dados de 100%. A grande questão não é garantir que a qualidade de dados seja perfeita, mas sim que seja suficientemente precisa, atualizada e consistente para que a organização possa sobreviver e tomar decisões razoáveis. (Orr, 1998: 2).

A qualidade de dados também pode ser definida como "grau de adequação ao uso", o que implica que o conceito de qualidade dos dados também tem caráter relativo. Assim, dados considerados de qualidade apropriada para um uso específico podem não ter qualidade suficiente para outro uso (Tayi; Ballou, 1998: 1).

As situações de mudança têm se mostrado dramáticas no tocante à qualidade de dados. Os dados armazenados são estáticos mas o mundo real está em constante mudança. Mesmo dados armazenados com 100% de conformidade com o mundo real num instante inicial, após um curto espaço de tempo apresentam ligeira defasagem, e após mais um tempo a defasagem será muito maior. É necessária uma constante realimentação para que um sistema possa refletir continuamente o mundo real. (Orr, 1998: 2).

O controle da qualidade em bases de dados envolve todas as etapas do tratamento da informação, desde a sua criação até o uso final.

1.4. Dimensões da qualidade de dados

A qualidade de dados pode ser medida através de algumas dimensões: acuracidade, disponibilidade em tempo, relevância, completude, compreensibilidade e confiabilidade (Olson, 2003: 24), a seguir detalhadas.

- Acuracidade: é a medida de quão próximo o dado está do dado real. É a principal dimensão de qualidade de um dado, e se esta dimensão não atende a um nível satisfatório, o pleno atendimento a todas as demais dimensões não é significativo. Para um dado acurado, espera-se que o mesmo seja representado de forma consistente e não ambígua.
- Disponibilidade em tempo: implica em quão atualizado o dado encontra-se no momento em que é necessário para utilização
- Relevância: diz respeito a quão útil o dado é no contexto em que se encontra inserido
- Completude: indica o quão de acordo com amplitude e profundidade esperadas o dado se encontra
- Compreensibilidade: indica o quão fácil é entender o dado, e o quão além o dado vai da sua real necessidade
- Confiabilidade: indica o quão corretas estão as fontes de dados, bem como as transformações pelas quais o dado sofreu

Estas são características observáveis e passíveis de serem medidas e avaliadas. Estas características são as mais disseminadas na literatura, embora outras classificações possam ser encontradas, como a encontrada em (Wang, 1998: 3), que traz 15 dimensões para a qualidade de dados.

1.5. Fontes de problemas de qualidade de dados

Segundo (Olson, 2003: 43), as principais fontes de problemas são:

- Entrada inicial de dados com problema: erros devido à pressa, falta de atenção ou intencionais
- Degradação: dados iniciais corretos que não foram atualizados
- Mudanças e reestruturação: problemas na migração dos dados para outro ambiente. Erros podem ocorrer nas fases de extração, limpeza, transformação, carga e integração. Campos que contêm múltiplos dados armazenados também costumam trazer problemas.
- Utilização dos dados: quando os dados são inseridos num contexto incorreto ou ambíguo, sua interpretação ou uso podem ser prejudicados

1.6. Impactos da baixa qualidade de dados

Segundo (Redman, 1998: 2), os impactos podem ser classificados em impactos operacionais, impactos táticos (média gerência) e impactos estratégicos (direção da empresa).

- Impactos operacionais: aumento no custo operacional com baixa satisfação de clientes e empregados;
- Impactos táticos: dificuldade para implementar *data warehouses*, decisões demoradas e sem grande fundamentação, desconfiança da capacidade gerencial;
- Impactos estratégicos: dificuldades em criar uma estratégia, implementá-la e efetuar seu alinhamento com toda a organização.

De acordo com (Olson, 2003: 12) a baixa qualidade de dados gera diversos custos:

- Custo para reprocessar transação: engloba custos de mão de obra de todos os envolvidos na transação, custos de multas de atrasos, custos de transporte. Estes custos podem ser minimizados com a padronização de procedimentos alinhados com sistemas com interface clara, validação de dados e realimentação imediata
- Custo de implementação de novos sistemas: custos adicionais de mão de obra especializada e custos adicionais de licença de sistemas antigos
- Custo de adaptação de dados para entrega a tomadores de decisão: custos de mão de obra para agir manualmente nos dados, bem como custos decorrentes das decisões sub-otimizadas ou incorretas baseadas em dados atrasados e pouco confiáveis
- Custo de perdas de clientes: custos adicionais para ações de marketing para promover produtos e serviços, bem como custos para recuperar a imagem da marca
- Custos de produção: erros nos materiais ou quantidades entregues paralisam o sistema de Supply Chain, afetando a linha de produção. Podem haver custos de estocagem no caso de produção excedente.

1.7. Relacionamento da qualidade dos dados com seu uso

De acordo com (Orr, 1998: 3), a qualidade dos dados de tem forte ligação com a freqüência de seu uso. Algumas regras gerais de qualidade de dados podem ser destacadas:

- Dados pouco usados tendem a se deteriorar
- A freqüência de uso dos dados influi mais na qualidade dos dados que do tipo de dado armazenado
- Os dados mais utilizados em um sistema tendem a ser os mais atualizados
- O passar do tempo aumenta os problemas de qualidade de dados
- Os dados que se supunham permanecer inalterados são difíceis de se tratar quando uma mudança é requerida
- Os metadados (dados sobre dados) também tendem a se comportar conforme as leis de qualidade de dados

Muitos dados são coletados sem real necessidade, simplesmente para minimizar alterações futuras no sistema. Entretanto, estes dados se deterioram rapidamente porque ninguém verifica a sua acuracidade. Na biologia, este fenômeno recebe o nome de atrofia: se uma parte de um organismo não é utilizada, ela se deteriora. Assim também acontece com lista de nomes e endereços de uma mala direta se ninguém atualiza a lista com base nas correspondências que voltam ou com base nos contatos feitos de forma direta.

