ARTIGO

Proposta de uma Arquitetura para o Processo de Melhoria da Qualidade

Marcello Sandi Pinheiro

***Resumo*—Esse artigo aborda sobre um processo de melhoria de qualidade de dados para um ambiente de tomada de decisão. Adota a perspectiva da informação como um produto dentro do ciclo em seu ciclo de vida dentro da organização As questão inerentes também visam o tempo em que a carga no *Data Warehouse*, uma vez que uma informação ela é relevante no momento que em que também é pontual. Foi demonstrado que o processo de melhoria produziu melhoras significativas no tempo da carga e na qualidade dos dados no repositório em questão.**

***Palavras-chave*—Qualidade de Dados TDQM, Processo, ETL, Metodologia.**

I. INTRODUÇÃO

DURANTE os últimos anos ocorreu um crescimento significativo dos Sistemas de Informações nas organizações, fruto de um processo natural de implantação de tecnologias visando a automatização das áreas e atividades dentro da organização e, com isso, a dinamização dos serviços dentro da organização como um todo. Tecnicamente, tais Sistemas de Informações (SI) mantém repositórios com os dados que, na verdade, representam o negócio da organização, e são, na realidade, o que há de mais precioso depois dos recursos humano da organização.

Além disso, sabendo que os repositórios podem ser acessados por diversos sistemas e considerando a suas características evolutivas, o problema principal está em garantir que a informação esteja íntegra e "saudável", quer dizer, livre de inconsistências e redundâncias, tanto as relacionadas à padronização e completude quanto as relacionadas às regras de negócio.

Nesse sentido, ao se levar em consideração que os dados são usados para diversas finalidades, inclusive para a tomada de decisão, há uma necessidade intrínseca de qualidade dos dados, pois há uma razão bem simples para que isso aconteça: garantir a confiabilidade da informação [1].

Em [2] afirma-se que a importância da qualidade dos dados nos processos decisórios e operacionais é reconhecido por várias instituições e organizações internacionais. O Data Warehousing Educational & Solution [3] afirma que a há uma diferença significativa entre a qualidade dos dados percebida e a real em muitas organizações. Tais problemas com qualidade dos dados geram um prejuízo de mais de 600 bilhões de dólares por ano nas empresas Norte-Americanas.

Tamanho é o problema que a má qualidade dos dados pode ocasionar, que algumas organizações criaram departamentos específicos para a Gestão da Qualidade dos Dados e da Informação. Apenas para citar alguns exemplos, no nosso país a Brasil Telecom (atual Oi Telecom), a Submarino e o Serasa. No exterior a Fedex, a Cedars-Sinai Medical Center e o Exército Norte-Americano ([4], [5] e [1]).

A qualidade da informação numa organização propicia a entrega da informação correta, no tempo adequado, no local indicado e às pessoas certas. Não há como o decisor fazer algum juízo efetivo a partir de dados falhos, incompletos e/ou imprecisos ([2] e [6]).

Considerando tais aspectos, nesse artigo serão abordados assuntos sobre metodologia de qualidade de dados, propondo-se uma arquitetura para melhoria da qualidade de dados nos repositórios utilizados no *Data Warehouse* (DW) do Sistema Integrado de Gestão (SIG) e o resultado de uma prova de conceito em uma ferramenta de qualidade de dados.

II. PROPOSTA METODOLÓGICA EM QUALIDADE DE DADOS

A garantia de qualidade dos dados e da informação exige uma infraestrutura metodológica validada e que propicie, dessa forma, garantias de que a implantação desse processo na organização tenha êxito. A metodologia Total *Data Quality Managment* (TDQM) (MIT, 2010) foi desenvolvida pelo *Massachusetts Institute of Techn*ology (MIT) e adota a perspectiva da informação como um produto. A TDQM prevê métricas de qualidade da informação, mede a qualidade ao longo do ciclo de vida da informação, analisa e identifica as causas que geram problemas de qualidade e define a implementação do processo de melhoria da qualidade dos dados.

A TDQM é um processo iterativo e incremental, onde são definidas etapas e fases bem específicas. Como resultado da aplicação da TDQM na organização cria-se o Plano de Qualidade de Dados. A Fig. 1 ilustra as quatro etapas dessa metodologia.

