Passos iniciais para um programa de qualidade de dados

O desafio da qualidade de dados

As organizações dependem de dados de qualidade para fundamentar as decisões de negócios, fornecer suporte aos clientes, desenvolver planos e administrar outras tarefas fundamentais. Mas com a proliferação das fontes e o crescimento exponencial dos volumes, manter dados de alta qualidade pode ser uma tarefa difícil. Essas deficiências por vezes resultam em histórias engraçadas, como o caso de um homem de 50 anos de idade que recebeu uma carta de seu médico, confirmando a próxima ultrassonografia para exame da sua gravidez. No entanto, dados de baixa qualidade também podem acarretar graves repercussões. Em um incidente, uma jovem mãe no Reino Unido faleceu em decorrência de câncer de mama, pois houve um atraso no diagnóstico e subsequente tratamento. O que causou o atraso? Um erro no prontuário da paciente, que informava o número de sua residência como "16" em vez de "1b". Consequentemente, as cartas enviadas pelo hospital nunca chegaram até a paciente, e ela perdeu consultas importantes.[1]

Tratando os problemas de qualidade de dados, as organizações evitam consequências lamentáveis e melhoram os resultados dos negócios. Excluindo as falhas humanas, a maioria dos problemas de qualidade de dados é decorrente da falta de padrões de informação em toda a empresa, que expliquem como os dados devem ser armazenados e identificados de forma exclusiva. A inconsistência entre as fontes dificulta a compreensão dos relacionamentos entre entidades de negócio essenciais, tais como partes (fornecedores, clientes, funcionários, etc.) e produtos. Em muitos casos, não existe uma chave confiável e persistente para recuperar toda a informação na empresa que esteja associada a uma única parte ou produto.

Dados de alta qualidade permitem que sistemas estratégicos integrem todos os dados relacionados, proporcionando uma visão completa da organização e de seus inter-relacionamentos. Sem dados de alta qualidade em toda a empresa, as organizações não podem contar com o retorno sobre os investimentos feitos em aplicativos de negócios críticos, tais como data warehouses, ferramentas de inteligência de negócios e sistemas de dados mestre. Ao implementar um programa de qualidade de dados, as organizações são capazes de reforçar a integridade dos dados e com isso aproveitar ao máximo seus ativos de informação.

Várias ferramentas fazem parte de uma solução de qualidade de dados: software de padronização de dados, mecanismos de combinação de dados, gerenciamento de metadados, consoles de terminologia de TI/negócios e software para integração de dados, entre outros. Integrar essas ferramentas de uma só vez em uma solução completa de qualidade de dados pode ser uma tarefa opressiva para muitas organizações de TI. Além disso, as organizações podem ter prioridades conflitantes em relação à qualidade de dados. Um departamento pode simplesmente querer fazer a limpeza de endereços em um arquivo mestre de clientes, enquanto outro está realizando a fusão de sistemas financeiros de uma empresa recém-adquirida, e ainda outro departamento está sendo pressionado pela administração devido à confiabilidade suspeita dos relatórios de vendas mensais.

A chave para responder a esses desafios é a flexibilidade. Uma plataforma comum de integração de dados que possa ser aplicada a vários problemas de qualidade de dados desempenha um papel importante na estratégia de governança de informações de uma organização. Tal plataforma pode ser um investimento melhor do que a compra de uma série de soluções pontuais não integradas para resolver problemas à medida que vão surgindo.

Governança eficaz da informação em toda a cadeia de suprimentos de informações

Uma organização típica pode hospedar centenas ou até milhares de sistemas diferentes. A informação pode ter origem em muitos lugares (sistemas de transação, sistemas operacionais, repositórios de documentos, fontes externas de informação) e em muitos formatos (dados, conteúdo, streaming). Muitas vezes existem relacionamentos significativos entre as várias fontes e tipos de dados. Essa cadeia de suprimentos de informações flui por toda a organização e para além de suas fronteiras (ver Figura 1). Ao contrário das entidades em uma cadeia de suprimentos tradicional, as entidades em uma cadeia de suprimentos de informações possuem um relacionamento de muitos-para-muitos. Os mesmos dados sobre uma pessoa – que pode se tratar de um cliente, funcionário e/ou parceiro, por exemplo – podem vir de várias fontes, e essas informações acabam em muitos relatórios e aplicativos. Além disso, sistemas distintos também podem definir a informação de uma forma diferente.