A falta de uso também afeta a qualidade dos metadados. As pessoas responsáveis pela entrada dos dados descobrem quais campos não são consistidos devidamente e deixam de atualizá-los corretamente ou utilizam os campos para outros propósitos. Assim, a definição dos dados deixa de corresponder ao seu conteúdo.

A proposta de (Orr, 1998: 6) para minimizar estes problemas é incentivar as pessoas a utilizar mais os dados, sobretudo para as pessoas que têm mais conhecimento sobre os mesmos.

Os programas de fidelidade das empresas aéreas são um excelente exemplo de manutenção de qualidade de dados: para obter o benefício das milhas, os próprios clientes se preocupam em manter atualizados os seus cadastros junto à empresa aérea.

1.8. Desafios específicos de qualidade de dados para data warehouse

Uma baixa qualidade de dados compromete totalmente uma implementação de data warehouse. Segundo (Redman, 2001: 31), são vários os desafios para se implementar um data warehouse de alta qualidade:

- Os clientes de data warehouse são diferentes dos clientes de sistemas operacionais, daí é natural que os requisitos de qualidade de dados sejam diferentes.
- As cadeias de processo de tomada de decisão costumam ser mais dificilmente entendidas que as cadeias de processo operacionais, o que dificulta a compreensão das necessidades do cliente.
- Em sistemas transacionais, os dados novos costumam ser bem mais importantes. Em data warehouses, os dados antigos também tem grande importância, então as necessidades do usuário envolvem também questões de histórico de informação.
- Grande parte dos dados de data warehouse provém de sistemas transacionais. Como os data warehouses exigem padronização e formatação de dados, a obtenção de dados em sistemas transacionais sem padronização definida pode esbarrar em complicadas questões políticas.

2. Metodologia

2.1 Roteiro para Qualidade de Dados com foco em Data Warehouse

Foi visto que um nível adequado de qualidade de dados é vital na minimização de perdas financeiras e de tempo (operacional e de decisão) em uma organização. Como os ambientes de Data Warehouse das empresas direcionam os executivos da organização na tomada de decisões estratégicas de longo prazo, a qualidade de dados destes ambientes é fator crucial para a sobrevivência da corporação.
Devido a isso, é apresentado a seguir um roteiro (Loshin, 2001: 463) de como implementar e manter um programa de qualidade de dados para o Data Warehouse da empresa. Devido à sua criticidade, a qualidade de dados deve ser pensada desde o nascimento do Data Warehouse. O roteiro apresentado a seguir também pode ser utilizado na implementação de um programa de qualidade de dados em bancos de dados transacionais já existentes. A figura 1 apresenta todas as atividades desse roteiro, em sua ordem de realização.

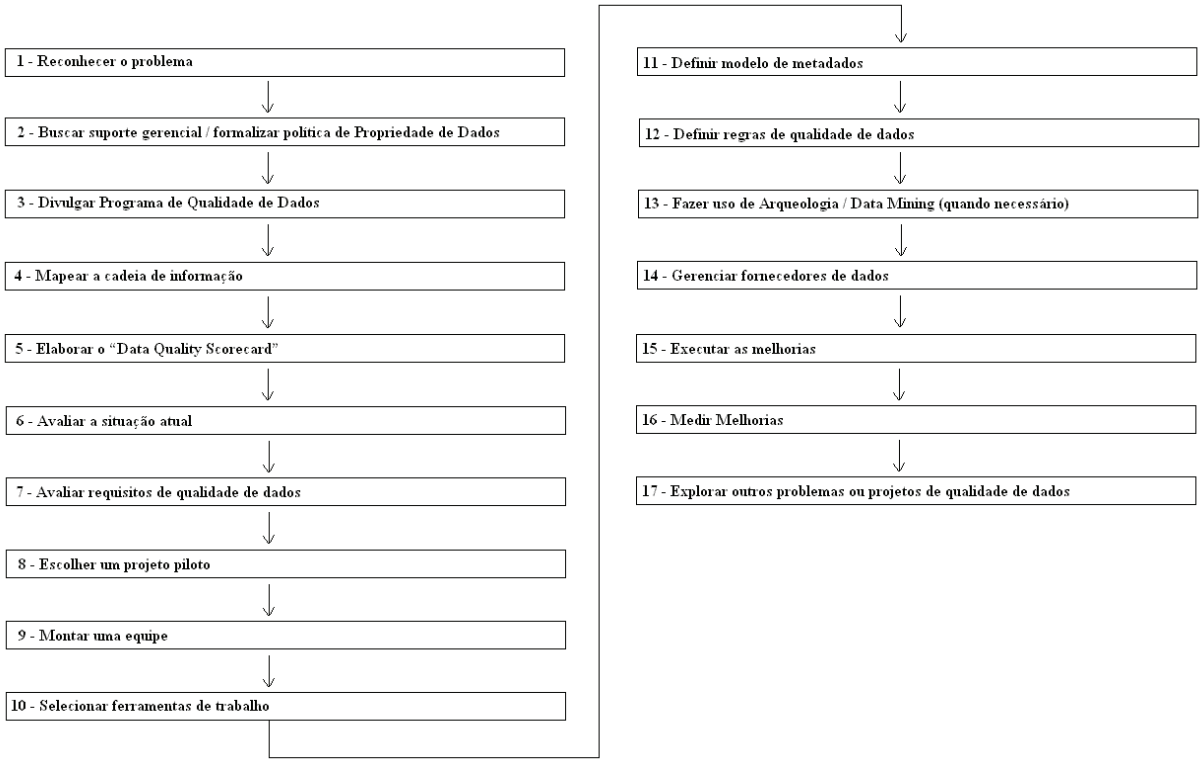

Figura 1 – Roteiro para Qualidade de Dados com Foco em Data Warehouse

2.1.1. Reconhecer o problema

Em primeiro lugar, a organização deve reconhecer que seu nível de qualidade de dados carece de melhorias, ou pelo menos, de medição formal. Dentre as evidências de má qualidade dos dados, pode-se citar:

- Freqüentes falhas e interrupções no sistema da empresa
- Queda na produtividade quando se aumenta o volume de dados a serem processados
- Grande giro de empregados devido à falta de confiança na organização
- Aumento dos requisitos de serviço do cliente, demonstrando que o cliente tem um nível de exigência e expectativa acima da demonstrada pela empresa

- Atrito com clientes, seja por atrasos, entregas incorretas ou má informação

Além disso, deve-se buscar identificar de forma geral quais são os maiores problemas de qualidade de dados da empresa. Estas informações podem ser obtidas através de conversas formais ou informais com clientes e empregados, ouvindo suas queixas e casos vividos.

