Data Management
Whitepaper
Series

Gestão de Dados para Empresas

Visão geral sobre o mundo dos dados

eBook – Master Data Management
www.datamotion.com.br
Maio 2017

Introdução

No mundo atual, o volume de dados cresce exponencialmente em todas as áreas, desde dados pessoais até dados empresariais e globais. Por isso torna-se extremamente importante entender e organizar conjuntos de dados. Disciplinas como integração, migração e sincronização de dados, assim como *business intelligence*, etc., permitem realizar isso. Este texto de Informações sobre Integração de Dados procura explicar e descrever essas ideias e conceitos dentro do complexo cenário de gestão de dados.

Informações sobre Integração de Dados

Integração de Dados

Integração de dados envolve a combinação de dados a partir de várias fontes distintas, armazenando-os com o uso de várias tecnologias e proporcionando uma visão unificada. A integração de dados é cada vez mais importante para consolidar as aplicações de uma empresa ou os sistemas de duas empresas distintas, de modo a apresentar uma visão unificada dos recursos de dados da companhia.

Migração de Dados

Migração de Dados é o processo de transferência de dados de um sistema para outro em que o armazenamento, o banco de dados ou o aplicativo são diferentes. Em relação ao processo de ETL (extração, transformação, carga), a migração de dados sempre exige pelo menos as etapas de Extração e Carga.

Quanto mais os volumes de dados aumentam, mais complexas as regras de negócio se tornam.

Sincronização de Dados

Sincronização de dados é o processo que estabelece a consistência entre sistemas, realizando atualizações contínuas para mantê-la. A palavra “contínuas” deve ser enfatizada, pois a sincronização de dados não deve ser realizada de forma pontual.

ETL

ETL vem da área de *Data Warehousing* e significa Extração-Transformação-Carga de dados. ETL abrange o processo de como os dados são carregados a partir de um sistema de origem para o *data warehouse*.

Business Intelligence

Business Intelligence (BI) é um conjunto de ferramentas para transformar dados brutos em informações úteis que possam ser usadas na tomada de decisões. *Business Intelligence* oferece funcionalidade de relatórios, ferramentas para identificar *clusters* de dados, suporte para técnicas de mineração de dados (*data mining*), gestão de desempenho de negócios e análise preditiva.

Master Data Management

Master Data Management (MDM) representa um conjunto de ferramentas e processos utilizados por uma organização para gerenciar seus dados não transacionais de forma consistente.

Integração de Dados

Integração de dados envolve a combinação de dados a partir de várias fontes distintas, armazenando-os com o uso de várias tecnologias e proporcionando uma visão unificada. A integração de dados é cada vez mais importante para consolidar as aplicações de uma empresa ou os sistemas de duas empresas distintas, de modo a apresentar uma visão unificada dos recursos de dados da companhia. Essa última iniciativa muitas vezes é chamada da *Data Warehouse*.

Talvez a mais conhecida implementação de integração de dados seja a elaboração do *Data Warehouse* de uma organização. A empresa pode realizar diversas análises com base nos dados contidos no *Data Warehouse* que não seriam possíveis trabalhando diretamente com os dados no sistema de origem. O motivo é que os sistemas de origem podem não conter dados correspondentes – mesmo que os dados tenham os mesmos nomes, podem se referir a entidades diferentes.

Áreas de Integração de Dados

Integração de dados é um termo que abrange diversas subáreas distintas, tais como:

Data Warehousing

Migração de Dados

Integração de aplicações/informações corporativas

Master Data Management

Este e-book foca no processo de integração de dados. Informações mais detalhadas sobre as áreas mencionadas acima podem ser encontradas em artigos relacionados.

Desafios da Integração de Dados

À primeira vista, o maior desafio é a implementação técnica da integração de dados de fontes diversas, muitas vezes incompatíveis. Entretanto, a integração de dados como um todo apresenta um desafio muito maior. Ela necessariamente inclui as seguintes fases:

Design

A iniciativa de integração de dados em uma companhia deve ser liderada pela área de negócios e não pela área de TI. Deve haver um campeão que entenda os ativos de dados da organização e possa liderar a discussão sobre a iniciativa de longo prazo para integração dos dados, tornando-a consistente, bem sucedida e benéfica.

Análise de requisitos, por que a integração de dados está sendo feita, quais são os objetivos e as entregas a serem feitas. Que sistemas servirão como fontes de dados? Todos os dados para atender aos requisitos estão disponíveis? Quais são as regras de negócios? Qual é o modelo de suporte e o SLA?

Tratando isoladamente a câmada da lógica de negócios é possível se obter uma vantagem clara.