Diante dessa complexidade, integrar as informações, garantir sua qualidade e interpretá-la corretamente são etapas fundamentais para apoiar as melhores decisões. As informações devem ser transformadas em um ativo confiável e controladas para manter a qualidade ao longo do seu ciclo de vida. Os sistemas subjacentes devem ser econômicos e fáceis de manter, assim como devem apresentar um bom desempenho na execução das cargas de trabalho, mesmo quando o volume de informações manipuladas crescer astronomicamente.

Uma governança eficaz das informações melhora a qualidade, a disponibilidade e a integridade dos dados de uma empresa ao estimular a colaboração entre organizações e a criação de políticas estruturadas. Isso gera equilíbrio entre os silos departamentais e o interesse organizacional, ajudando a aumentar a confiança nos dados – o que pode influenciar diretamente interesses de negócio importantes, como o aumento de receita, a diminuição de custos e a redução de riscos. Dados de má qualidade causam efeitos como: falha em processos de negócios, redução da produtividade e desperdício de materiais. Informações perdidas, imprecisas ou incompletas também podem gerar altos custos e trabalho extra, como buscar ou conciliar informações.

Características dos dados de alta qualidade

- **Integridade:** Todos os dados relevantes estão vinculados. Por exemplo, um registro de cliente completo deve incluir todas as contas, endereços e relacionamentos que a empresa possui desse cliente.
- **Exatidão:** Problemas de dados comuns, como erros de ortografia, erros de digitação, abreviações aleatórias e outros semelhantes foram resolvidos.
- **Disponibilidade:** Os dados requeridos já foram descobertos e estão disponíveis sob demanda; os usuários não precisam procurar manualmente pela informação.
- **Atualidade:** Qual valor um relatório de vendas terá se ele não tiver os dados do mês mais recente?

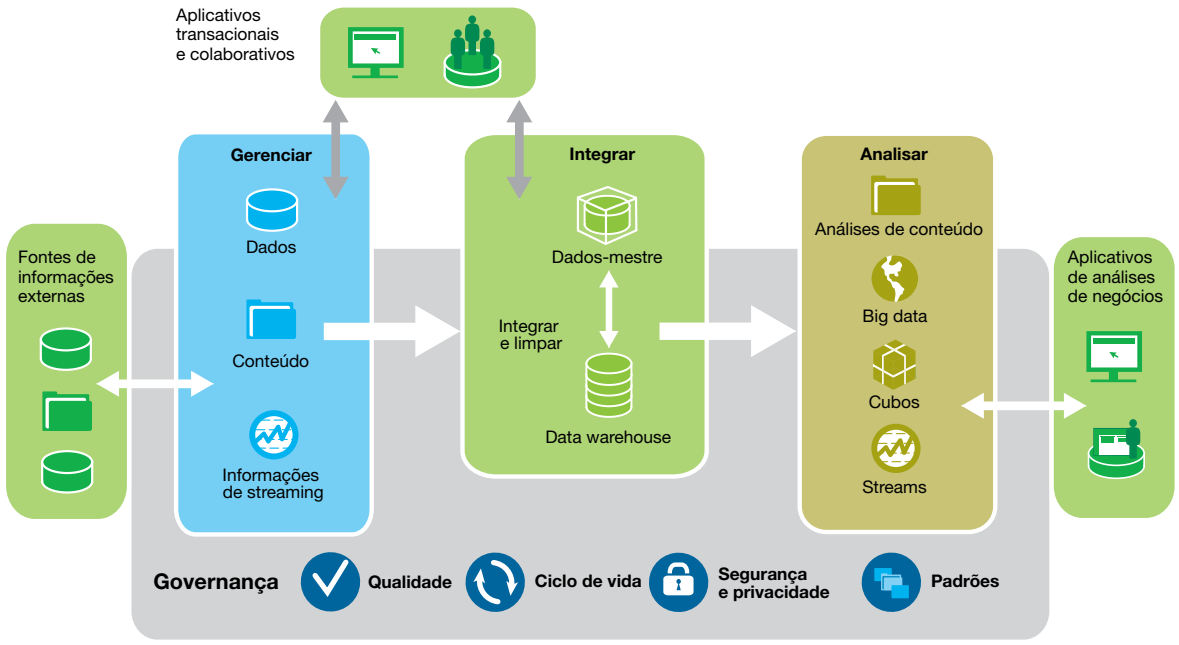

Figura 1: A cadeia de suprimentos de informações.

A excelente qualidade dos dados é essencial para o sucesso. Por exemplo, ela ajuda a proporcionar uma compreensão clara dos clientes, parceiros e fornecedores, o que pode representar a diferença entre o crescimento dos negócios e a sua incapacidade de competir.