2.1.2. Buscar suporte gerencial e formalizar política de Propriedade de Dados

Para melhoria da qualidade, é vital buscar apoio dos gerentes seniores da organização (Loshin, 2001: 464). Como estratégia de convencimento, é importante apresentar os impactos (sobretudo financeiros) sofridos pela organização devido à má qualidade dos dados. Após convencer a gerência, deve-se elaborar e apresentar um plano sumarizado de melhoria de qualidade de dados, demonstrando o retorno financeiro que o projeto traria à organização. Um relatório do tipo ROI (Return on Investiment) mostra-se bastante adequado.
Após esta etapa, deve-se atribuir papéis e responsabilidades aos chamados "proprietários de dados" da organização. Estas pessoas ou grupos terão as seguintes responsabilidades: definição de dados, autorização de acesso e segurança, suporte aos usuários de dados, gerenciamento de regras de negócio, metadados, padrões e fornecedores.
Um proprietário de dados não é obrigado a assumir todas as responsabilidades acima para um conjunto de dados, porém todas as responsabilidades devem estar atribuídas a algum proprietário de dados. Estas responsabilidades devem ser formalizadas em um documento formal assinado de papéis e responsabilidades, que deve conter as seguintes informações: gerentes apoiadores da política, conjuntos de dados em questão, políticas de resolução de disputas, papéis e responsabilidades.
Os passos para estabelecimento de uma política de propriedade de dados podem ser os seguintes: identificar partes interessadas nos dados, catalogar conjuntos de dados, determinar modelos de propriedade dos dados, definir papéis e responsabilidades de cada proprietário de dados e manter registro atualizado da política de propriedade.

2.1.3. Divulgar Programa de Qualidade de Dados

Após obter apoio dos gerentes sêniores e definir os proprietários dos dados, deve-se comunicar o plano de qualidade de dados aos demais membros da organização (Loshin, 2001: 467). Para atrair o interesse dos ouvintes, convém primeiro mostrar questões de negócio envolvidas e depois apresentar aspectos de implementação. Recomenda-se elaborar um treinamento para os membros da organização contendo o seguinte escopo:

- Impacto econômico de qualidade de dadosCriação, uso e armazenamento da informaçãoPropriedade de dadosConceitos de qualidade e ciclo de melhoria qualidadeDimensões da qualidade, domínios e regras de negócioMétricas para medição e análiseLimpeza e padronização de dadosDetecção de erros, correção, análise da causa

2.1.4. Mapear a cadeia de informação

Após toda organização estar ciente não só da importância da qualidade de dados mas principalmente do que significa qualidade de dados e como medi-la, tem-se como próximo passo mapear a cadeia de informação da empresa em nível macro (Loshin, 2001: 468).

Em um fluxo de dados operacional típico, onde pode-se identificar os seguintes estágios de processamento: suprimento, aquisição, criação de dados, processamento, empacotamento, entrega de dados e seu consumo

Na figura 2, é apresentado um fluxo de dados estratégico típico. A diferença básica é que há estágios de decisão e implementação de decisão antes dos estágios de entrega de dados e de consumo de dados.

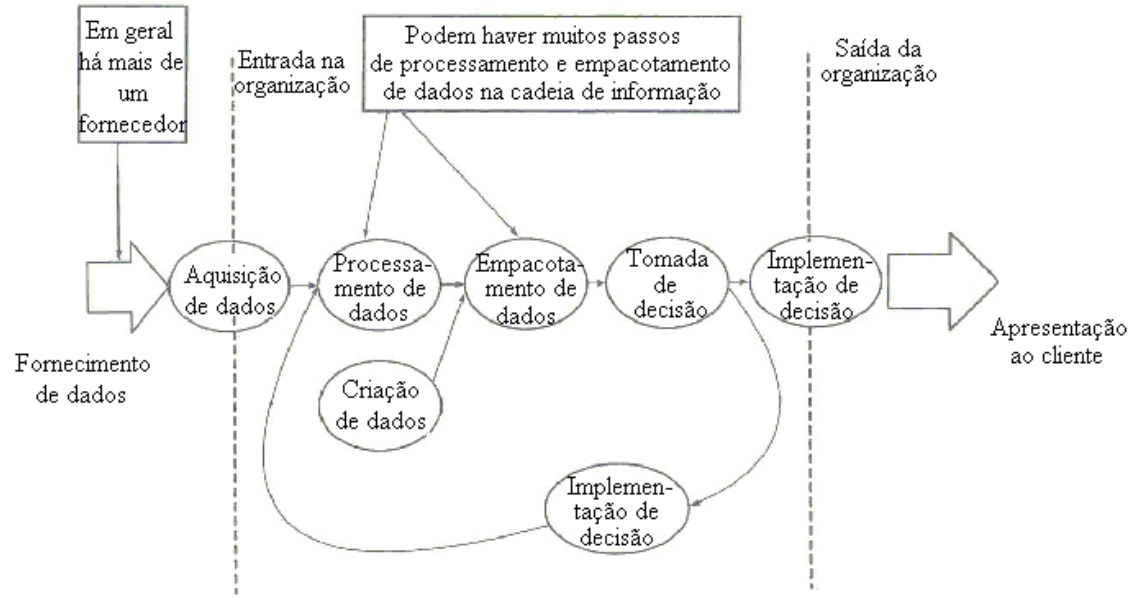

Figura 2 – Fluxo de dados estratégico típico (LOSHIN, 2001: 79)

Após a identificação dos estágios de processamento, deve-se determinar a conectividade entre os estágios, ou seja, quais estágios estão interligados, e também em que direção a informação flui. Após tudo isso, deve-se mapear a cadeia de informação por completo, tendo uma visão macro de todo o fluxo de informação da organização.