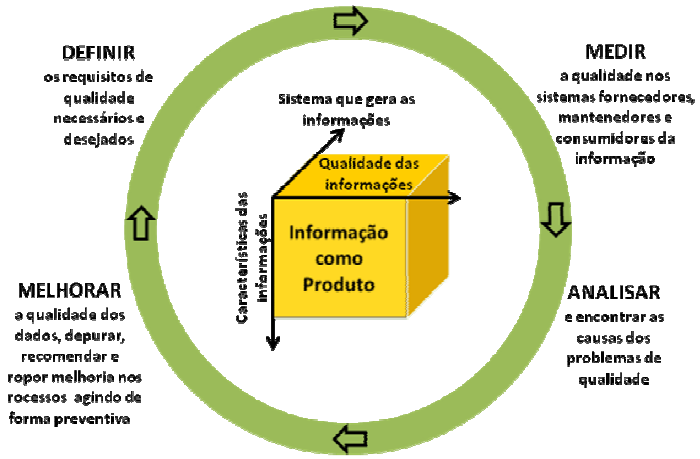

Fig. 1. Ciclo da Metodologia TDQM.

III. Arquitetura para a Melhoria da Qualidade

Trata-se um processo suportado pela TDQM, onde os repositórios operacionais "alvo" passam pelo tratamento e melhoria da qualidade dos dados. Começa após o mapeamento das tabelas que serão replicadas em uma área denominada STAGE, com o intuito de, primeiro, manter os repositórios de origem na produção e, segundo, não onerar os Sistemas de Gerenciamento de Banco de Dados (SGBD) operativos desses repositórios, conforme ilustrado na Fig. 2.

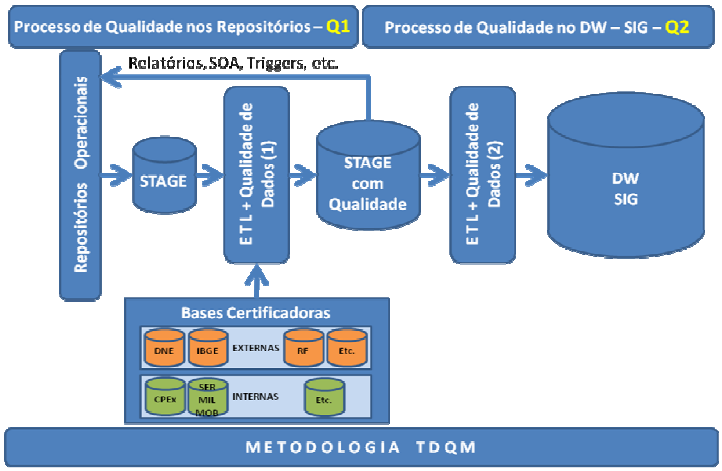

Fig. 2. Arquitetura Proposta para o Projeto de Qualidade de Dados.

As tabelas na STAGE irão passar pelas etapas da TDQM com o intuito de gerar o Plano de Qualidade de Dados. Dessa forma, os problemas são mapeados e as regras de tratamento e melhoria de qualidade de dados são definidas e implementadas em uma ferramenta específica para qualidade de dados, a qual dá suporte a técnicas de Extração, Transformação e Carga (conhecido como ETL) e também a recursos específicos para a qualidade de dados utilizando técnicas estatísticas, probabilísticas e de inteligência artificial.

A proposta é criar um modelo contínuo de melhoria da qualidade, com processos bem definidos e com uma equipe específica responsável pelo tratamento e melhoria da qualidade dos dados.

Em um primeiro momento, após a aplicação de regras de qualidade de dados pela ferramenta, os dados com qualidade serão armazenados na "*STAGE* com Qualidade" (Processo Q1). Os dados serão replicados sistematicamente aos repositórios de origem e, também, servirão de fonte para a carga do *Data Warehouse* (DW) do Sistema Integrado de Gestão (SIG) (Processo Q2).

Dessa forma, esperam-se como benefícios, não somente a melhoria dos dados nos repositórios de origem, como também a criação de relatórios para reportar inconsistências, relatórios de acompanhamento da qualidade dos dados, propensão para criar repositórios com dados de interesse corporativo e unificado, dentre outros, tudo de forma contínua e incremental.

IV. Utilização de uma Ferramenta de Qualidade de Dados

Até o presente momento, foram realizadas duas provas de conceito (PoC) com software de qualidade de dados. No entanto, somente com o IBM *InfoSphere DataStage/Quality*Stage Versão 8.1.2 foi possível implementar um teste onde os processos de ETL e qualidade de dados foram implementados para simular um carga completa do *Data Mart* (DM) de Pessoal do SIG.

Nesse PoC foram implementadas 42 tarefas de criação de arquivos (*datasets*), 36 tarefas de carga em tabelas dimensionais, 12 tarefas de ETL para o ciclo completo de carga da tabela fato, tendo esse último abrangido todas as regras de negócio e, inclusive, utilizando um arquivo em formato texto com dados de pessoal proveniente do Centro de Pagamento do Exército (CPEx). Nesse processo foram manipulados 218.906.640 de registros. Todo o processo de ETL e qualidade de dados foram processados em uma máquina virtual VMWare 3.0.0 de 32 Bits e com 3 GB de memória virtualizada, com sistema operacional Red Hat 5.0 e espaço em disco de 160 GB.