Análise dos sistemas origem, i.e. quais são as opções para extrair os dados dos sistemas (notificação de atualização, extrações incrementais, extrações totais), qual é a frequência necessária/disponível das extrações? Qual é a qualidade dos dados? Os campos de dados necessários estão populados de forma correta e consistente? A documentação está disponível? Quais são os volumes de dados sendo processados? Quem é o dono do sistema?

Quais são os requisitos não funcionais, tais como janela de processamento de dados, tempo de resposta do sistema, número estimado de usuários (concorrentes), política de segurança de dados, política de backup.

Qual é o modelo de suporte para o novo sistema? Quais são os requisitos de SLA?

Finalmente, quem será o dono do sistema e qual será a verba disponível para as despesas de manutenção e atualizações?

Os resultados das etapas anteriores devem ser registrados em um documento SRS (especificação de requisitos de software), confirmado e aprovado por todas as partes que participarão do projeto de integração de dados.

Implementação

Com base no conteúdo exposto, deve-se realizar um estudo de viabilidade para selecionar as ferramentas para implementação do sistema de integração de dados. Pequenas companhias e organizações que iniciam com o *Data Warehousing* precisam decidir quais ferramentas irão utilizar para implementar a solução. O processo é mais simples em empresas maiores ou aquelas que já iniciaram outros projetos de integração de dados, pois já possuem experiência e podem ampliar o sistema atual ou utilizar o conhecimento existente na organização para implementar o novo sistema de maneira mais eficiente. Existem casos, entretanto, em que o uso de uma nova plataforma ou tecnologia mais adequada à situação proporciona maior eficiência ao sistema do que permanecer com os padrões existentes na companhia. Por exemplo, encontrar uma ferramenta mais adequada e que proporcione melhor escalabilidade para crescimento/expansão futura; uma solução que reduza o custo de implementação ou suporte; redução dos custos de licenciamento; migração do sistema para uma nova e mais moderna plataforma, etc.

Testes

Juntamente com a implementação, é essencial realizar testes adequados para garantir que os dados unificados estejam corretos, completos e atualizados.

As áreas de TI e de negócios devem participar dos testes para assegurar que os resultados estejam conforme esperado e exigido. Portanto, os testes devem incluir pelo menos um teste de estresse de desempenho (PST), um teste de aceitação técnica (TAT) e de aceitação de usuário (UAT).

Técnicas de Integração de Dados

Há diversos níveis organizacionais nos quais a integração pode ser realizada. À medida que descemos no nível organizacional, o nível de automatização da integração aumenta.

Integração Manual ou Interface Comum de Usuário – os usuários trabalham com toda a informação relevante acessando todos os sistemas origem ou interfaces via web. Não existe uma visão unificada dos dados.

Integração Baseada em Aplicativos – exige que os aplicativos implementem individualmente todos os recursos de integração. Esta abordagem só é gerenciável enquanto o número de aplicativos for muito limitado.

Integração de Dados por *Middleware* – transfere a lógica de integração dos aplicativos individuais para uma camada de *middleware*. Ainda que a lógica de integração não seja mais implementada nos aplicativos, ainda há a necessidade de os aplicativos participarem parcialmente do processo de integração de dados.

Acesso Uniforme aos Dados ou Integração Virtual – deixa os dados nos sistemas origem e define um conjunto de *views* para apresentar e acessar uma visão unificada do cliente em toda a organização. Por exemplo, quando um usuário acessa a informação do cliente, os dados específicos do cliente são obtidos do respectivo sistema de forma transparente. Os principais benefícios

da integração virtual são a latência praticamente inexistente na propagação das atualizações do sistema de origem para a visão consolidada, sem necessidade de armazenamento separado para dados consolidados. Entretanto, as desvantagens incluem a possibilidade limitada de gestão do histórico e versões dos dados, limitação em aplicar o método apenas a fontes de dados 'similares' (p.ex. no mesmo tipo de banco de dados) e o fato de que o acesso aos dados do usuário gera uma carga adicional nos sistemas de origem que eles podem não estar preparados para aceitar.

Armazenamento Comum dos Dados ou Integração Física dos Dados – em geral significa a criação de um novo sistema que mantém uma cópia dos dados provenientes dos sistemas de origem para que sejam armazenados e processados de forma independente do sistema original. O exemplo mais conhecido desta abordagem é chamado de *Data Warehouse* (DW). Os benefícios incluem a gestão da versão dos dados, a combinação de dados de diferentes fontes (*mainframes*, bancos de dados, *flat files*, etc.). A integração física, entretanto, exige um sistema separado para lidar com os grandes volumes de dados.