2.1.5. Elaborar o "Data Quality Scorecard"

O Data Quality Scorecard consiste numa ferramenta de auxílio para identificar as melhores oportunidades de melhoria (Loshin, 2001: 470). Ele é baseado no custo total de cada estágio da cadeia de informação devido à má qualidade de dados. Para construir o Data Quality Scorecard, os passos são os seguintes:

- Analisar a cadeia de informação
- Entrevistar empregados e clientes
- Isolar falhas de dados por local na cadeia de informação
- Associar o domínio do impacto com cada problema de qualidade de dados
- Caracterizar custos econômicos com cada impacto
- Agregar totais para determinar o impacto econômico real

Através da análise no Data Quality Scorecard, pode-se então escolher os pontos na cadeia de informação cuja melhoria irá trazer maior impacto econômico à organização. Em geral, os locais mais sujeitos a problemas de qualidade na cadeia de informação são: junções críticas, coletores, dispersadores, pontos de acesso aos dados e pontos de alto consumo de recursos (Loshin, 2001: 472).

Por junção crítica entende-se qualquer estágio de processamento com um alto grau de carga de informação onde admite-se que sejam manipuladas informações provenientes de diferentes fontes de dados. Por coletor entende-se que é um local onde a informação é agregada e preparada para armazenamento ou para geração de relatórios. Por dispersador entende-se que é um estágio de processamento que prepara a informação para muitos consumidores. Para pontos de acesso aos dados, a relevância diz respeito à facilidade de acesso aos dados. Por ponto de alto consumo de recursos, entende-se que é um estágio de processamento que consome uma alta percentagem dos recursos da organização que provavelmente fornece informações úteis para o processo de medição.

2.1.6. Avaliar a situação atual

Uma vez identificado o local da cadeia de informação que receberá o foco de melhoria, deve-se escolher sob quais dimensões da qualidade de dados o local da cadeia será analisado e melhorado (Loshin, 2001: 472). Após a melhoria ser realizada, ela deverá permanecer em constante medição.

2.1.7. Avaliar requisitos de qualidade de dados

No início desta etapa, serão combinados os resultados do Data Quality Scorecard e da avaliação da situação atual para descobrir-se o percentual de custo e impacto que cada problema representa para a organização (Loshin, 2001: 473).
Neste estágio, tem-se uma idéia melhor do escopo e magnitude dos problemas a serem enfrentados, e o escopo de cada problema deverá então se adequar em uma unidade de tarefa gerenciável. Cada problema então deverá ser priorizado (em geral de acordo com o percentual de custo e impacto) e deverão ser definidos os papéis e responsabilidades relativos a cada problema.
Após isto, serão então definidos e refinados os requisitos de cada problema. Cada especificação de um requisito de qualidade de dados deverá conter o seguinte: único identificador do requisito, nome do requisito, nome da parte responsável, local na cadeia de informação onde o requisito é aplicável, referência para a dimensão de qualidade de dados que está sendo medida, descrição do método de medição, regra de medição (se possível), mínimo requisito para entrada do dado no sistema, requisito ideal para entrada do dado no sistema e fator de escala como percentagem da qualidade de dados total do sistema.

2.1.8. Escolher um projeto piloto

Não se pode esperar que todos os problemas de qualidade sejam resolvidos de uma vez só. Nesta etapa será definido o escopo do projeto, ou seja, quais problemas serão atacados pelo projeto (Loshin, 2001: 474). A seguir, listam-se algumas sugestões para selecionar um projeto piloto.

- Escolher problemas cuja resolução trará grande impacto (sobretudo financeiro, seja por aumento de receita ou diminuição de custo)
- Para os problemas selecionados, garantir que as questões políticas estão sobre controle
- Certificar-se que o apoio dos gerentes sêniores está garantido
- Escolher problemas factíveis de serem resolvidos por sua futura equipe

2.1.9. Montar uma equipe

Sugere-se que a equipe seja composta pelos seguintes profissionais (Loshin, 2001: 475):

- Gerente de projeto: define conjunto de tarefas, cria cronogramas, atribui tarefas aos membros da equipe, acompanha cronogramas, dá condições aos membros realizarem as tarefas, preocupa-se com questões políticas e administrativas
- Arquiteto de sistema: compreende como o sistema funciona de acordo com o contexto do problema, tem visão técnica mais global sobre o sistema e os problemas
- Especialistas: experts no domínio do problema que entendem e validam os requisitos do usuário e atuam com os engenheiros de regras de software
- Engenheiro de regras de software: converte requisitos do usuário em regras de qualidade de dados e regras de negócio, através de consulta a especialistas
- Engenheiro de "Quality Assurance" (QA): define padrões e níveis de qualidade para as tarefas e monitora qualidade do projeto

2.1.10. Selecionar ferramentas de trabalho

Cada problema tem suas peculiaridades, mas o uso de uma ou mais das ferramentas automatizadas confere mais rapidez às atividades (Loshin, 2001: 476). Tais ferramentas devem executar as seguintes tarefas: limpeza automática dos dados, padronização de dados, verificação e validação de bases de dados, definição de regras amigáveis para o usuário, execução de regras e combinação aproximada

2.1.11. Definir modelo de metadados

Nesta etapa, deve-se definir o modelo de metadados, os quais servirão de base para a definição da qualidade de dados e regras de negócio (Loshin, 2001: 479). As seguintes informações deverão ser levantadas:

- Elementos genéricos: contato, descrição, palavras chave, versão
- Tipos de dados e domínios: alias, mapeamento, estrutura, tipos base
- Esquema de metadados: tabelas, atributos, programas de carga, visões, *queries*, transformações, índices, *triggers*, fontes de dados
- Uso e sumarização: restrições, usuários, permissões, agregações, relatórios
- Histórico

2.1.12. Definir regras de qualidade de dados

A inteligência do processo de melhoria de qualidade de dados estará principalmente concentrada nas regras de qualidade de dados, daí a importância vital desta etapa (Loshin, 2001: 479).