Estabelecendo um programa de qualidade de dados: Passos iniciais

Para conseguir a governança adequada das informações, as organizações devem estabelecer um programa de qualidade de dados, com base nos objetivos e prioridades do negócio. Nem todos os problemas de qualidade de dados têm o mesmo impacto nos resultados da empresa; tentar resolver todos esses problemas em uma organização pode ser exaustivo e ineficiente. Considere as seguintes questões para estabelecer o retorno sobre o investimento (ROI) para cada iniciativa de qualidade de dados na organização:

- Quais são os processos de negócios mais críticos que dependem de informações?
- Quais informações são mais importantes para estes processos?
- Qual é o custo da informação deficiente para a efetividade destes processos?
- Qual é o custo de se manter informações de alta qualidade para estes processos?
- Qual é o benefício líquido para a organização por manter a qualidade dos dados para estes processos?

Seja qual for o caminho escolhido, as necessidades de qualidade de dados provavelmente mudarão ao longo do tempo, e por isso é importante investir em tecnologias que possam ser dimensionadas e aproveitadas em toda a empresa. Uma solução pontual que resolva os problemas de hoje – como uma solução de limpeza de endereços para aprimorar a exatidão e a consistência – poderá estar mal equipada para cumprir os requisitos de qualidade de dados de amanhã.

Além de planejar as necessidades futuras, um programa de qualidade de dados deve abordar duas questões fundamentais: Primeiro, a organização deve concordar com a definição de qualidade. O que são dados "bons"? São os dados que produzem uma taxa de erro menor que 1 por cento? Ou a organização pode tolerar uma taxa de erro de 10 por cento?

Vejamos o caso de um órgão do governo que precise de informações altamente precisas nos postos de fronteira. Erros de dados nessa situação podem ter consequências catastróficas. No entanto, é menos provável que erros na base de endereços de uma loja de roupas tenham consequências tão terríveis. Entender o objetivo (isto é essencial para a alocação de recursos e a gestão dos custos do programa) e não esperar perfeição é necessário ou até mesmo desejável em todos os casos.

A próxima questão envolve relatórios: Dada a definição de qualidade, quais métricas devem ser monitoradas para garantir que o limite mínimo de qualidade está sendo mantido? O programa de qualidade de dados deve incluir uma abordagem sistemática e orientada aos negócios para capturar e relatar essas métricas, e a organização deve articular as prioridades a serem formalmente documentadas e monitoradas ao longo do tempo.

O foco inicial dos esforços de qualidade de dados dependerá das respostas que a organização oferecer para estas perguntas.

O primeiro ponto de entrada para a qualidade de dados: A avaliação de qualidade dos dados

Seja embarcando em novas iniciativas ou resolvendo problemas de governança de dados e mitigação de riscos, muitas organizações acham que um bom ponto de partida é a avaliação de qualidade dos dados, que estabelece uma linha de base: Qual o nível de qualidade dos seus dados, e onde estão as maiores oportunidades de melhoria?

Muitas organizações acham difícil obter uma compreensão consistente dos seus dados em toda a empresa, especialmente porque os sistemas e aplicativos são adaptados para mudanças de requisitos e porque fusões e aquisições ampliam os conjuntos de dados existentes. Unidades de negócios e o setor de TI usam diferentes semânticas e dados de registro de aplicativos, de inúmeras formas – com identificadores diferentes (IDs de cliente e de conta), diferentes formatos (datas armazenadas em forma de data em um banco de dados, mas em forma de string em um arquivo) ou diferentes valores (gênero registrado como "M" ou "F" em um sistema e "0" ou "1" em outro). A abordagem da organização em relação à gestão do conhecimento pode complicar ainda mais a situação.

Muitas vezes o conhecimento sobre os dados é apenas tácito, guardado na cabeça das pessoas com base em seu trabalho específico. Ou pode estar registrado em documentos que não foram mantidos atualizados com os últimos processos de negócios ou alterações de sistemas. Quando as pessoas mudam de setor ou saem da organização, muito desse conhecimento acaba sendo perdido ou fragmentado.

A avaliação de qualidade dos dados destina-se a oferecer um insight deste quadro complexo, estabelecendo uma prática fundamental para trabalhos posteriores e uma base de conhecimento em um repositório de metadados compartilhado, que podem ser usados e reutilizados em vários projetos e iniciativas. Uma avaliação de qualidade dos dados tem como foco e ilustra um problema de negócios com base nos dados subjacentes. Este esforço ajuda a trazer à luz os problemas de qualidade de dados, mas um analista de dados ou de negócios ainda deve analisar os resultados e tirar as conclusões, principalmente para definir o impacto nos negócios.