Migração de Dados

Migração de Dados é o processo de transferência de dados de um sistema para outro em que o armazenamento, o banco de dados ou o aplicativo são diferentes. Em relação ao processo de ETL (extração, transformação, carga), a migração de dados sempre exige pelo menos as etapas de Extração e Carga.

A migração de dados normalmente ocorre durante uma atualização de hardware existente ou durante a transferência para um sistema totalmente novo. Exemplos: migração de ou para plataforma de hardware; atualização de banco de dados ou migração para novo software; fusão de companhias onde os sistemas paralelos de ambas precisam ser combinados em um só. Há três opções principais para realizar a migração de dados:

Consolidar os sistemas de duas empresas para formar um novo sistema.

Migrar um dos sistemas para o outro.

Deixar os sistemas como estão, mas criar uma visão comum sobre eles – um *Data Warehouse*.

Desafios da Migração de Dados

Vamos descrever os desafios da migração de dados em mais detalhe. A migração de dados pode ser um processo simples, mas há desafios que podem ser encontrados na implementação.

Migração de Armazenamento

A migração de armazenamento pode ser realizada de forma transparente à aplicação, desde que esta utilize apenas interfaces genéricas para acessar os dados. Na maioria dos sistemas isso não constitui problema. Entretanto, é preciso ter atenção especial com aplicativos antigos que rodam em sistemas proprietários. Muitas vezes, o código-fonte do aplicativo não está disponível, e o fornecedor pode ter deixado de existir. Em tais casos, a migração do armazenamento pode ser complicada e deve ser corretamente testada antes de a solução ser colocada em produção.

Migração de Banco de Dados

A migração do banco de dados é relativamente simples quando ele é usado apenas como armazenamento. A migração "apenas" envolve a mudança dos dados de um banco de dados para outro. Entretanto, essa tarefa pode ser difícil. Os principais problemas encontrados incluem:

Tipos de dados diferentes (números, datas, sub-registros)

Conjuntos de dados diferentes (codificação)

A questão dos tipos de dados diferente pode ser resolvida facilmente pela aproximação do tipo de dados mais similar no banco de dados de destino, a fim de manter a integridade dos dados. Se o banco de dados de origem suporta formatos complexos de dados (como sub-registros), mas o banco de dados destino não, será necessário alterar os aplicativos que utilizam o banco de dados. Da mesma forma, se o banco de dados de origem suporta codificações diferentes em cada coluna para uma tabela específica, mas o banco de dados destino não, será necessário fazer uma revisão completa dos aplicativos que utilizam o banco de dados.

Quando um banco de dados não é utilizado apenas como armazenamento, mas também para representar lógica de negócio na forma de *procedures* armazenadas e *triggers*, é preciso tomar muito cuidado ao se realizar o estudo de viabilidade da migração. Novamente, se o banco de dados de destino não oferece suporte a alguns recursos, pode ser necessário fazer alterações nos aplicativos ou no *middleware*.

Ferramentas de ETL são indicadas para a tarefa de migrar dados entre bancos de dados, particularmente quando a migração ocorre entre bases que não possuem conexão direta ou uma interface implementada.

Migração de Aplicativos

Você pode perceber que o processo envolvido nesses dois últimos casos é relativamente simples. Entretanto, isso é muito raro quando se trata da migração de aplicativos. O motivo é que os aplicativos, mesmo quando desenvolvidos pelo mesmo fornecedor, utilizam estruturas e formatos muito diferentes para armazenar os dados, o que torna impossível uma simples transferência de dados. Como a etapa de transformação de dados nem sempre é simples, é preciso aplicar o processo completo de ETL. É claro que a migração de aplicativos normalmente inclui a migração do armazenamento e do banco de dados. A vantagem de uma ferramenta de ETL nesse caso está em sua conectividade a diferentes fontes/destinos de dados.

Pode ocorrer dificuldade ao migrar dados de sistemas de *mainframe* ou aplicativos que utilizam armazenamento proprietário. Sistemas de mainframe utilizam formatos de armazenamento baseados em registros. Esses formatos são de fácil manuseio, mas muitas vezes o formato do armazenamento de dados no *mainframe* inclui otimizações que podem complicar a migração de dados. As otimizações típicas incluem armazenamento numérico em formato BCD (*binary coded decimal*), valores negativos/positivos armazenados de forma não padronizada ou o armazenamento de sub-registros mutuamente exclusivos dentro do mesmo registro. Consideremos, por exemplo, o *data warehouse* de uma biblioteca. Há dois tipos de publicação – livros e artigos.

