CURSO DE CIÊNCIA DA COMPUTAÇÃO

Fernando Wendland

INTEGRAÇÃO DE DADOS DE FONTES HETEROGÊNEAS ATRAVÉS DE *MASTER DATA MANAGEMENT*

Santa Cruz do Sul
2016

Fernando Wendland

INTEGRAÇÃO DE DADOS DE FONTES HETEROGÊNEAS ATRAVÉS DE *MASTER DATA MANAGEMENT*

Trabalho de Conclusão II apresentado ao Curso de Ciência da Computação da Universidade de Santa Cruz do Sul para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Me. Eduardo Kroth

Santa Cruz do Sul
2016

AGRADECIMENTOS

Agradeço principalmente à minha mãe Traudi e meu pai Eldon, por terem se preocupado com minhas noites em claro e terem me apoiado, desde o início, com meus estudos e assim tornarem possível essa realização. Vocês são o alicerce disso tudo.

Agradeço ao corpo docente do Curso de Ciência da Computação, que compartilhou comigo muitos de seus conhecimentos no decorrer dos anos em que estive na graduação.

Agradeço também, de forma especial, ao meu orientador, professor Me. Eduardo Kroth, que fez com que esse trabalho fosse possível e me acompanhou de forma excepcional no decorrer do mesmo. O meu agradecimento e gratidão.

Aos meus amigos, que muitas vezes reclamaram da minha ausência, mas também puderam entender o quão importante essa etapa era.

Por fim, agradeço a Deus por tudo ter dado certo e ter assim chegado ao final de mais uma etapa importante da minha vida.

RESUMO

Integrar dados é uma tarefa complexa e minuciosa. Exige que os mesmos sejam extraídos de diversas origens seguindo regras de transformação específicas para serem, em seguida, apresentados homogeneamente. Este trabalho apresenta o desenvolvimento, as características e as funcionalidades de uma ferramenta para integração, através do uso de dados mestres, de dados provenientes de fontes heterogêneas. Focado no *middleware,* a camada intermediária entre essas bases de dados e a interface de apresentação dos dados já globalizados, o *software* desenvolvido tem como objetivo dar ao usuário total dinamicidade sobre como será realizada cada uma das etapas de integração, permitindo, no mapeamento, a definição dos dados mestres vistos como importantes pelo usuário dentro do propósito em que a ferramenta estiver sendo utilizada. Partiu-se do princípio de que quantidades enormes de dados são produzidas diariamente por empresas e organizações. Dentre esse montante de dados, o gerenciamento de dados mestres (MDM) identifica os dados mais importantes. A ferramenta desenvolvida realiza todas as etapas de integração de dados, baseada em técnicas de MDM, além de permitir a normalização dos dados no repositório homogêneo criado, eliminando assim problemas como redundância de informações.

Palavras chave: Integração de bancos de dados heterogêneos, dados mestres, *Master Data Management, Middleware*.

ABSTRACT

Integrate data is a complex and detailed task. It requires to extract data from various sources following specific transformation rules and after be shown homogeneously. This work presents the development, characteristics and functionalities of an integration tool, through the use of master data, of data from heterogeneous sources. Focused on the middleware, the intermediate layer between these databases and the presentation interface of that data already globalized, the developed software has the objective of giving the user total dynamism on how each of the stages of integration will be carried out, allowing, in the mapping, the definition of the master data seen as important by the user within the purpose in which the tool is being used. It was assumed that a lot of data are produced daily by companies and organizations. Within this amount of data, Master Data Management (MDM) identifies the most important data. The developed tool performs all the steps of data integration, based on MDM techniques, besides allowing the normalization of the data in the created heterogeneous repository, eliminating so problems like redundancy of information.

Keywords: Integration of heterogeneous databases, master data, Master Data Management, Middleware.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

BD – Base de Dados

DDL – *Data Definition Language*

DML – *Data Manipulation Language*

DW – *Data Warehouse*

EG – Esquema Global

EL – Esquema Local

ETL – *Extract, Transform, Load*

GD – Governança de Dados

IU – Identificador Universal

MDC – Modelo de Dados Comum

MDM – *Master Data Management*

SGBD – Sistema Gerenciador de Banco de Dados

SQL – *Structured Query Language*

XML – *eXtensible Markup Language*

SUMÁRIO

1. INTRODUÇÃO

(OZSU e VALDURIEZ, 1999) definem que bancos de dados são projetados para atenderem a um único sistema. Porém, é a integração desses bancos de dados de sistemas distintos que ganha cada vez mais força.

Não é de hoje essa necessidade de combinar fontes de dados heterogêneas em uma única interface de consulta. Integrar dados provenientes de fontes distintas, como bases de dados, planilhas EXCEL e arquivos XML, visa fornecer uma visão única e global das informações, evitando redundância e dados dispersos.

Empresas necessitam integrar seus dados por diversas razões. Uma delas é a integração com clientes, fornecedores e parceiros objetivando ampliar o alcance dos seus serviços e produtos (DEGAN, 2005).

Esse processo de integração pode ser facilitado através do MDM (Gerenciamento de Dados Mestres), que identifica as informações mais importantes de uma organização, buscando uma visão única dos dados. Assim, pode-se, por exemplo, com a integração de dados, chegar à conclusão de que Beto Fernandes e Roberto Fernandes são a mesma pessoa. Outra situação possível, que pode ser resolvida com o uso de MDM, é a possibilidade de um mesmo cliente, cadastrado em dois sistemas heterogêneos da empresa, possuir dois endereços diferentes associados a ele, um em cada fonte de dados.

Além disso, MDM permite uma visão completa das relações, em que se podem observar dependências entre dados e informações que antes, em bases de dados separadas, não possuíam relevância.

Porém, sabe-se também que esse processo de integração é dificultado por vários motivos. Um deles é o fato de que um mesmo conceito pode ter sido representado de maneiras totalmente distintas entre duas empresas que estão integrando seus dados. De um lado, por exemplo, lucro pode representar um valor e no outro, pode significar uma quantidade.

São poucas as opções de *softwares* disponíveis que permitem a integração de fontes de dados distintas, principalmente através do gerenciamento de dados mestres, que objetiva também a qualidade dos dados durante o processo de integração. Somente integrar dados, sem analisá-los, pode não trazer uma mudança significativa para a organização, já que informações duplicadas e erradas podem aparecer.

Assim, o objetivo principal da ferramenta desenvolvida é permitir a integração de dados provenientes de fontes heterogêneas, com o uso de MDM, a fim de unificar o acesso às informações, buscando alcançar resultados mais satisfatórios em consultas realizadas, além de mesclar dados que possuem relacionamentos entre si e eliminar dados duplicados.