No início desta etapa, deve-se identificar os domínios de dados (se ainda não estão detalhadamente formalizados), que poderão ser enumerados ou descritivos. Recomenda-se iniciar os trabalhos pelos domínios enumerados, que em geral são mais simples, com poucos valores distintos e raramente nulos. Para levantar mais informações sobre os domínios, sugere-se analisar os padrões de conteúdo que surgem nos dados. Após a identificação do domínio, deve atribuir um nome único e significado semântico. Para os valores válidos do domínio, deve-se atribuir a eles tabelas de referência ou validação. Após isto, os domínios

devem ser documentados para que, em futuras identificações de domínio, permita-se a verificação de aderência deste domínio para outros atributos.
Após a identificação dos domínios, deve-se identificar os mapeamentos dos domínios, ou seja, o relacionamento que há entre os mesmos. Para o mapeamento, deve-se identificar os domínios origem e destino, elaborar uma lista de enumeração dos mapeamentos e definir as regras de mapeamento.
Após os mapeamentos estarem definidos, deve-se definir as regras de qualidade de dados. As regras podem ser de diferentes tipos: permissão de valores nulos, restrição de faixa de valor, restrição por domínio, restrição por mapeamento, regras que envolvem valores de outros campos, regras baseadas em outros registros ou tabelas, regras baseadas na etapa da cadeia informação e regras de transformação ou de atualização.

2.1.13. Fazer uso de Arqueologia / Data Mining (quando necessário)

Para melhor compreensão dos metadados mais complexos, pode-se utilizar de algoritmos complexos, como o caso de Data Mining (Loshin, 2001: 480). Assim, é possível descobrir-se conhecimento que estava embutido nestes dados. Há técnicas de data mining como: clusterização, descoberta de regras de associação, árvores de classificação e regressão, árvores de decisão e análise de ligação.

2.1.14. Gerenciar fornecedores de dados

Para evitar que fornecedores externos de dados “poluam” o ambiente com dados incorretos, deve gerenciar o trabalho efetuado por estes fornecedores.
Para executar esta tarefa, as seguintes observações são válidas (Loshin, 2001: 481): impedir a entrada de dados externos incorretos no sistema, especificar requisitos de qualidade de dados, estabelecer métricas para aferir qualidade de dados, realizar auditorias das métricas, divulgar resultados de auditorias e definir penalidades para má qualidade de dados.

2.1.15. Executar as melhorias

Neste ponto, toda a solução pode ser elaborada e colocada em prática (Loshin, 2001: 481).
A execução efetiva da melhoria se inicia com o desenho da arquitetura da solução. São explicitados todos os melhores locais na arquitetura para validar mecanismos de validação e reconciliação. Para um data warehouse, uma arquitetura possível é apresentada na figura 3:

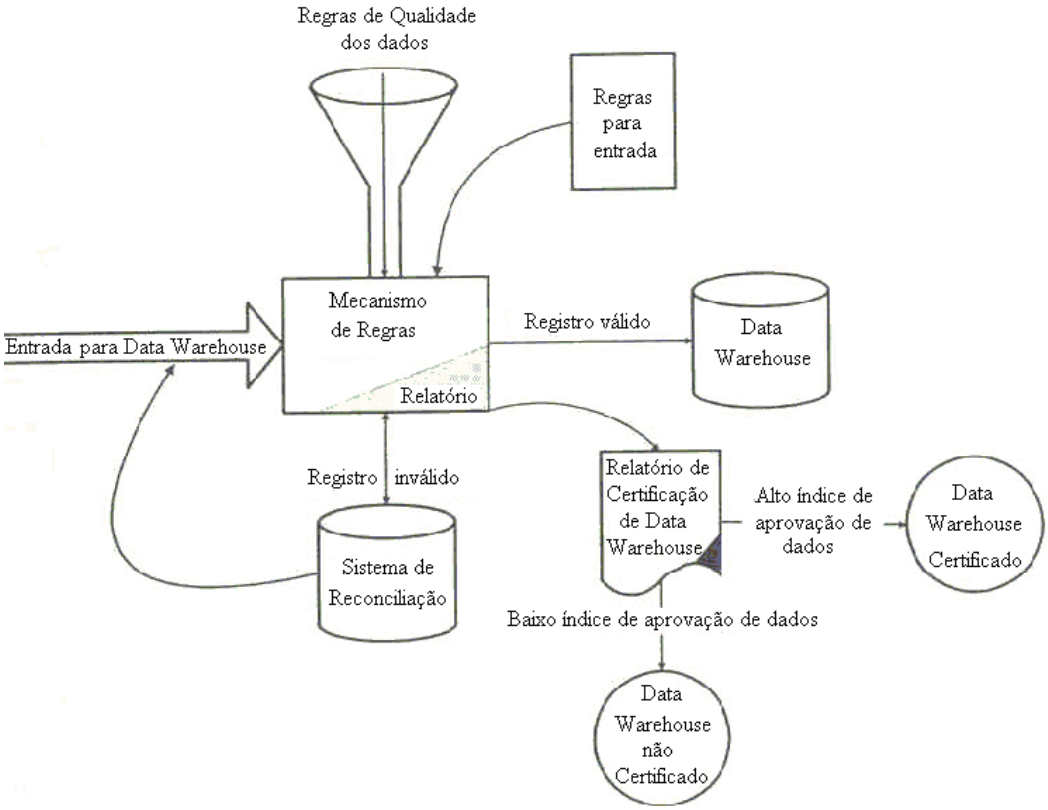

Figura 3 – Proposta de arquitetura de qualidade de dados com certificação de data warehouse (LOSHIN, 2001: 455)

Uma vez desenhada a arquitetura, deve-se executar uma limpeza estática nos dados já existentes. Existem vários aspectos a serem considerados em relação à limpeza estática, conforme descrito a seguir:

- Padronização - permite conformidade para comparação
- Parsing de Registros – permite dividir um dado composto em seus componentes
- Limpeza de Metadados – análise de conformidade dos dados com seus metadados
- Correção de Dados e Melhoria – execução dos ajustes necessários
- Similaridades: Medida, Fonética, Gramática – uso de algoritmos complexos
- Consolidação – eliminação de informação duplicada ou fortemente relacionada
- Campos não preenchidos – garantir que toda a informação necessária seja fornecida

Após a limpeza estática, as regras propostas devem ser implementadas e testadas em ambiente de teste. O Sistema de Reconciliação então deve ser implementado, caracterizando-se como um sistema de resolução de não conformidade. Em caso de entrada de registro incorreto, é gerada uma tarefa para a equipe proprietária daquele dado. Após todas as validações dos usuários, o sistema de regras pode ser migrado para o sistema de produção. Nesta etapa, deve-se garantir que todos os métodos de medição foram também migrados.