Uma publicação pode ser um livro ou um artigo, mas não ambos. Há diferentes tipos de informação armazenados para os livros e os artigos. As informações sobre os livros e artigos são mutuamente exclusivas. Portanto, no armazenamento, os dados possuem um formato diferente de sub-registro para um livro e para um artigo, apesar de ocuparem o mesmo espaço. Mesmo assim, os dados são armazenados com uma codificação mais ou menos padrão. Por outro lado, sistemas de armazenamento proprietário tornam a etapa de extração de dados ainda mais complicada. Em ambos os casos, a maneira mais eficiente de extrair os dados é realizar a extração no próprio sistema de origem e, então, convertê-los para um formato que possa ser impresso e posteriormente analisado com ferramentas padrão.

Codificação de Caracteres

A maioria dos sistemas desenvolvidos em plataformas de PC utilizam codificação ASCII ou uma extensão nacional baseada em ASCII. A extensão mais recente é a UTF-8, que mantém o mapeamento ASCII para caracteres alfabéticos e numéricos, mas permite o armazenamento de caracteres para a maioria dos alfabetos nacionais, incluindo Chinês, Japonês e Russo. Sistemas de *mainframe* em geral utilizam codificação EBCDIC, que é incompatível com ASCII, de forma que uma conversão torna-se necessária para a visualização dos dados. Ferramentas de ETL necessitam suportar as conversões de codificação, incluindo EBCDIC.

Sincronização de Dados

Sincronização de dados é o processo que estabelece a consistência entre sistemas, realizando atualizações contínuas para mantê-la. A palavra "contínuas" deve ser enfatizada, pois a sincronização de dados não deve ser realizada de forma pontual. Realmente é um processo que necessita de planejamento, responsabilidade, gestão, programação e controle.

Motivação

Vamos apresentar dois cenários nos quais a sincronização de dados é essencial para uma organização.

Em qualquer empresa, em geral há no mínimo 10 sistemas que compartilham os mesmos dados – dados de clientes, produtos, funcionários – incluindo sistemas de suporte ao cliente, cobrança, faturamento, etc. Para que o processo de manufatura da companhia seja auditável, cada atividade deve ser adequadamente registrada. Por exemplo, se a companhia fabrica carros, para cada carro é preciso registrar os componentes utilizados, seus números de série, os números de lotes, os fornecedores, o ID do funcionário que montou o componente; no fim do processo é preciso registrar para quem o carro foi vendido e o histórico de serviço do carro, incluindo a concessionária e até mesmo os técnicos e peças utilizadas. Cada um dos sistemas de produção da companhia, entretanto, contém apenas parte da informação – os dados que a companhia necessita para funcionar. Por exemplo, o sistema que registra a montagem do carro tem a informação sobre os funcionários; da mesma forma, ele precisa acessar a informação sobre fornecedores e peças em estoque. Ainda que diversos aplicativos/sistemas utilizem os mesmos dados, estes são capturados por um único aplicativo. Os dados então devem ser sincronizados para os outros sistemas.

> Data Quality a cada dia se tornar mais importante nos processos de integração de dados.

Com o advento da internet e com o crescimento dos negócios internacionais, muitas companhias decidiram distribuir seus sistemas geograficamente para reduzir a latência e o custo da rede e para aumentar a confiabilidade (ao reduzir o risco de, p.ex., um desastre natural afetar o local). Os sistemas distribuídos, entretanto, precisam ter os mesmos dados, ainda que estes sejam modificados em diversos locais de maneira paralela. Os dados precisam ser sincronizados em todos os locais.

Planejamento de Processos

Os requisitos da sincronização de dados devem ser determinados na fase de planejamento, que precisa abranger conteúdo de dados, formatos, carga inicial e frequência das atualizações. Requisitos não funcionais, como desempenho, tempo e segurança também devem ser abordados.

Responsabilidade

Mesmo que a ideia da sincronização de dados venha do departamento de TI, um responsável ou "campeão" da área de negócios da companhia são necessários para que a iniciativa tenha continuidade. Afinal, a área de negócios é que vai se beneficiar da iniciativa de sincronização de dados.

Programação

A programação e frequência das atualizações é um dos itens que precisam ser analisados durante a fase inicial de planejamento. Os requisitos muitas vezes mudam nessa fase e o cronograma de atualizações precisa ser ajustado. Obviamente, a granularidade do cronograma de atualizações/sincronizações não pode ser menor do que o sistema de origem é capaz de fornecer. Mas a programação também deve levar em conta aspectos de desempenho.

Monitoramento

O processo de sincronização deve ser monitorado para avaliar se a programação e a frequência de atualização atendem às necessidades da companhia.