O gerenciamento de dados mestres como forma de integração de dados de fontes heterogêneas é uma dentre muitas opções de integração que podem ser encontradas. Destaca-se sua utilização pelas vantagens existentes em utilizar um repositório de dados, para consulta, contendo dados organizados e de qualidade. Tal fato propicia melhores resultados em buscas, uma vez que se eliminam variáveis negativas para a análise do usuário, como dúvidas quanto à origem e veridicidade dos dados apresentados, já que um sistema integrado utilizando dados mestres conta apenas com os dados mais importantes para a organização, cautelosamente selecionados pelo analista responsável.

Uma aplicação para integração de dados possui vários ambientes e está mapeada em três principais camadas. A primeira, de mais baixo nível, corresponde à camada inferior onde se encontram as fontes de dados heterogêneas e o ambiente para reconhecimento e carregamento destas fontes de dados. A segunda camada que pode ser identificada em uma ferramenta de integração de dados MDM é a camada intermediária, responsável por receber a definição dos dados mestres de cada uma das fontes de dados e mapear tais dados para um único repositório, além de permitir várias etapas de limpeza e normalização de dados. Por fim, a camada de interface é a que possui um ambiente para a realização de consultas pelo usuário. Nesse ambiente, o usuário define as consultas, baseadas no modelo de dados global MDM.

Há inúmeros artigos e trabalhos relacionados à integração de dados, porém poucos abordam MDM, por se tratar de uma técnica ainda recente e pouco explorada. Portanto, justifica-se o trabalho pela necessidade de combinar dados de fontes heterogêneas, garantindo assim a qualidade dos mesmos, mantendo as informações sempre organizadas. Com isso, tem-se a vantagem de dados padronizados e consistentes.

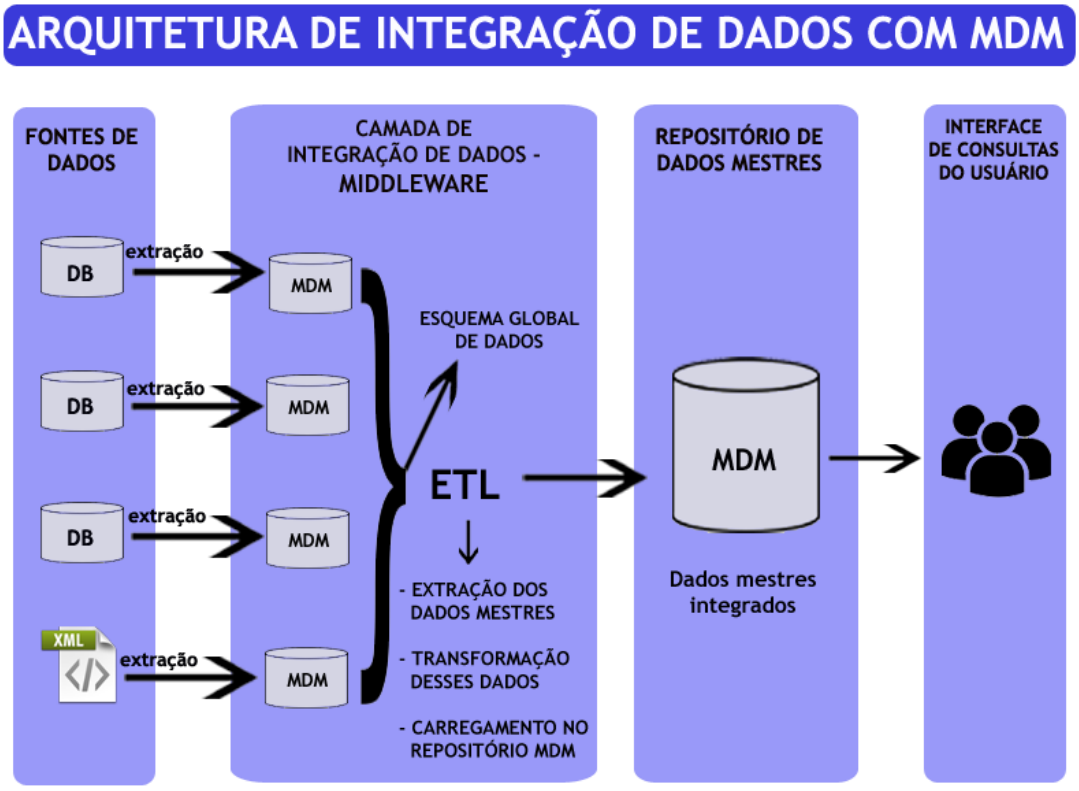
O texto se apresenta organizado em capítulos, sendo o capítulo 2 uma fundamentação teórica de vários temas abordados nesse trabalho, com comentários do autor a respeito de cada um deles. O capítulo 3 busca apresentar trabalhos relacionados ao tema, explanando o que já existe de pesquisa na área. O capítulo 4 objetiva descrever detalhadamente a solução desenvolvida neste trabalho, juntamente com sua validação, sendo o capítulo mais importante. Já o capítulo 5 apresenta propostas de trabalhos futuros que podem dar continuidade ao presente trabalho. Por fim, o capítulo 6 apresenta as conclusões finais obtidas.

2. FUNDAMENTAÇÃO TEÓRICA

A integração de dados, utilizando o gerenciamento de dados mestres, possui várias etapas e componentes envolvidos desde a extração dos dados das fontes de origem até a interface de consultas do usuário.

A Figura 1 exibe toda a arquitetura de integração de dados com os componentes e processos que compõem o modelo de implementação seguido neste trabalho. Podem ser observadas as diferentes fontes de dados, onde os dados mestres de cada repositório são extraídos na camada intermediária, onde ocorre o processo de ETL (*Extraction Transformation Load*), que consiste justamente na extração dos dados das fontes de origem citadas, a transformação desses dados para o posterior carregamento no repositório de dados mestres, onde estão finalmente disponíveis para consultas realizadas pelos usuários, em uma arquitetura que utiliza o gerenciamento de dados mestres como técnica de integração de dados.

Figura 1 - Arquitetura de Integração de Dados com MDM



Fonte: Do Autor (2016)

Baseado nessa arquitetura de integração, este capítulo traz conceitos de técnicas e métodos importantes utilizados no desenvolvimento do *software*, conectando sempre o conceito com a aplicação desenvolvida. Cada seção conta também com comentários do autor a respeito do tópico.

2.1. Middleware

Um *middleware*, também chamado de mediador, é uma ferramenta que se posiciona entre dois (ou mais) *softwares* para permitir que se comuniquem. Sem ele tais sistemas distintos não poderiam se comunicar. Pode-se chamá-lo de conector.