2.1.16. Medir melhorias

Após todo este trabalho, é necessário comprovar, através de evidências, que as melhorias obtiveram sucesso, daí a importância de garantir uma correta migração dos mecanismos de medição para o ambiente de produção. Um mecanismo interessante de controle histórico de comportamento de dados é o Controle Estatístico de Processo (CEP). As medições são feitas, e é estabelecido um número mínimo percentual de problemas de qualidade e um número

máximo percentual aceitável. Enquanto os dados com problemas estiverem nesta faixa, pode-se dizer que os problemas estão sob controle. Se os problemas crescem acima do limite, suspeita-se de falhas nos processos de qualidade, que devem ser então revistos. Entretanto, se os problemas estão muito abaixo do esperado, deve-se suspeitar que o sistema de medição está com problemas.

2.1.17. Explorar outros problemas ou projetos de qualidade de dados

Uma vez conseguida a aprovação da organização devido ao sucesso da implementação do projeto, fica mais fácil obter aprovação para posteriores projetos de melhoria para outras dimensões de qualidade ou para outros sistemas. Mesmo se isto não ocorrer, a experiência obtida garante à equipe bagagem suficiente para a resolução de problemas mais complexos.

3. Resultados

3.1 Proposta de caso prático

O caso prático aqui analisado refere-se à implementação de um novo sistema com alta qualidade de dados de indicadores financeiros, substituindo um sistema já existente no cliente. A geração dos indicadores é realizada logo após a finalização do processo de encerramento de período contábil executado no sistema transacional. Uma parte dos dados origem necessários é disponibilizada com razoável antecedência, porém a outra parte dos dados origem necessários é disponibilizada apenas no término do processo de encerramento de período contábil.
A empresa cliente é de grande porte, presente em vários países do mundo, sendo que atualmente o autor trabalha neste cliente como prestador de serviços através de outra empresa multinacional de consultoria de tecnologia e processos. O sistema deveria ser concebido como uma solução regional aplicável para todo o continente. As informações geradas seriam enviadas para a matriz mundial da companhia.
Para o caso prático, as atividades da metodologia foram seguidas, com exceção da atividade 5 (Elaborar o "Data Quality Scorecard"), uma parte da atividade 7 (Avaliar requisitos de qualidade de dados) e do passo 13 :(Fazer uso de Arqueologia / Data Mining). A seguir, é apresentada uma descrição da execução de cada passo da metodologia, ajudando a vislumbrar o grau de aderência do caso prático à metodologia.
Em relação à atividade 1 (Reconhecer o problema), foram identificados os seguintes problemas no sistema antigo:

- Cargas completas de todos os dados necessários após o término de encerramento contábil, gerando graves problemas de desempenho no sistema transacional
- Processo complicado, de difícil compreensão pelo usuário e pela equipe de suporte
- Difícil rastreamento do processo
- Conceitos de negócio de cada país embutidos nos programas de interface, levando a definições ambíguas
- Programas de interface diferenciados por tipo de informação contábil
- Obsolescência do sistema de geração de indicadores financeiros, cujo fornecedor previa o fim da garantia e suporte para a versão em uso

Em relação à atividade 2 (Buscar suporte gerencial e formalizar política de Propriedade de Dados), a implementação recebeu apoio da alta administração da empresa em nível continental. Os usuários envolvidos receberam suas responsabilidades e ganharam acesso ao

sistema com perfis distintos por grupo, restringindo assim suas permissões de acesso a suas atividades específicas.
Em relação à atividade 3 (Divulgar Programa de Qualidade de Dados), foram realizadas várias reuniões gerenciais e treinamentos com usuários para divulgar a decisão da nova implementação, bem como os benefícios esperados da mesma. Posteriormente, os usuários receberam treinamento técnico, de acordo com a tecnologia empregada.
Em relação à atividade 4 (Mapear a cadeia de informação), foram mapeados os diversos estágios envolvidos, desde a obtenção dos dados gerados no encerramento contábil do sistema transacional até a visualização da informação para análise gerencial.
Em relação à atividade 5 (Elaborar o "Data Quality Scorecard"), que ajudaria a identificar as melhores oportunidades de melhoria, não houve montagem formal do "Data Quality Scorecard". Não fazia muito sentido prático escolher pontos específicos na cadeia de informação, uma vez que a idéia era substituir todo o sistema de geração de indicadores financeiros, cujo fornecedor tinha declarado a obsolescência da versão. Era obrigatório decidir entre implementar um novo sistema ou fazer atualização de versão de um sistema com os problemas encontrados na atividade 1.
Em relação à atividade 6 (Avaliar a situação atual), o cenário foi avaliado e decidiu-se dar maior ênfase às seguintes dimensões de qualidade de dados (em ordem decrescente de prioridade): acuracidade, disponibilidade em tempo e compreensibilidade.
Em relação à atividade 7 (Avaliar requisitos de qualidade de dados), não havia uma definição de um "Data Quality Scorecard" a ser confrontado com a avaliação da situação atual. Entretanto, os requisitos de cada problema foram definidos e refinados, e são apresentados a seguir de forma resumida, de acordo com sua relação com as dimensões de qualidade de dados priorizadas.