Da perspectiva técnica, a sincronização pode ser implementada em qualquer nível:

Nível de sistema/aplicativo

Nível de arquivo (pode até incluir controle de versões)

Nível de sincronização de registro

Complexidade dos Formatos de Dados

À medida que a organização cresce e evolui, novos sistemas de diferentes fornecedores são implementados. Os formatos de dados para funcionários, produtos, fornecedores e clientes variam entre os diferentes segmentos de mercado, o que torna necessário não só construir uma interface simples entre os dois aplicativos (origem e destino), mas também realizar a transformação dos dados à medida que são passados para o aplicativo de destino. Os formatos de dados variam desde formatos proprietários até texto simples ou XML. Alguns dos aplicativos oferecem APIs para transferir os dados diretamente. As ferramentas de ETL podem ajudar neste caso.

Tempo real

Hoje em dia os sistemas precisam atuar em tempo real. Clientes querem ver o status de seus pedidos no comércio eletrônico; o status de sua entrega – o acompanhamento em tempo real do pacote; o saldo de sua conta bancária, etc. As organizações também precisam atualizar seus sistemas em tempo real para

assegurar a eficiência do processo de manufatura, p.ex. fazer pedidos de materiais quando os estoques estiverem baixos, sincronizar os pedidos de clientes com o processo de manufatura, etc. Há milhares de exemplos da vida real em que o processamento em tempo real se torna uma vantagem ou uma necessidade para o sucesso e a competitividade.

O principal desafio com a sincronização de dados em tempo real está em trabalhar com sistemas que não oferecem uma API para identificar as mudanças. Nesses casos, desempenho pode ser o fator limitador.

Segurança

Sistemas diversos podem ter políticas diferentes para assegurar os níveis de segurança de dados e de acesso. Mesmo que a segurança seja mantida corretamente no sistema de origem que captura os dados, a segurança e privilégios de acesso à informação precisam ser assegurados também no sistema de destino, a fim de prevenir qualquer uso indevido da informação. Isso se torna um problema principalmente em se tratando de informações pessoais ou informações confidenciais sob um acordo de confidencialidade (NDA). Quaisquer resultados intermediários da transferência de dados, assim como a própria transferência de dados devem ser criptografados.

Qualidade de Dados

As melhores práticas para gestão e melhoria da qualidade dos dados indicam mantê-los em um único local e compartilhá-los com outros aplicativos. Isso previne inconsistências nos dados, causadas pela atualização do mesmo dado em mais de um sistema.

Desempenho

O processo de sincronização de dados possui, basicamente, cinco fases:

Extração de dados a partir do sistema de origem/mestre

Transferência de dados

Transformação de Dados

Transferência de dados

Carga de dados no sistema destino

No caso de muitos dados, cada uma dessas etapas pode ter um impacto sobre o desempenho. Portanto, a sincronização deve ser planejada cuidadosamente, a fim de evitar qualquer impacto negativo, p.ex. durante horários de pico.

Manutenção

Como qualquer outro processo, a sincronização deve ser monitorada para assegurar que esteja correndo conforme programado e corretamente tratando quaisquer erros que surjam durante o processo de sincronização, como registros rejeitados ou dados malformados.

ETL

ETL vem da área de *Data Warehousing* e significa Extração-Transformação-Carga de dados. ETL abrange o processo de como os dados são carregados a partir de um sistema de origem para o data warehouse. Atualmente, ETL inclui uma etapa de limpeza em separado. A sequência assim se torna Extrair – Limpar – Transformar – Carregar. Vamos descrever rapidamente cada etapa do processo de ETL.

Extração

A etapa de Extração abrange a extração de dados do sistema origem, disponibilizando-os para processamento posterior. O principal objetivo da etapa de extração é recuperar todos os dados necessários do sistema origem, utilizando a menor quantidade de recursos possível. A etapa de extração deve ser planejada de forma a não afetar o sistema origem negativamente em termos de desempenho, tempo de resposta ou qualquer tipo de travamento.

Há diversas maneiras de se realizar a extração:

Notificação de atualização – se o sistema de origem puder apresentar uma notificação de que um registro foi alterado e descrever a alteração, essa é a maneira mais fácil de obter os dados.

Extrato incremental – alguns sistemas podem não oferecer notificações de atualização, mas são capazes de identificar quais registros foram alterados e proporcionar uma extração desses dados. Durante etapas posteriores de ETL, o sistema precisará identificar as alterações e propagá-las. Observe que ao utilizar uma extração diária pode haver problemas em lidar com registros que foram excluídos.