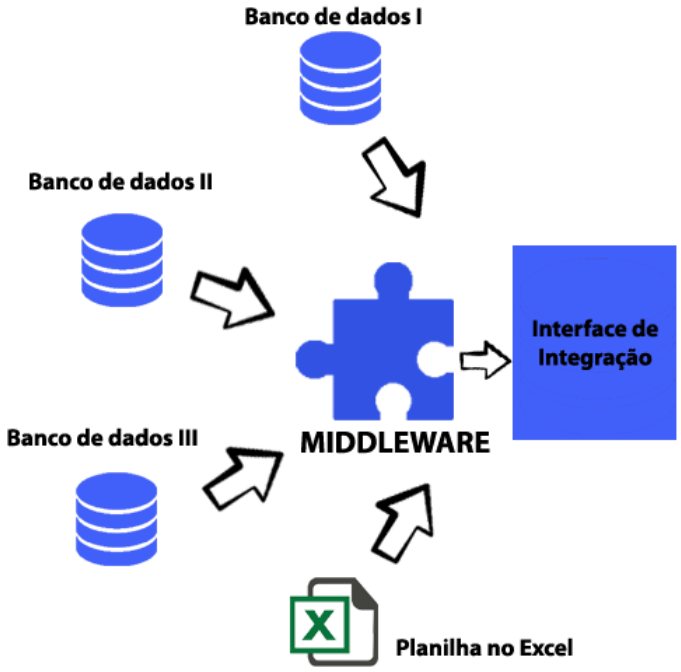
Segundo (BARBOSA, 2001), *middleware* é um componente cuja finalidade é unir processos, fornecendo um conjunto de serviços e visando reduzir a complexidade do desenvolvimento e da execução de uma aplicação.

No contexto de bancos de dados, é a peça responsável por realizar a conexão entre as bases de dados heterogêneas e a interface de integração, que apresenta os dados homogeneamente ao usuário.

Diferentes formas de armazenamento são hoje utilizadas para persistir dados. Essas formas incluem, por exemplo, bancos de dados, planilhas e arquivos no formato XML (*eXtensible Markup Language*). Para possibilitar que informações desses diferentes meios sejam integrados e acessados através de uma mesma interface, o *middleware* é componente que intermedia esse processo.

A Figura 2 exemplifica o posicionamento desse componente diante do processo de integração de dados. Em um cenário com diferentes meios de armazenamento e uma interface de integração desses dados, o *middleware* é a ‘peça’ que aparece na camada intermediária do processo de integração, entre as fontes de dados distintas e a camada de interface do usuário.

Figura 2 - Posicionamento do *Middleware* no processo de integração de dados



Fonte: Do Autor (2016)

Segundo (JOSIFOVSKI e KIRISCH, 2002), um fator de força de uma empresa moderna é a capacidade da mesma armazenar e processar informações. Isso acaba gerando, em grandes empresas, muitos repositórios isolados de dados. E esse número continua a crescer devido a razões organizacionais dessas empresas. De acordo com os autores, a incapacidade de esses sistemas fornecerem ao usuário uma visão unificada dos dados e recursos de toda a organização é um dos principais obstáculos para alcançar o próximo nível de eficiência.

Sabe-se, porém, que há sempre uma solução (ou a busca de uma) para problemas desse contexto. A possibilidade de unir departamentos ou setores, diferentes BD (Bases de Dados) e informações díspares da empresa, para um acesso unificado, de forma rápida, fácil e eficiente, traz como resultado mais do que relatórios e gráficos mais precisos, traz consigo a possibilidade de planejamento prévio para tomadas de decisões. Análise de variáveis de risco, busca de fatores de sucesso no passado, implicações de informação presente na BD X em resultados de consultas na BD Y. Mais dados começam a fazer sentido e deixam de ser apenas dados para virarem, também, informações. Isso tudo possível através da entrada do *middleware* como camada intermediária no processo de homogeneização.

O usuário não envia consultas diretamente para as fontes de dados. Com isso, tira-se do usuário a necessidade de ter que conhecer todas as fontes de dados e ter que interagir com cada uma delas individualmente (BARBOSA, 2001).

Um *Middleware* tem aplicabilidade nas mais diferentes áreas da tecnologia. Nas aplicações de integração de dados, aparece como peça chave fundamental para possibilitar a prospecção do projeto e o alcance do resultado esperado.

A crescente necessidade de interagir com várias plataformas e buscar informações em diferentes meios de armazenamento é suprida com a chegada dessa tecnologia intermediária, uma solução que fica exatamente entre as fontes de armazenamento de dados e a aplicação que o usuário enxerga. É o componente que possibilita o desenvolvimento deste trabalho, realizando todos os processos internos de mapeamento das tabelas e colunas das fontes de dados da camada inferior para o repositório de dados mestres criado, deixando para o usuário somente a tarefa de definir o que será feito, e não como será feito.

A apresentação de forma homogênea ocorre apenas na interface com a qual o usuário interage. Os dados continuam separados, cada um em seu real meio de acesso, assim como foram armazenados quando criados. A homogeneização acontece nessa camada intermediária, não exigindo do usuário nenhum conhecimento extra e detalhado de cada fonte de dados isolada.