- Acuracidade: Os dados passariam a ser extraídos de modo uniforme através de programas de carga vindos do fabricante, que era o mesmo do sistema transacional
- Disponibilidade em tempo: O processo de geração de indicadores financeiros deveria ser concluído em até 3 dias (exigência da matriz mundial)
- Compreensibilidade: Os usuários deveriam poder tratar de dados sem ambigüidades, de maneira uniforme para todo o continente, fazendo uso de relatórios de consulta no estilo Web.

Em relação à atividade 8 (Escolher um projeto piloto), foi escolhido apenas um país para implementação. O sucesso desta implementação seria vital para a posterior implementação nos demais países do continente.
Em relação à atividade 9 (Montar uma equipe), foram alocados um gerente de projeto do lado do cliente e um gerente de projeto do lado da equipe de implementação. Os dois usuários mais experts atuaram como arquitetos de sistema. Do lado da equipe de implementação, foram designados dois especialistas, dois engenheiros de regras de software e um engenheiro de "Quality Assurance".
Em relação à atividade 10 (Selecionar ferramentas de trabalho), foram utilizadas ferramentas para verificação e validação de banco de dados, definição de regras amigável para o usuário e execução de regras.
Em relação à atividade 11 (Definir modelo de metadados), o modelo criado continha:

- Tipos de dados e domínios: alias, mapeamento, estrutura, tipos base
- Esquema de metadados: tabelas, atributos, programas de carga, visões, *queries*, transformações, índices, *triggers*, fontes de dados
- Uso e sumarização: restrições, usuários, permissões, agregações, relatórios
- Histórico

Em relação à atividade 12 (Definir regras de qualidade de dados), todos os domínios, seus mapeamentos e regras de qualidade de dados foram definidas e padronizadas já pensando num modelo regional, que atendesse a todo o continente, não só ao país do projeto piloto.

Em relação à atividade 13 (Fazer uso de Arqueologia / Data Mining), entendeu-se que os metadados não tinham um nível de complexidade alto o suficiente para justificar a utilização de programas que explorassem técnicas de Arqueologia ou Data Mining.
Em relação à atividade 14.(Gerenciar fornecedores de dados), os dados provenientes de sistemas externos eram validados durante cada processo de carga, apresentando o resultado das validações em um relatório de acesso a todos os envolvidos.
O detalhamento das demais atividades é apresentado nas seções seguintes.

3.2 Tecnologia empregada

Complementando a descrição da atividade 10 (Selecionar ferramentas de trabalho), para o caso prático decidiu-se empregar a tecnologia presente no software de gestão empresarial SAP, cuja solução nativa de Data Warehouse recebe o nome de SAP BW (Business Warehouse). O sistema transacional já empregava a tecnologia SAP R/3.

No SAP BW, a extração de informações é feita basicamente com três tipos de objetos:

- Fontes de Dados: Contém a definição física da localização dos sistemas legados
- Fontes de Informação: Contém o mapeamento de informação entre origem e destino
- Provedores de Informação: Armazenam informação em caráter definivo, como cubos

O processo de extração se inicia com a leitura das fontes de dados, e sobre elas são aplicadas as chamadas regras de transferência, que são regras mínimas de aceitação para entrada de informação no sistema, constituindo-se nas regras mais simples. As informações aceitas são armazenadas nas fontes de informação, de acordo com mapeamento pré-definido de campos. A partir das fontes de informação, o fluxo de dados passa então pelas chamadas regras de atualização, regras mais complexas, com fórmulas e composição, e em caso de aceitação as regras chegam aos provedores de informação.

3.3 Resultados obtidos

Em relação à atividade 15 (Executar as melhorias), a solução foi desenvolvida, testada, aprovada e enviada para ambiente de produção. Durante os primeiros três meses, os valores obtidos foram confrontados com o sistema transacional e também com o antigo sistema, verificando-se que os valores financeiros apresentados estavam corretos.
Alguns ajustes foram feitos após os primeiros processos de encerramento de período contábil, apresentando boa estabilidade no momento atual, onde os dados são confrontados apenas com o sistema transacional.
Em relação à atividade 16 (Medir melhorias), foram observadas as seguintes melhorias:

- A parte dos dados disponibilizados com antecedência passou a ser obtida à noite com freqüência diária, sem onerar o sistema transacional. Somente os dados disponibilizados apenas após o término de encerramento contábil são carregados pouco antes do processo de geração de indicadores financeiros, aliviando em várias horas a carga de processamento no sistema transacional após o término de encerramento contábil
- O processo foi bastante simplificado, aumentando a compreensão por parte do usuário e da equipe de suporte
- O rastreamento do processo foi facilitado
- Os conceitos de negócio foram uniformizados para todos os países do continente, eliminando definições ambíguas

- Os programas de interface foram eliminados, uma vez que o novo sistema disponibiliza um pacote pronto para extração de dados
- Foi obtido um moderno sistema de geração de indicadores financeiros, com garantia de suporte para um período muito mais extenso

Em relação à atividade 17 (Explorar outros problemas ou projetos de qualidade de dados), o sucesso da implementação contribuiu para que outro país do continente optasse pela implementação, que no momento presente encontra-se em andamento. Representantes de vários países de outro continente (Europa) vieram ao Brasil para aprender a solução para poder implementá-la por várias filiais da Europa.
Atualmente o autor atua na área de suporte à solução implementada no país piloto, e espera-se que também dê suporte à solução para todos os países onde for implementada.