Conforme mostrado na Figura 2, o analista de dados está no centro da avaliação de qualidade de dados. Os analistas devem compreender todo o panorama: o escopo, os objetivos e os resultados da avaliação. Se o analista não souber o que a organização está tentando alcançar, como poderá identificar uma verdadeira anomalia ou problema? O grande número de tabelas e atributos pode ser assombroso. Simplesmente selecionar várias colunas e executar uma tarefa de análise não é o fim do processo – a avaliação posterior é fundamental.

O IBM® InfoSphere® Information Server oferece recursos que ajudam os analistas a aprenderem e aplicarem técnicas de análise de dados. O treinamento no uso desses recursos – disponível na IBM – concentra-se em etapas de análise essenciais e melhores práticas:

- Identificação e abordagem das fontes de dados em questão
- Vantagens das funções de dados automatizadas baseadas em conteúdo
- Uso de classificações de dados para focar a análise
- Validação dos formatos e domínios de dados
- Emissão de relatórios e entrega de resultados e conclusões
- Retenção dos resultados da análise ao longo do tempo

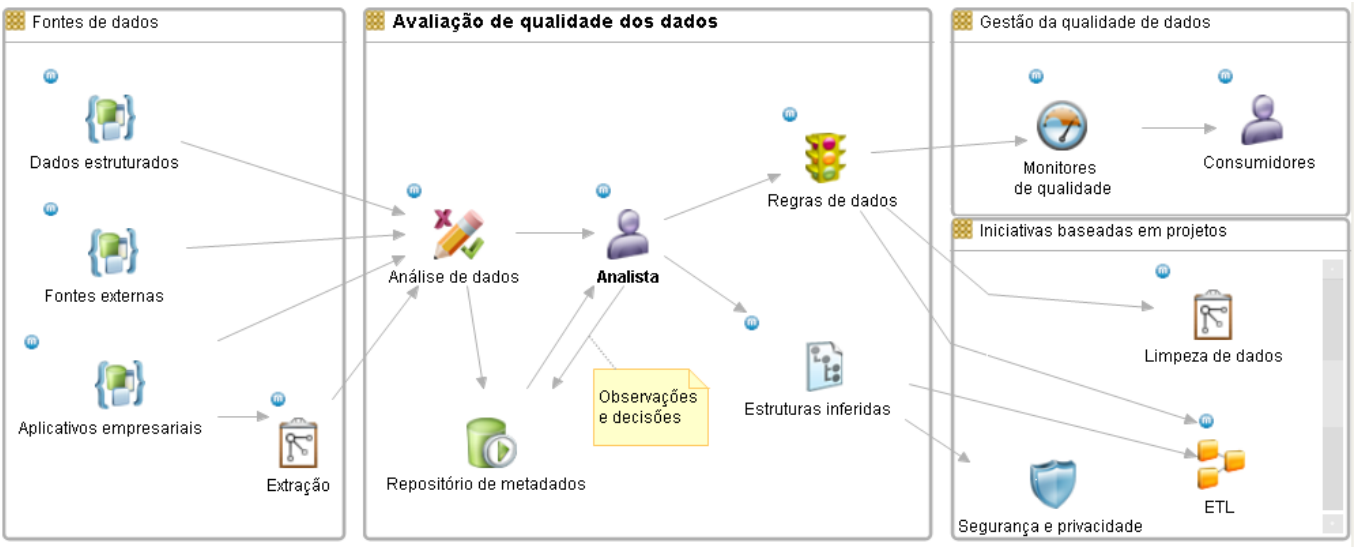

Figura 2: Cenário para uma avaliação de qualidade dos dados

A segunda etapa: uma plataforma abrangente de avaliação de qualidade dos dados

O InfoSphere Information Server suporta recursos de análise e qualidade de dados que permitem a criação de uma avaliação de qualidade de dados. Usando uma arquitetura de múltiplas camadas, serviços comuns, metadados compartilhados e um mecanismo de processamento paralelo, ele oferece uma plataforma comum capaz de analisar uma ampla gama de fontes de dados, processando grandes volumes de dados, armazenando resultados abrangentes e capturando insights do analista em um ambiente protegido e baseado em projeto (ver Figura 3).