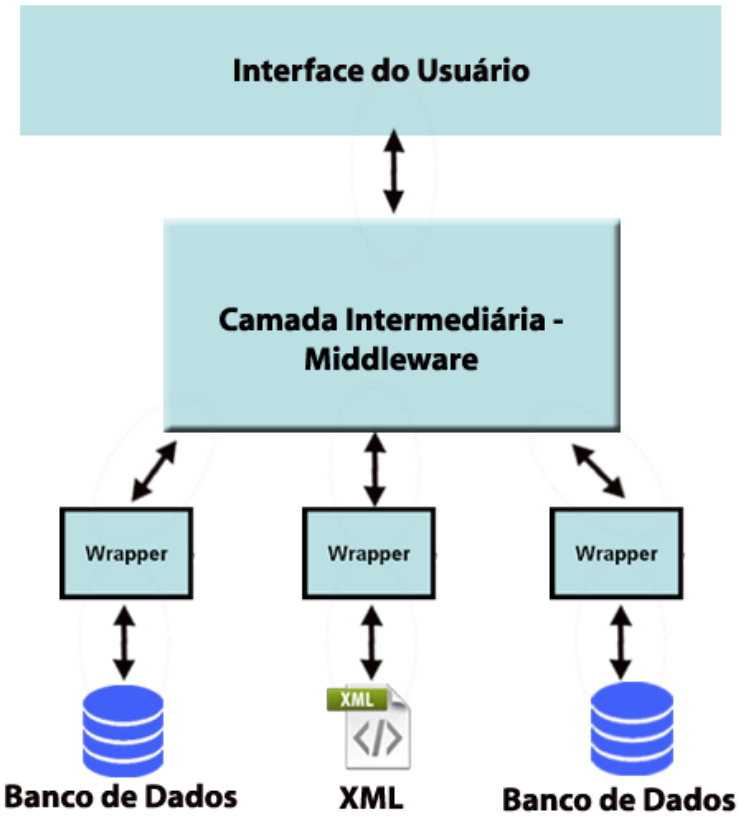
2.1.1. *Wrappers* ou tradutores

Wrappers são componentes responsáveis por converter/traduzir dados. Um *middleware* solicita ao *wrapper* consultas em uma linguagem de consulta comum, como SQL (*Structured Query Language*), e o mesmo converte esta consulta para uma linguagem de consulta suportada pela BD ao qual está ligado. Em seguida, recebe o resultado dessa consulta e o converte para um formato suportado pelo mediador (FERRANDIN, 2002).

Assim, pode-se dizer que é responsável por traduzir o modelo de dados de cada fonte heterogênea para o modelo de dados global, utilizado pela *middleware.*

A Figura 3 a seguir ilustra a arquitetura de um *middleware* com o auxílio de *wrappers*. Como se observa na imagem apresentada, a função principal do tradutor é fazer o meio de campo entre as fontes de dados e o *middleware*, traduzindo as consultas recebidas.

Figura 3 - *Wrappers* em um sistema de integração de dados



Fonte: Do Autor (2016)

2.1.2. Comentários do Autor

O *middleware* é uma tecnologia de extrema importância para a realização deste trabalho e o subsequente desenvolvimento prático do mesmo.

Como já citado anteriormente, integrar dados é hoje cada vez mais um requisito de instituições de quaisquer segmentos. Tem se tornado cada vez mais comum também a integração de informações entre diferentes empresas, principalmente através da *Web*, onde uma pesquisa em determinado serviço ou plataforma pode buscar dados em várias bases de dados para trazer uma resposta mais coerente e completa ao usuário. Essas bases de dados não precisam estar centralizadas, apenas integradas.

A construção de um *middleware* capaz de lidar bem com essas necessidades de integrar meios de armazenamentos, sejam eles arquivos nos mais diferentes formatos ou então bancos de dados, estejam em uma empresa privada ou pública, armazenados de forma local ou *online*, com certeza é a chave para uma homogeneização completa desses dados.

2.2. Integração de Esquemas

Um paradigma para acesso integrado a bancos de dados heterogêneos é o mapeamento das diferentes representações de uma entidade no mundo real, possibilitando que os usuários tenham acesso não apenas ao objeto, mas que também recebam as informações disponíveis na forma mais adequada (KANTORSKI, 1999).

Nem sempre as organizações possuem a mesma estrutura de dados, já que muitas vezes bancos de dados distintos são modelados por pessoas diferentes, em períodos diferentes, com necessidades diferentes.

Quando o processo de integração cede lugar ao mapeamento das diferentes representações de uma mesma entidade do mundo real, torna-se possível a visualização global das diferentes formas modeladas pelos diferentes bancos de dados autônomos (RIBEIRO, 1995).

Este processo, segundo (GOIS e ZAUPA, 2010), é fundamental para possibilitar a integração de esquemas criados de forma independente e de BD diferentes, uma vez que o principal problema encontrado no processo de integração de dados é o transporte dos dados armazenados de um esquema para o outro.

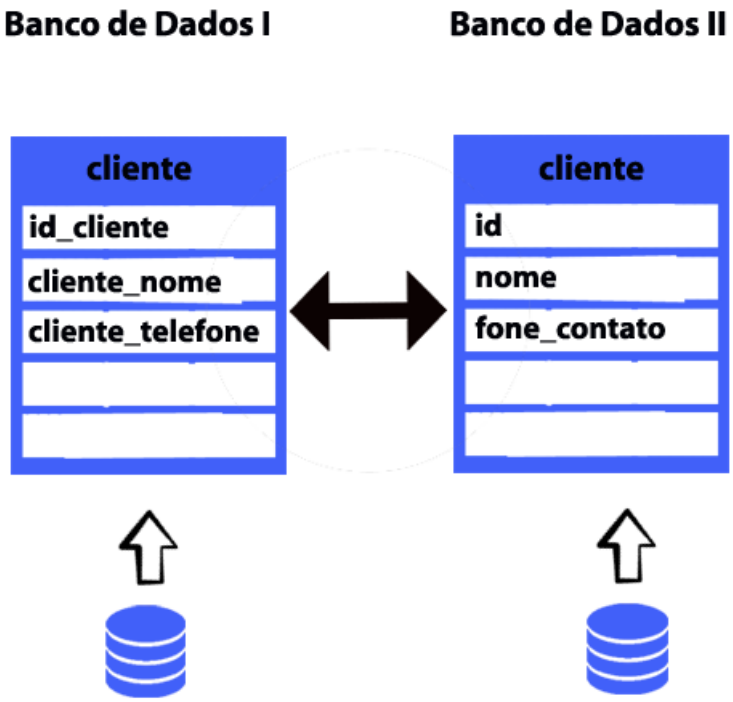
Assim, a integração de dados tem como requisito fundamental a integração de todos os esquemas dos bancos de dados, que precisam estar definidos em um mesmo modelo de dados. Surge com isso uma das maiores dificuldades desse processo de integração, já que na maioria das vezes, esses esquemas estão desenhados de maneiras distintas entre si.

Essa integração de esquemas é também chamada de casamento de esquemas (*schema matching*), que consiste em avaliar os atributos das tabelas dos bancos de dados, a fim de agrupar colunas comuns que, mesmo estruturas em modelos diferentes, possuem o mesmo tipo de informação.

Para ilustrar, a Figura 4 traz um exemplo de casamento de esquemas entre duas tabelas de bancos de dados diferentes. Na imagem pode-se observar que as tabelas possuem as informações mapeadas de formas diferentes, mas ambas apresentam as mesmas informações a respeito do cliente.

Assim, se identifica o casamento de esquemas nos três atributos apresentados: **id_cliente** da BD 1 = **id** da BD 2; **cliente_nome** da BD 1 = **nome** da BD 2; **cliente_telefone** da BD 1 = **fone_contato** da BD 2.