Extração completa – alguns sistemas não conseguem identificar quais dados foram alterados, portanto uma extração completa é a única maneira de obter os dados do sistema. A extração completa exige que uma cópia da última

extração seja mantida, no mesmo formato, para que as alterações possam ser identificadas. A extração completa trata corretamente os registros excluídos.

Ao usar extrações incrementais ou completas, a frequência de extração é de extrema importância, especialmente no caso de extrações completas, pois os volumes podem chegar a dezenas de gigabytes.

Limpeza

A etapa de limpeza é uma das mais importantes, pois assegura a qualidade dos dados no *data warehouse*. A limpeza deve executar regras básicas de unificação de dados, tais como:

Tornar únicos os identificadores (categorias de gênero Masculino/Feminino/Desconhecido, M/F/nulo, Homem/Mulher/ND – são todos traduzidos para Masculino/Feminino/Desconhecido)

Converter valores nulos em valores padronizados Não Disponível/Não Fornecido

Converter números de telefone e CEPs para um modelo padrão

Validar campos de endereço, convertendo-os aos nomes corretos, p.ex. Rua/R/R./Rua.

Validar campos de endereço uns contra os outros (Estado/País, Cidade/Estado, Cidade/CEP, Cidade/Rua).

Transformação

A etapa de transformação aplica um conjunto de regras para transformar os dados a partir da origem para o destino. Isso inclui converter quaisquer dados de mensuração para que usem a mesma dimensão (i.e. dimensão conformada) utilizando as mesmas unidades, para que possam ser integrados mais tarde. A etapa de transformação também exige que dados de várias fontes sejam consolidados, gerando dados agregados e chaves substitutas, fazendo classificações, derivando novos valores calculados e aplicando regras avançadas de validação.

Carga

Durante a etapa de carga, é necessário garantir que ela seja realizada corretamente e com a menor quantidade possível de recursos. O objetivo do processo de carga muitas vezes é gerar um banco de dados. Para que o processo de carga seja eficiente, é útil desabilitar quaisquer restrições e índices existentes antes da carga, tornando a habilitá-los apenas após o processo de carga. A integridade referencial deve ser mantida pelo ETL de forma a assegurar a consistência.

Gestão do Processo de ETL

O processo de ETL parece bastante simples. Como qualquer outro processo, existe a possibilidade de o processo de ETL falhar. Isso pode ocorrer pela falta de extrações de um dos sistemas, valores faltando em uma das tabelas de referência, ou simplesmente uma falha de conexão ou de energia. Portanto, é necessário projetar o processo de ETL com previsão de recuperação de falhas.

Staging

Deve ser possível reiniciar pelo menos algumas das fases de forma independente das outras. Por exemplo, se a etapa de transformação falhar, não deve ser necessário reiniciar a etapa de Extração. Isso pode ser assegurado através de uma área temporária (*staging*). *Staging* significa que os dados são simplesmente colocados em uma área (chamada de área temporária, ou "*staging area"*) para que possam ser lidos pela próxima fase de processamento. A *staging area* também pode ser utilizada durante o processo de ETL para armazenar os resultados intermediários do processamento. Isso funciona bem para processos ETL. Entretanto, a *staging area* deve ser acessada apenas pelo processo de carga do ETL. Nunca deve estar disponível para outros processos e especialmente para usuários finais, pois sua função não é apresentar dados para os usuários finais; ela pode conter dados incompletos ou em fase de processamento.

Implementação da Ferramenta de ETL

Quando você está prestes a usar uma ferramenta de ETL, há uma decisão fundamental a ser tomada: a companhia construirá sua própria ferramenta de ETL ou utilizará uma ferramenta existente?

Desenvolver sua própria ferramenta de transformação de dados – normalmente um conjunto de scripts de *shell* – é a abordagem preferida quando há um pequeno número de fontes de dados que se encontram em meios de armazenamento similares. O motivo é que o esforço de implementar a transformação necessária é pequeno, pois a estrutura de dados é similar e a arquitetura de sistemas é comum. Além disso, essa abordagem economiza os custos de licenciamento e não há necessidade de treinar pessoas no uso de uma nova ferramenta. Entretanto, essa solução é perigosa do ponto de visa de TOC. Se as transformações se tornarem mais sofisticadas com o tempo ou houver necessidade de integrar outros sistemas, a complexidade do sistema de ETL aumenta e a facilidade de gerenciamento cai significativamente. Da mesma forma, implementar sua própria ferramenta muitas vezes se parece com reinventar a roda.