Ao usar o InfoSphere Information Server para realizar uma análise essencial, o analista pode descobrir uma ampla gama de problemas em vários domínios, como dados omissos, valores ausentes e chaves não-exclusivas ou duplicadas. O analista pode usar técnicas adicionais para focar em certos tipos de condições, que normalmente podem ser expressas como regras de validação de dados. Essas regras podem ser usadas para testar combinações de valores válidos, fórmulas e agregações corretas ou requisitos de formato complexos, bem como para obter uma avaliação completa de todos os registros e tabelas. Os testes adicionais podem ser relatados, retidos e ter suas tendências avaliadas ao longo do tempo no InfoSphere Information Server.

Os metadados compartilhados obtidos a partir da análise essencial realizada no InfoSphere Information Server ficam diretamente disponíveis para os usuários de outros recursos na plataforma. Modeladores de dados e administradores de banco de dados podem usar as estruturas inferidas e classificações identificadas para estabelecer áreas de teste com a estrutura correta, visando refinar as políticas de privacidade e governança. Desenvolvedores concentrados na transformação ou limpeza de dados podem utilizar as estatísticas e anotações para ajudar a garantir que as rotinas adequadas de limpeza sejam aplicadas aos dados, e eles podem incorporar tabelas de referência geradas a partir da análise. Além disso, os desenvolvedores podem conectar as regras de validação de dados da análise diretamente na limpeza de dados, e em processos de extração, transformação e carregamento (ETL), através de um estágio integrado que aplica as regras em tempo real. Este recurso ajuda a garantir que condições de dados problemáticas e inválidas sejam resolvidas antes que os dados sejam carregados nos ambientes de destino.

O InfoSphere Information Server permite que a empresa se concentre primeiro em um sistema de origem e, então, amplie continuamente a avaliação básica da qualidade de dados em iniciativas por toda a empresa, incluindo limpeza de dados, integração de informações e projetos de governança de dados. Processos de análise, regras de validação de dados e relatórios podem ser

Figura 3: O InfoSphere Information Server é construído em uma base de metadados compartilhados, processamento paralelo e outros serviços.

agendados para serem executados regularmente, fornecendo monitoramento contínuo da qualidade de dados. O insight dos domínios de dados de uma empresa proporcionado pelo InfoSphere Information Server suporta e responde aos desafios inerentes à contínua expansão e aquisição de dados, sistemas e aplicativos que são a base de todos os negócios de uma organização.

Outros pontos de entrada de qualidade de dados suportados pelo InfoSphere Information Server

Além da avaliação de qualidade de dados e da análise e profiling (pesquisa de perfil) das informações, o InfoSphere Information Server suporta outros pontos de entrada para a realização de um programa de qualidade de dados em grande escala, que pode ser apropriado dependendo das prioridades da organização.

Definir uma linguagem de negócios comum

Dificuldades para entender e interpretar dados, determinar quais dados são importantes e gerenciar os dados podem criar obstáculos, visto que os usuários de negócios e TI tentam colaborar na integração eficaz de informações. O problema da inconsistência de definição de negócios entre os ambientes empresariais é geralmente atribuído à ausência de um dicionário de dados que abranja toda a empresa e de um programa de compromisso.

A funcionalidade de glossário de negócios do InfoSphere Information Server ajuda as organizações a criar, gerenciar e compartilhar um vocabulário controlado para toda a empresa. Criar essa linguagem comum entre negócios e TI é um passo fundamental para o alinhamento da tecnologia com os objetivos de negócio. Além de um vocabulário controlado, os sistemas de hierarquia e classificação oferecem um contexto de negócios adicional.

Entender os dados e os relacionamentos de dados

Antes de implementar um programa de governança de informações ou projeto centrado em informações, as organizações devem ter uma imagem completa dos seus dados: quais dados eles possuem, onde estão localizados e como eles se relacionam entre os sistemas. Para a maioria das organizações, o processo de descoberta de dados é manual, exigindo meses de envolvimento humano para descobrir objetos, dados confidenciais, relacionamentos de dados entre fontes e lógica de transformação. O resultado: um processo demorado e propenso a erros, que reduz o tempo de obtenção de valor, gera dúvidas sobre a exatidão dos dados no novo sistema e cria a possibilidade de o novo sistema jamais entrar em operação.