Figura 4 - Demonstração de casamento de esquemas entre tabelas de bancos de dados distintos



Fonte: Do Autor (2016)

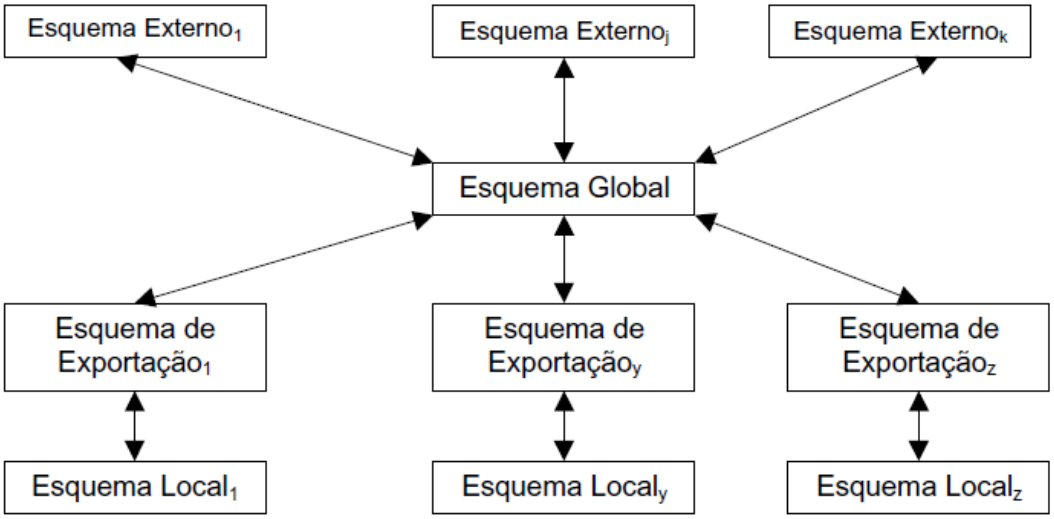
Dessa forma, como exemplificado na Figura 4, é necessário que todos os esquemas sejam transformados para um mesmo modelo de dados, o que é chamado de **Modelo de Dados Comum (MDC)**. Para formas de armazenamento diferentes de bancos de dados, como arquivos XML e planilhas, deve-se simular a existência de um esquema para que o processo de integração possa ser realizado (BARBOSA, 2001).

O autor ainda define que o esquema conceitual de cada banco de dados a ser integrado é chamado de **esquema local**. No processo de integração de dados, esse esquema é transformado e descrito no MDC. O resultado disso é o que se chama de **esquema de exportação**. Todos os esquemas de exportação gerados são posteriormente integrados para a geração de um **esquema global**, único.

Ainda existe uma etapa posterior que consiste em, a partir do esquema global gerado, definir esquemas externos, utilizados para atender a solicitações específicas.

A Figura 5 exibe o funcionamento dessa hierarquia de esquemas, conforme descrito anteriormente.

Figura 5 - Etapas da integração de esquemas



Fonte: (BARBOSA, 2001)

Escolhas errôneas a respeito do nome, tipo, *constraints* de integridade, etc. podem resultar em entradas errôneas no processo de integração de esquemas. Assim, uma boa metodologia de integração de esquemas deve conduzir à detecção de tais erros (MARCÍLIO, 2002).

Nem toda a integração de esquemas é simples e rápida. Em muitos casos, é exigido que transformações sejam feitas nesses esquemas para que seja possível sua integração.

Quando, mesmo após passarem por processos assim, esquemas não puderem ser integrados por razões de inconsistência, os conflitos precisam ser informados aos usuários responsáveis, já que assim, os bancos de dados não podem ser integrados.

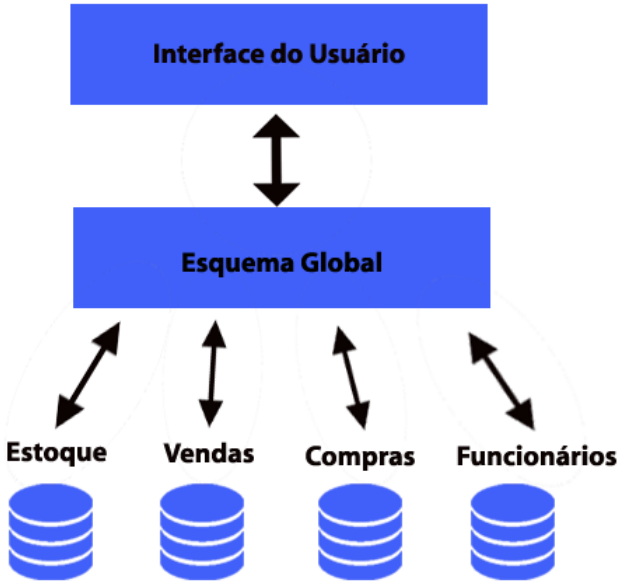
Existe ainda outra fase importante, chamada de minimização. Seu objetivo consiste em eliminar informações repetidas. Isso simplifica o esquema global gerado, já que geralmente a integração de esquemas locais encontra redundâncias que podem (e devem) ser eliminadas.

Esse esquema é o conjunto de nomes de atributos e demais características das tabelas que são posteriormente utilizadas para a criação das consultas.

A Figura 6 mostra o posicionamento do esquema global na estrutura do processo de integração de dados. É importante ressaltar que a imagem é meramente ilustrativa para apresentação deste componente, sendo que além das estruturas apresentadas na imagem, há

muitos outros artefatos que compõem essa estrutura em um modelo completo, porém aqui se toma foco no esquema global.

Figura 6 - Esquema Global no papel da integração de esquemas



Fonte: Do Autor (2016)