4. Conclusões

A qualidade de dados é responsável por perdas de tempo, dinheiro e oportunidades. Como muitas organizações não conseguem vislumbrar a dimensão destes valores, elas continuam executando tarefas que não agregam valor. Se elas se conscientizassem da gravidade do problema, provavelmente lhe dariam a devida prioridade e partiriam para ações que levassem a soluções definitivas.
Um dos grandes desafios do CIO (Chief Information Officer) moderno é sensibilizar os executivos da importância da garantia de qualidade de dados. (Tayi; Ballou, 1998: 3).
Entendeu-se que é essencial que se determine o nível de qualidade de dados necessário para a organização, a fim de que se dimensione corretamente a quantidade de esforço e recursos a serem exigidos. Esta tarefa muitas vezes é complicada, uma vez que diferentes usuários têm diferentes necessidades. Além disso, há o problema da semântica, pois um usuário que tem profundo contato com os dados e suas nuances pode atribuir à informação um significado muito diferente do que outros usuários (Tayi; Ballou, 1998: 1). Este fato foi observado no caso prático, uma vez que usuários de cada país tinham uma interpretação diferente para os dados.
Entendeu-se que a qualidade de dados também indica quão aderentes os dados do sistema estão em relação aos dados do mundo real (ORR, 1998:2). Observou-se que a qualidade dos dados pode ser melhor garantida através do incentivo contínuo aos usuários da utilização dos dados (ORR, 1998: 6). Para o caso prático, a substituição de sistema obriga a todos os usuários envolvidos no processo a entenderem e utilizarem a nova solução.
Ficou evidente que a qualidade de dados não é um conceito subjetivo: ela pode ser medida, avaliada, analisada e ter requisitos próprios. Além disso, a qualidade de dados pode ser projetada, implementada, aperfeiçoada e documentada. Por tudo isso, conclui-se que a qualidade de dados tem um significado concreto e que ela pode ser melhorada continuamente, sendo implementada através de metodologia bem definida e obtendo bons resultados comprovados através de indicadores.

5. Recomendações

Alguns erros que podem ser evitados para garantir que o data warehouse disponibilize informações de alta qualidade são (English, 2002: 1):

- Assumir que as fontes de dados são seguras porque os sistemas operacionais parecem funcionar bem: como exemplo, cita-se uma companhia de seguros onde 80% dos diagnósticos de saúde constavam como “perna quebrada”. Como ninguém analisava

esta informação, os digitadores decoraram o código e o utilizavam para não precisarem procurar o código correto. O dado era válido, porém incorreto

- Colocar muito foco em desempenho em detrimento da qualidade de dados do data warehouse: no pior cenário, erros serão cometidos pela organização mais rápido do que nunca. No melhor cenário, os clientes estarão insatisfeitos, deixando de explorar toda a potencialidade do data warehouse, abalando a credibilidade para futuros projetos. Performance sempre deve ser um requisito secundário em relação à qualidade de dados

Algumas recomendações que podem ser formuladas para uma modelagem eficiente do data warehouse são (English, 1996: 6):

- Modele somente dados cujo valor aumenta com o tempo: nem todos os dados operacionais devem ir para o data warehouse. Devem ser modelados dados que permitam identificar tendências ou dados que respondam a perguntas relevantes ou dados que são requeridos devido a propósitos legais
- Mantenha os dados base dos dados derivados ou sumarizados. A verificação futura dos dados derivados deverá ser possível através dos dados base, garantindo a reputação de confiabilidade do data warehouse

No que se refere à "limpeza" dos dados, algumas observações a serem feitas são (English, 1996: 7):

- Inicie a limpeza com um pequeno grupo de dados de alta importância
- Analise e descubra o significado, valores e regras de negócio da fonte de dados: este processo pode evidenciar regras ainda não percebidas
- Conduza auditoria eletrônica nos dados para verificar a conformidade com as regras de negócio
- Conduza auditoria física nos dados para verificar o seu real nível de acuracidade
- Explore a automatização em larga escala
- Desenvolva regras de transformação com cuidado e confira os dados de saída
- Envolva os usuários relacionados aos dados nos processos de limpeza e auditoria: isto fortalece uma cultura de prevenção de erros
- Limpe os dados na fonte se eles ainda são usados: não restrinja a limpeza ao data warehouse. Os relatórios da fonte de dados e do data warehouse devem servir para reconciliação de dados
- Registre o tempo e os custos envolvidos na limpeza: isto ajuda a justificar ações para prevenção de erros

Finalmente, algumas considerações que podem ser feitas sobre o gerenciamento da qualidade da informação são (English, 2002: 3):

- Entenda a qualidade de informação como um problema de negócios, não como um problema de sistemas, e resolva-o como um processo de negócio, não como um processo de sistema
- Dê também foco a clientes e fornecedores da informação, e não somente nos dados
- Dê foco a todos os componentes da informação, incluindo definição, conteúdo e apresentação
- Implemente processos de gerenciamento de qualidade de informação, e não somente software de qualidade de informação
- Meça a acuracidade dos dados, e não apenas na sua validade

- Meça custos – e não somente percentuais – de informações de baixa qualidade e resultados de negócio de qualidade de informação
- Enfatize a melhoria e manutenção preventiva, e não apenas manutenção corretiva
- Melhore os processos na fonte dos dados, e não apenas em áreas isoladas
- Providencie treinamento em qualidade de dados a gerentes e produtores de informação
- Não apenas implemente atividades, mas transforme a cultura local de forma ativa

6. Referências

ENGLISH, Larry – "The Essentials of Information Quality Management", White paper at Web Site www.information-quality.com, setembro de 2002

ENGLISH, Larry – "Mistakes to Avoid If Yor Data Warehouse is to Deliver Quality Information", White paper at Web Site www.information-quality.com, junho de 2002

ENGLISH, Larry – "Data Quality in The Data Warehouse: a Key Critical Success Factor", White paper at Web Site www.information-quality.com, outubro de 1996

LOSHIN, David – "Enterprise Knowledge Management – The Data Quality Approach", Academic Press, 2001

OLSON, Jack E – "Data Quality – The accuracy dimension", Morgan Kaufmann Publishers, 2003.

ORR, Ken – "Data Quality and Systems Theory", Communications of the ACM, V. 41, N° 2, fevereiro de 1998.

REDMAN, Thomas C. – "Data Quality: The Field Guide", Digital Press, 2001.

REDMAN, Thomas C. – "The Impact of Poor Data Quality on the Typical Enterprises", Communications of the ACM, V. 41, N° 2, fevereiro de 1998.

TAYI, Giri T.; BALLOU, Donald P. – "Examining Data Quality", Communications of the ACM, V. 41, N° 2, fevereiro de 1998.

WANG, Richard Y. – "A Product Perspective on Total Data Quality Management", Communications of the ACM, V. 41, N° 2, fevereiro de 1998.