O InfoSphere Information Server oferece uma gama completa de recursos para automatizar o processo de descoberta de dados. Ele aborda uma fonte única de profiling, a análise de sobreposição de dados entre fontes, a descoberta de chaves correspondentes, a geração de protótipos e a realização de testes para consolidação de dados e a descoberta de transformação automatizada. Ele também usa heurísticas e algoritmos sofisticados que automatizam a análise, ajudando as organizações a economizarem tempo e recursos em comparação com a realização das mesmas tarefas utilizando uma solução de geração de perfis.

Limpar, padronizar e combinar informações

Para garantir a qualidade e consistência em tarefas como limpeza de endereços e desduplicação de registros, as organizações precisam de ferramentas que incluam funções de padronização e combinação confiáveis e fáceis de usar, assim como integração de dados, especialmente se várias fontes e/ou múltiplos destinos estão envolvidos. O InfoSphere Information Server permite que as empresas criem e mantenham uma visão exata das entidades de dados mestre, tais como clientes, fornecedores, locais e produtos. Ele também oferece um ambiente de desenvolvimento com um conjunto de recursos poderoso e flexível.

- Oferece um conjunto único de regras de padronização, limpeza, correspondência e sobrevivência para as principais entidades de negócio – executadas em batch, em tempo real ou como um web service
- Combina dados usando algoritmos de probabilidade destinados a garantir que as informações necessárias para administrar uma empresa sejam precisas, completas e confiáveis
- Processa dados globais em uma plataforma paralela altamente escalável para desempenho ideal em ambientes exigentes
- Facilita a criação e manutenção de dados mestres de alta qualidade, gerando benefícios em várias iniciativas empresariais críticas, incluindo gerenciamento de dados mestres e governança de dados
- Traz os recursos de qualidade de dados para as situações de integração de dados através da integração contínua do fluxo de dados
- Emprega uma interface do usuário intuitiva do tipo “crie à medida que as ideias forem surgindo”

Manter a origem dos dados

O InfoSphere Information Server é projetado para integrar e enriquecer as informações entre sistemas de origem distintos. Utilizando uma camada de repositório de metadados ativa e compartilhada, ele suporta uma gama completa de atividades de integração e funções de usuário com os princípios de colaboração e reutilização. Esses artefatos incluem metadados técnicos sobre as várias fontes de informação, metadados de negócios que descrevem o significado dos negócios e o uso das informações e metadados operacionais que descrevem o que acontece no processo de integração.

A plataforma InfoSphere Information Server oferece uma interface poderosa para o gerenciamento de metadados que suporta não apenas os metadados do próprio InfoSphere, como também outros metadados que desempenham papéis críticos nos processos de integração de dados. A plataforma oferece uma visão holística e centralizada de todo o cenário de processos de integração de dados, com visibilidade das transformações de dados que operam dentro e fora do InfoSphere Information Server. Essa visibilidade permite que as organizações rastreiem as informações até as fontes originais, estabelecendo confiança e segurança nas informações recebidas – algo especialmente crítico em situações de descoberta envolvendo assuntos legais ou de auditoria.

Uma solução de qualidade de dados flexível e escalável

As decisões de negócios são cada vez mais baseadas em informações de clientes, parceiros e operacionais. Para bons resultados, estas decisões devem ser baseadas em dados de alta qualidade. Estabelecer prioridades claras de negócios com antecedência, apoiadas por um programa de qualidade de dados abrangente, permite que as organizações se concentrem no planejamento de investimentos. O IBM InfoSphere Information Server oferece a flexibilidade para lidar com os problemas atuais e de alta prioridade na qualidade de dados, enquanto dimensiona facilmente seus recursos para suportar os requisitos futuros. Para as organizações que acabam de dar os primeiros passos em iniciativas de qualidade de dados e governança de informações, ele oferece total flexibilidade, em uma plataforma de integração de dados comum e abrangente.

Para obter mais informações

Para saber mais sobre a qualidade de dados e o papel que ela representa na sua estratégia de governança de informações, entre em contato com o seu representante de vendas IBM ou Parceiro de Negócios IBM, ou acesse:

- **ibm.com**/software/data/integration/capabilities/cleanse.html
- **ibm.com**/software/data/db2imstools/solutions/data-governance.html

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produzido nos Estados Unidos da América
Março de 2012

[1] "Mãe com filho pequeno morre de câncer aos 38 anos após hospital enviar cartas ao endereço errado devido a erro de digitação" The Daily Mail, 14 de março de 2011 (em inglês). www.dailymail.co.uk/news/article-1366056/Mistyped-address-leaves-mother-dead-cancer-son-8-orphan.html

Por favor, recicle

IMW14623-BRPT-00