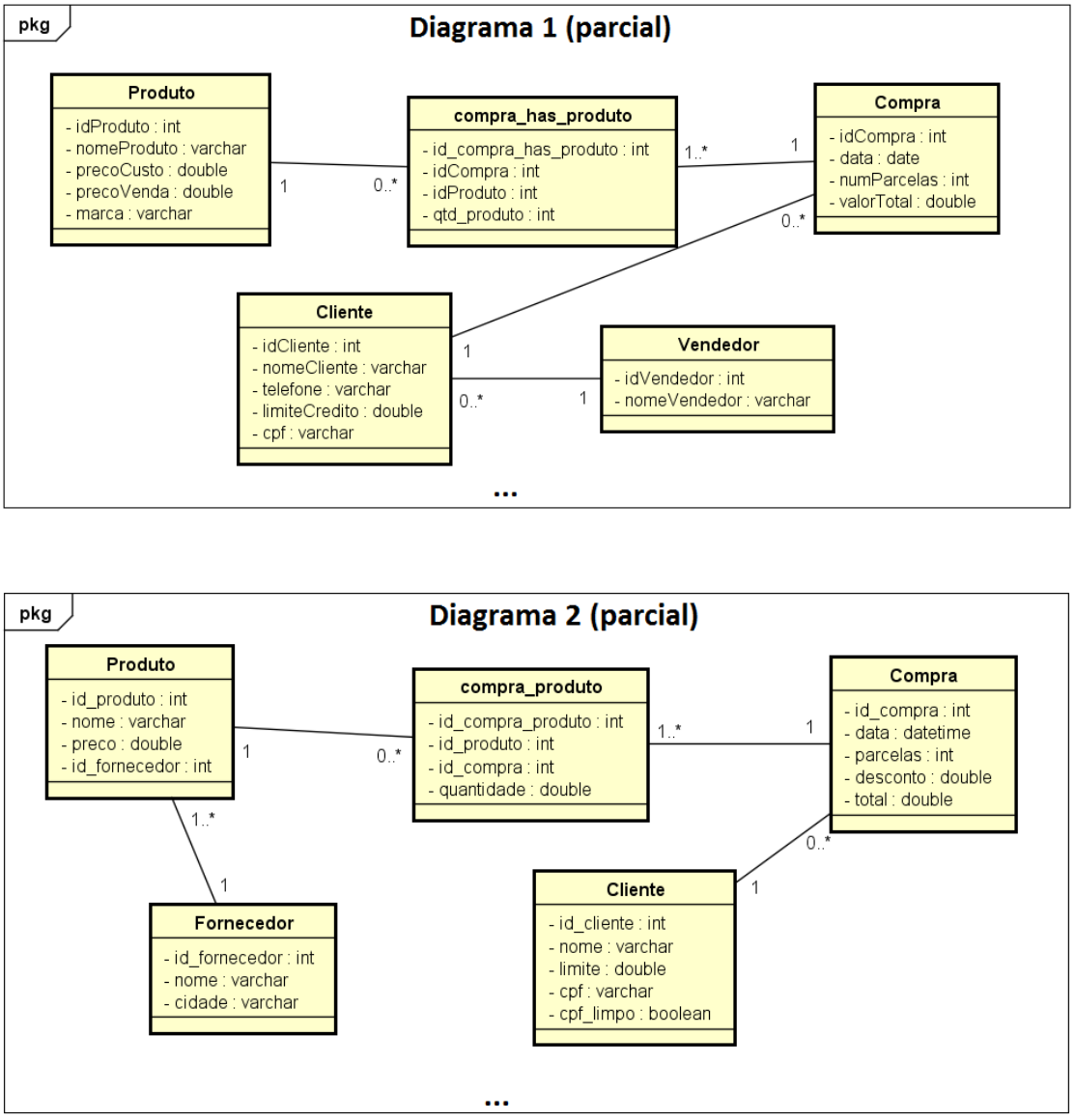
Como se observa, as diferentes bases de dados dos mais diversos setores, como no exemplo da figura, possuem cada uma seu esquema de dados local. Cada uma com suas características específicas e estrutura própria. Após passar pelo esquema de exportação obtém-se o esquema global apresentado, com uma integração de esquemas das bases da camada inferior. Assim é possível a integração de dados para que o usuário, em sua interface, possa fazer acessos (consultas) a essas fontes de forma integrada.

Segundo (RAM e RAMESH, 1999), para facilitar a obtenção do esquema global é recomendável representar todos os esquemas das bases de dados a serem integradas em um único modelo de dados, mesmo levando-se em consideração que as bases de dados iniciais foram baseadas em modelos de dados diferentes.

Para isso, quanto mais completo for o modelo de dados criado, melhor será a representação das bases de dados e a relação com os objetos no mundo real. Esse modelo precisa conter, ao menos, as entidades, atributos, chaves primárias, chaves estrangeiras, relacionamentos entre entidades e grau de relacionamento.

Na Figura 7, têm-se dois exemplos de modelos de dados para duas bases de dados diferentes, onde cada BD foi elaborada e criada de acordo com critérios específicos e isolados. O máximo de detalhes apresentados procura aproximar a representação ao máximo do mundo real. Para essa exemplificação e melhor visualização, foram utilizadas apenas algumas entidades e não o modelo de dados completo.

Figura 7 - Estruturas de Modelos de Dados de Bancos de Dados diferentes



Fonte: Do Autor (2016)

Para tal, a Tabela 1 objetiva desmontar cada diagrama e identificar claramente as entidades presentes, seus atributos, chaves primárias e chaves estrangeiras.

É possível identificar com clareza dados comuns, indicando que, provavelmente, as duas fontes de dados que operam de forma isolada possuem inúmeros registros duplicados referentes ao mesmo dado mestre.

Tabela 1: Identificação das características dos modelos de dados apresentados

Característica	Diagrama 1	Diagrama 2
Entidades	Produto, Cliente, Vendedor, Compra, compra_has_produto	Produto, Fornecedor, Cliente, Compra, compra_produto
Atributos	nomeProduto, precoCusto, precoVenda, marca, qtd_produto, data, numParcelas, valorTotal, nomeCliente, telefone, limiteCredito, cpf, nomeVendedor	nome, preco, nome, cidade, quantidade, data, parcelas, desconto, total, nome, limite, cpf, cpf_limpo
Chaves Primárias	idProduto, idCliente, idVendedor, idCompra, id_compra_has_produto	id_produto, id_fornecedor, id_cliente, id_compra, id_compra_produto
Chaves Estrangeiras	idCompra, idProduto	id_produto, id_compra

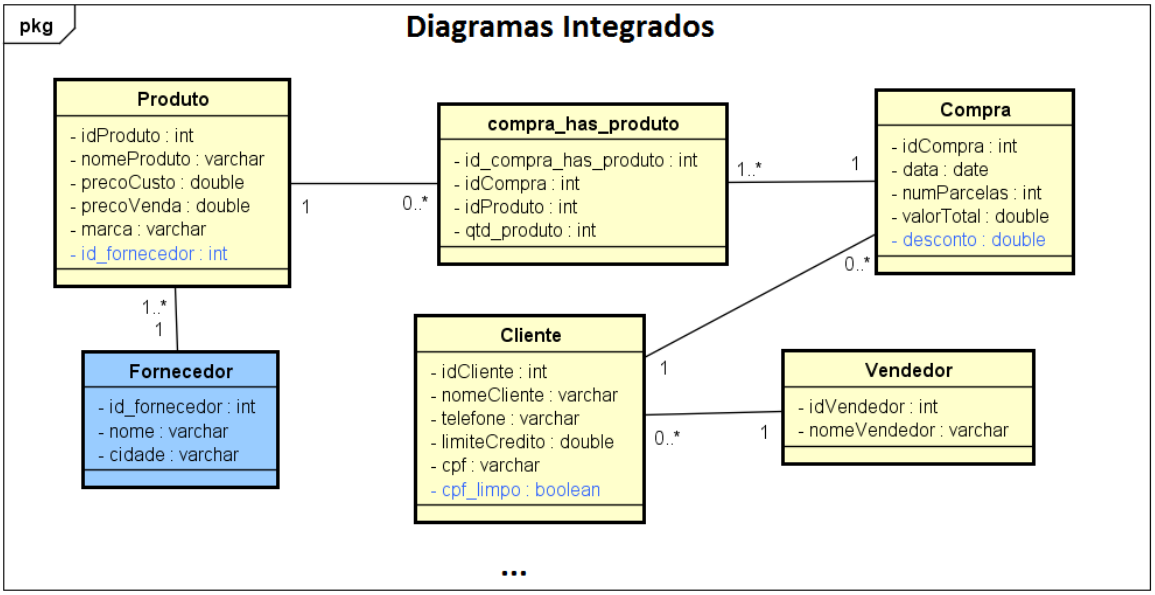
Fonte: Do Autor (2016)