Há muitas ferramentas de ETL prontas para uso no mercado. O principal benefício de utilizar ferramentas ETL "de prateleira" é que elas são otimizadas para o processo de ETL, oferecendo conectores às fontes de dados mais comuns, como bancos de dados, *flat files*, sistemas de mainframe, XML, etc. Elas permitem implementar as transformações de dados de maneira fácil e consistente em todas as fontes de dados. Isso inclui filtrar, reformatar, classificar, unificar, integrar, agregar e outras operações prontas para uso. As ferramentas também suportam a programação da etapa de transformação, o controle de versões, o monitoramento e a gestão unificada de metadados. Algumas ferramentas de ETL são até integradas a ferramentas de BI.

Business Intelligence

Business Intelligence (BI) é um conjunto de ferramentas para transformar dados brutos em informações úteis que possam ser usadas na tomada de decisões. *Business Intelligence* oferece funcionalidade de relatórios, ferramentas para identificar clusters de dados, suporte para técnicas de mineração de dados (*data mining*), gestão de desempenho de negócios e análise preditiva.

O objetivo de *Business Intelligence* é apoiar a tomada de decisões. Na verdade, ferramentas de BI são muitas vezes chamadas de Sistemas de Suporte à Decisão (SSD), ou sistemas de suporte baseados em fatos, pois proporcionam aos usuários empresariais as ferramentas para analisar seus dados e extrair informação.

Ferramentas de *Business Intelligence* muitas vezes buscam os dados em *data warehouses*. O motivo é simples: um *data warehouse* já contém dados de diversos sistemas de produção da organização; os dados são limpos, consolidados, padronizados e armazenados em um único local. Desta forma, as ferramentas de BI podem se concentrar em analisar os dados.

Visualização de Dados

Quando dados são armazenados como um conjunto ou matriz de números, eles são precisos, mas difíceis de interpretar. Por exemplo, as vendas estão aumentando, diminuindo ou se mantendo estáveis? Ao analisar mais de uma dimensão dos dados, isso se torna ainda mais difícil. Por isso a visualização dos dados em gráficos é uma forma prática de entender rapidamente como os dados devem ser interpretados.

Data Mining

O *data mining* (mineração de dados) é um método apoiado por computador para revelar relacionamentos entre entidades de dados que antes eram desconhecidos ou haviam passado despercebidos. Técnicas de *data mining* são utilizadas de diversas formas: análise de cesta de compras – a determinação de que produtos são comprados conjuntamente para se promover outros produtos; no segmento bancário, avaliação de risco do cliente é usada para determinar a probabilidade de o cliente pagar o empréstimo, com base em dados históricos; no segmento de seguros, a detecção de fraudes é feita com base em dados históricos e de comportamento; em medicina e saúde, a análise de complicações ou doenças comuns pode ajudar a reduzir o risco de infecções cruzadas.

Relatórios

Uma área onde ferramentas de BI auxiliam aos usuários empresariais é o desenvolvimento, programação e geração de relatórios de desempenho, vendas, balanço e economias. Relatórios gerados por ferramentas de BI reúnem e apresentam informações de forma eficiente para dar suporte aos processos de gestão, planejamento e tomada de decisão. Uma vez que o relatório tenha sido projetado, ele pode ser enviado automaticamente para uma lista de distribuição predefinida, apresentando estatísticas diárias, semanais ou mensais.

Análise de séries temporais, incluindo técnicas preditivas

Quase todos os *data warehouses* e todos os dados empresariais possuem uma dimensão de tempo. Por exemplo: vendas de produtos, chamadas telefônicas, hospitalização de pacientes, etc. É extremamente importante revelar as mudanças no comportamento dos usuários ao longo do tempo, assim como a relação entre produtos e mudanças em contratos de vendas com base em promoções. Com base em dados históricos, também podemos procurar prever tendências ou resultados futuros.

OLAP - *On-line Analytical Processing* (processamento analítico online)

OLAP é mais conhecido pelos cubos OLAP, que oferecem uma visualização multidimensional dos dados. Cubos OLAP mostram dimensões nas arestas do cubo (p.ex. tempo, produto, tipo de cliente, idade do consumidor, etc.). Os valores dentro do cubo representam os fatos medidos (p.ex. valores de contrato, número de produtos vendidos, etc.). O usuário pode navegar nos cubos OLAP usando recursos para fazer *drill-down* (detalhar), *drill-up* (resumir) e *drill-across* (analisar). A funcionalidade de *drill-up* permite que o usuário navegue facilmente para um nível mais alto, de detalhes mais macro. Da mesma forma, o *drill-down* permite ver a informação com mais detalhes. Finalmente, fazer um *drill-across* significa que o usuário pode navegar para outro cubo OLAP para ver os relacionamentos em outras dimensões. Toda a funcionalidade é oferecida em tempo real.