O tempo de processamento para a carga de todas as 17 dimensões foi inferior a 5 minutos. Todo o processo de preparação da carga, com as regras de negócio, foi realizado em 2 horas e todo o fluxo, inclusive com a carga no DW e reconstrução dos índices no Oracle 10g Standard foi feito em quase 4 horas, um contraste significativo em relação ao processo atual que leva um pouco mais de 24 horas usando *Oracle Warehouse Builder* (OWB), sem abranger técnicas de qualidade de dados. No processo de qualidade de dados, foi realizada uma tarefa de saneamento de dados através de funcionalidades e técnicas específicas da ferramenta para investigação, de duplicação e padronização de dados. Nesse processo, 18,97 % de dados discrepantes foram detectados e saneados.

V. CONCLUSÃO

Ted Friedman, vice-presidente de pesquisa do *Gartner Group*, afirma que qualidade de dados não é um problema da Tecnologia da Informação, e sim, um problema de gestão de negócio e processos. Quer dizer, a área de negócio deve assumir as responsabilidades e conduzir os processos de melhoria continuamente dentro da organização apoiando-se em uma ferramenta de qualidade de dados. Como mostrado no PoC, espera-se que os benefícios sejam a curto e médio prazo, com a implantação de um Plano de Qualidade suportado por uma ferramenta de qualidade de dados. Nesse processo de qualidade, a ferramenta da IBM citada disponibiliza uma camada de serviços baseada nos processos batch de qualificação, fazendo com que os dados sejam qualificados ainda na origem, além de possuir um recurso que transforma uma tarefa, seja de ETL ou de qualidade de dados, em um *Web Service* para ser disponibilizado no barramento de serviços da organização, ampliando, dessa maneira, as possibilidades de tratamento dos dados e ratificando a necessidade da aquisição de uma ferramenta desse nível para o Exército Brasileiro.

REFERÊNCIAS

[1] HUANG, K. T., LEE, Y. W., WANG, R. Y. *Quality Information and Knowledge*. Prentice Hall, 2001.
[2] BATINI, Carlo, SCANNAPIECA, Monica. *Data Quality: Concepts, Methodologies and Techniques*. Springer-Verlag, 2006.
[3] TDWI - Data Warehousing Educational & Solution. (05/05/2010). Acessível em:http://tdwi.org/.
[4] MIT - Massachusetts Institute of Technology. (23/06/2010). Acessível em: http://web.mit.edu/tdqm/.
[5] OLSEN, Jack E. *Data Quality: The Accuracy Dimension*. Morgan Kaufmann, 2003.
[6] McGILVRAY, Danette. *Executing Data Quality Projects: Ten steps to quality data and trusted information*. Morgan Kaufmann, 2008.

Autor: Professor da Faculdade Projeção, é doutor em Ciências. Área de Concentração: Sistemas Computacionais. Sub-áreas: Mineração de Textos; Lingüística Computacional; Inteligência Computacional. Linha de Pesquisa: Análise de Informação Não Estruturada. Programa de Pós-graduação em Engenharia Civil - COPPE/UFRJ. Tem mestrado em Ciência da Computação. Área de Concentração: Inteligência Artificial. Sub-áreas: Mineração de Dados; Inteligência Computacional; Tecnologias de Apoio a Decisão; Inteligência Artificial. Linha de Pesquisa: Mineração de Dados no Apoio a Decisão. Centro de Estudios em Ingeniería de Sistemas CEIS, Instituto Superior Politécnico José Antônio Echeverría ISPJAE. Ciudad de La Habana Cuba, apostilado pelo Departamento de Ciência da Computação da UFRGS. Graduado em Processamento de Dados pela Universidade Católica de Brasília - UCB. Iniciação científica em Inteligência Artificial atuando no projeto Meta-Tutores Inteligentes na UCB. É Consultor de BI, Qualidade de Dados e Desenvolvedor de Sistemas de Informação. Tem experiência na área de Ciência da Computação, com ênfase em Mineração de Dados e Textos, Data Warehouse e Recuperação de Informação. Participou no desenvolvimento de Sistemas envolvendo Banco de Dados Oracle e SQL Server usando Java, Delphi, C/C++. Na área acadêmica desenvolveu atividades sobre: Modelos de Classificação Não-Supervisionada e Supervisionada, Modelos Preditivos com RNN, Modelos de Análise de Cestas de Mercado, Sistemas Baseados em Conhecimento; Formalismos para a Representação do Conhecimento; Algoritmos de Busca e Otimização Multi-Objetivo; Algoritmos Genéticos; Rede Neurais; Análise e Técnicas de Algoritmos; Otimização do Desempenho de Atletas em Esportes de Alto-rendimento. Email: msandipinheiro@yahoo.com.br