Após as identificações das características de cada fonte de dados terem sido realizadas e os diagramas devidamente entendidos, avança-se para a etapa seguinte, de integração destes dados, onde se objetiva eliminar as informações repetidas encontradas, através de regras de transformação.

Nessa segunda etapa, esses modelos das bases de dados são unidos em um único modelo, integrados, justamente para facilitar a obtenção do esquema global. É o que revela a

Figura 8: o processo de integração de modelos de dados de bancos de dados diferentes para a criação de um esquema comum.

Figura 8 - Integração de Modelos de Dados



Fonte: Do Autor (2016)

Todo esse processo não é simples e rápido. Problemas de incompatibilidade entre entidades e atributos, podem facilmente aparecer. Todo o cuidado tomado na construção do modelo de dados integrado, ainda é pouco. Nos dois exemplos anteriores, usou-se apenas porções de estruturas e modelos para melhor representar os conceitos expostos, porém, geralmente tais esquemas são compostos por dezenas de entidades, relacionamentos, atributos, etc. Quando, ainda não, por centenas. Em uma ideia de integrar diversas BD, tem-se essa quantidade aumentada ainda mais, já que os modelos de cada uma das bases devem ser cautelosamente analisados.

Isso resume o nível de detalhamento dessa etapa do processo de integração de dados de fontes heterogêneas e a necessidade de ser realizada com o máximo cuidado possível, já que erros aqui implicam diretamente no desenvolvimento dos níveis superiores, como o esquema global, por exemplo.

Pode-se concluir que a integração de esquemas levou à necessidade de descer vários níveis na estrutura de um banco de dados até chegar, finalmente, no processo de planejamento e criação dos modelos que compõem a BD. Para realizar a integração de esquemas de bancos

de dados diferentes, deve-se primeiramente conhecer a estrutura de cada base individualmente para assim tornar possível uma integração correta e precisa. Isso torna ainda mais imprescindível o processo de documentação de *software*, já que na maioria das vezes, cada BD é criada em épocas diferentes por pessoas diferentes.

2.2.1. Comentários do Autor

Bancos de dados distintos podem ter elementos simbolizando a mesma coisa, porém estruturados de formas diferentes e com terminologias diferentes. É a forma própria de cada projetista criar a estrutura do banco de dados.

Para criar uma interface única de integração de fontes de dados heterogêneas, o núcleo do problema está muito distante da camada de apresentação desses dados. Como no momento da integração essas bases já estão criadas e sendo utilizadas, não existe a possibilidade de alteração na estrutura de tabelas (entidades) ou relacionamentos. Resta assim, como solução, unir informações comuns e eliminar redundâncias, exportando tais modelos para um modelo único e global.

A integração de esquemas precisa ser realizada detalhadamente e minuciosamente. Precisa se ter cuidado para interpretar corretamente cada etapa deste processo. O que está sendo integrado são os esquemas das bases distintas a fim de unir dados comuns e relacionar informações. As fontes de dados físicas continuam exatamente da mesma forma, descentralizadas. Fala-se em integrar, não em centralizar.

Os exemplos ilustrados sempre traziam uma “solução perfeita” para melhor e mais fácil entendimento. Sabe-se que em situações reais, não é exatamente assim que as informações são apresentadas.

Fica ainda uma questão a ser resolvida: Quais são os dados (entidades, tabelas, atributos) realmente importantes para a organização dentre todo o montante que está persistido? Quais são as informações cruciais dentro dos esquemas que merecem atenção especial para uma homogeneização? Conclui-se que somente integrar esquemas, sem observar atentamente o que está sendo integrado e o que é realmente necessário ser integrado, ainda não é suficiente.

2.3. Integração de Dados

Muitas empresas possuem fontes de dados distribuídas setorialmente atendendo requisitos específicos, porém com informações comuns, como já foi citado nos itens anteriores. Esse fato é causado por que não ocorreu um planejamento na criação dos bancos de dados, muitas vezes pela falta de análise ou da não antecipação dos inter-relacionamentos que poderiam ocorrer com o passar do tempo. Com isso, os sistemas de integração de dados objetivam fornecer aos usuários uma interface uniforme para as diversas fontes de dados, autônomas, distribuídas e heterogêneas (OZSU e VALDURIEZ, 1999).

Integrar dados é uma das tarefas mais complexas dentro de uma organização. Uma aplicação que integra dados tem que ser capaz de comunicar com as bases de dados e ler diversos formatos de arquivos utilizados por toda a organização.

Segundo (KALINICHENKO, 1999), o acesso integrado a bancos de dados heterogêneos é um dos grandes problemas das organizações. Daí a necessidade de estabelecer uma visão global para dados e serviços.

Com a integração de dados é possível também, por exemplo, integrar empresas, através da integração de suas bases de dados.

Esse processo de integração exige que os mesmos sejam extraídos de determinada origem, transformados segundo determinadas regras de negócio e finalmente carregados em seu novo destino.

Isso tudo feito sem haver a necessidade de reestruturar o projeto e as implementações já existentes (UCHÔA e MELO, 1999).

Muitas vezes temos quantidades enormes de informações e são necessárias consultas complexas e vários filtros para se buscar o que é desejado. Nem sempre apenas pela quantidade de dados, mas também pela quantidade de tipos diferentes de informações.

Como o objetivo de um sistema de integração de dados é oferecer ao usuário uma interface global de acesso, sua construção é pensada de forma que na aplicação o usuário defina consultas especificando o que deseja saber e o sistema fique encarregado de determinar onde tal informação pode e deve ser encontrada, para em seguida apresentar as respostas para as consultas realizadas pelo usuário.