Análise Estatística

Análise estatística utiliza fundamentos da matemática para qualificar a significância e confiabilidade das relações observadas. Os recursos mais interessantes são análise de distribuição, intervalos de confiança (por exemplo, para mudanças em comportamentos, etc.). A análise estatística é utilizada para desenvolver e analisar os resultados do *data mining*.

Master Data Managment

Master Data Management (MDM) representa um conjunto de ferramentas e processos utilizados por uma organização para gerenciar seus dados não transacionais de forma consistente. MDM em geral se aplica a entidades como Consumidor, Cliente, Produto, Conta e dados internos de referência. MDM também possibilita a fácil manutenção da linhagem e histórico dos dados para fins de auditoria.

Problemas

Em uma organização existem diversos sistemas gerenciando os mesmos dados. O papel de MDM é centralizar a gestão dos dados em uma única cópia mestre dos dados, que é então sincronizada para todos os aplicativos que utilizam os dados. Com esta abordagem, ao se referir a um cliente da organização, por exemplo, todos os sistemas se referem ao mesmo cliente.

Há basicamente dois motivos para haver dados duplicados de forma inconsistente:

Os sistemas de produção em uma organização, quando implementados, não foram projetados para ser parte de um grupo maior de sistemas de produção com os quais devem cooperar. Portanto, cada sistema gerencia os dados por conta própria.

As filiais ou departamentos da companhia existem de forma independente, sem forte cooperação com outros departamentos. Por exemplo, o departamento de hipotecas lida com clientes e gerencia contratos de hipoteca. Ao mesmo tempo o departamento de marketing planeja uma promoção para aumentar as vendas de hipotecas. Se os dois departamentos não cooperarem (compartilharem os dados), o departamento de marketing pode oferecer uma hipoteca a um cliente que já possui uma hipoteca. Isso resulta tanto em um desperdício de dinheiro na promoção quanto em um aborrecimento para o cliente.

Aquisições e fusões de companhias são outros exemplos de como uma organização pode obter diversos sistemas em paralelo que gerenciam dados similares ou até duplicados.

Soluções

Para resolver esses problemas, as soluções fundamentais de *Master Data Management* compreendem os seguintes processos:

Identificação de fonte – o "sistema oficial" (*system of record*) precisa ser identificado inicialmente. Se o mesmo registro está armazenado em vários sistemas, aquele que possui a cópia mais relevante (mais válida, atual ou completa) é chamado de "sistema oficial".

Coleta de dados – os dados precisam ser coletados de diversas fontes (pois algumas fontes podem incluir uma nova informação), ao mesmo tempo em que partes que não interessam são descartadas.

Transformação – a etapa de transformação ocorre durante a entrada, quando os dados são convertidos em um formato que possibilite o processamento de MDM, e também durante a saída, quando os registros mestre são distribuídos aos sistemas e aplicativos específicos.

Consolidação de dados – os registros de vários sistemas que representam a mesma entidade física são consolidados em um único registro: um registro mestre. O registro recebe um número de versão para permitir que um mecanismo verifique qual versão do registro está sendo usada em sistemas específicos.

Deduplicação de dados – muitas vezes existem registros distintos nos sistemas da companhia que identificam o mesmo cliente. Por exemplo, o banco pode ter um registro que identifica o cliente e a subsidiária ou departamento de seguros pode ter um banco de dados independente de clientes, contendo um registro para o mesmo cliente. É essencial que esses dois registros sejam deduplicados e mantidos como um único registro mestre.

Detecção de erros – com base em regras e métricas, registros incompletos ou contendo dados inconsistentes devem ser identificados e enviados a seus respectivos donos antes de serem publicados para os aplicativos restantes. Caso sejam fornecidos dados com erros, a credibilidade do MDM da companhia pode ser comprometida.

Correção de dados – relacionado à detecção de erros, esse passo notifica o dono do registro de dados de que há a necessidade de revisar o registro manualmente.

Distribuição/sincronização de dados – os registros mestre são distribuídos aos sistemas na organização. O objetivo é fazer com que todos os sistemas utilizem a mesma versão do registro tão logo possível após a publicação do novo registro.

Processo

Nos parágrafos acima, mencionamos que cada registro de dados deve possuir um dono ou responsável – uma pessoa que entenda os dados e seja responsável por sua manutenção. O responsável deve ser da área de negócios da companhia. O motivo é que apenas uma pessoa da área de negócios pode compreender os dados e tomar decisões sobre consolidação, atualizações, correções e validade dos dados. Por outro lado, o processamento necessário pode ficar a cargo do usuário de negócios através de uma interface de usuário ou sob a responsabilidade de TI.