É necessário também ressaltar que fontes de dados integradas são dinâmicas, ou seja, mesmo depois de integradas continuam a suportar suas aplicações locais, podendo atualizar os dados que disponibilizam para integração. Isso permite que os esquemas, mesmo integrados, ainda possam ser modificados e alterados.

O problema de integração começou a ser considerado crítico a partir do momento em que se disseminou a utilização de sistemas gerenciadores de bancos de dados.

Um dos principais problemas encontrados no processo de integração são as diferentes formas como os dados foram organizados e armazenados. Tamanha sua relevância, é assunto do item 2.3.1. a seguir.

2.3.1. Heterogeneidade

Vários tipos de heterogeneidade podem ser encontrados quando se deseja extrair informações contidas em mais de uma fonte. Segundo (SHETH, 1999), é possível classifica-las em: **Heterogeneidade de Sistemas, Estrutural, Sintática e Semântica**.

Assim sendo, a heterogeneidade entre sistemas se refere às diferentes características estruturais que as fontes de dados possuem por terem sido modeladas cada uma com propósitos diferentes e específicos dentro da organização.

A heterogeneidade estrutural, como o próprio nome indica, se refere à estrutura dos esquemas de diagramas de dados de cada fonte de dados. A sintática corresponde a diferenças na forma de representação de um dado, como por exemplo, representar sexo usando M e F ou usando 0 e 1. Já a semântica diz respeito à escrita, ao significado dos dados.

De acordo com (DEGAN, 2005), o legado existente na empresa não é amigável para integração, tendo as bases de dados sido implementadas em momentos ímpares. Segundo o autor, empresas que se associam geralmente possuem diferentes abordagens para representar um mesmo conceito. Essa heterogeneidade causada por essas diferenças impõe barreiras à integração.

A integração de dados pode ser realizada de duas diferentes formas: **virtual e materializada**. A diferença está na forma como os dados são tratados. As duas formas possuem vantagens e desvantagens, sendo cada uma adequada para situações diferentes.

2.3.2. Integração de Dados Virtual

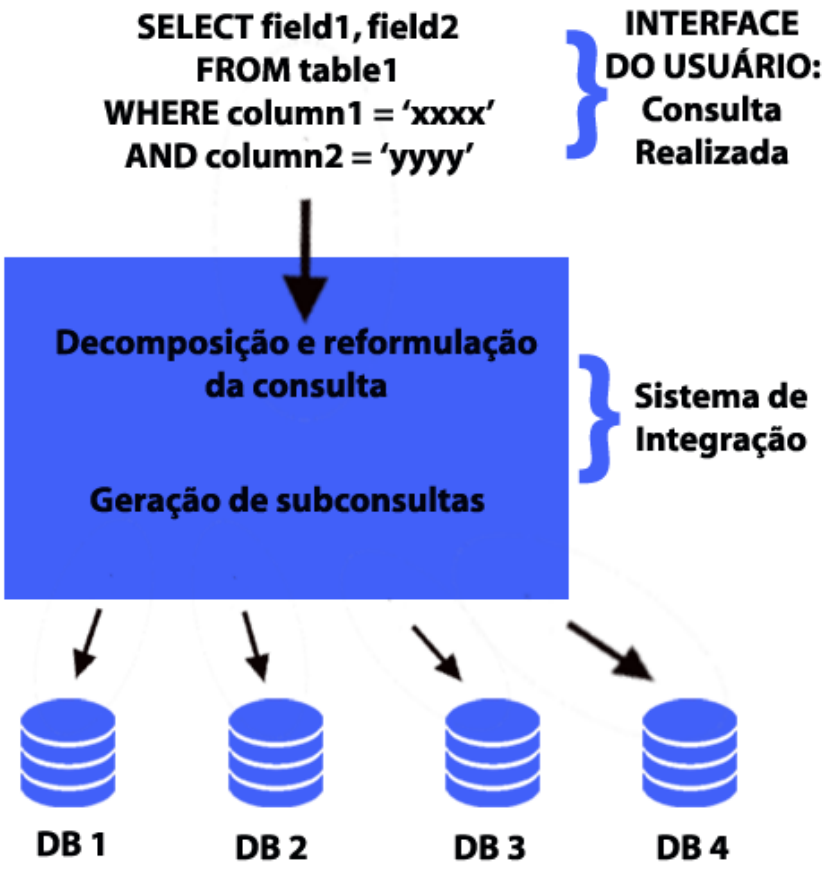
É baseada em consultas. O sistema de integração recebe uma consulta, determina quais fontes de dados são necessárias para atender a essa consulta e gera subconsultas apropriadas para cada base de dados. Ou seja, uma consulta é decomposta em várias menores, cada uma direcionada à BD específica.

Os dados coletados em cada uma dessas bases são integrados para serem então exibidos ao usuário de forma global, juntos.

Com esse processo, se tem a desvantagem de que o retorno pode ser lento, já que as consultas são "quebradas" e posteriormente os resultados encontrados são integrados. Por outro lado, pode-se dizer que a grande vantagem é a certeza de que os dados estão sempre atualizados, já que são buscados na hora em que a consulta é realizada.

A Figura 9 ilustra, de forma simples, um exemplo de integração de dados de forma virtual.

Figura 9 - Integração de Dados Virtual



Fonte: Do Autor (2016)

Na interface de consulta o usuário define a consulta que pretende realizar, com todas as condições. A mesma chega ao sistema de integração para então ser reformulada e decomposta em *sub queries* onde cada ‘pedaço’ da consulta é direcionado especificamente para a BD necessária. Não havendo necessidade de utilizar todas as fontes de dados na consulta.

No retorno ocorre o processo inverso, as respostas obtidas pelas *sub queries* são agrupadas e o resultado integrado é levado ao usuário, de forma a satisfazer assim o ciclo de acesso integrado a informações presentes em bancos de dados isolados.

O presente trabalho é focado no segundo tipo de integração, descrito no próximo tópico, que possui as características que permitiram a integração de várias fontes heterogêneas em um único meio de acesso.

2.3.3. Integração de Dados Materializada

Nessa abordagem, as informações consideradas importantes em cada uma das fontes de dados são extraídas, traduzidas e filtradas. Em seguida, essas informações são integradas entre si e então armazenadas em um repositório central.

Quando o usuário realiza uma consulta, a mesma é avaliada diretamente nesse repositório criado, sem haver a necessidade de acessar os bancos de dados originais, de onde as informações foram em um primeiro momento, extraídas.

Como desvantagem, deve se considerar que os dados podem se encontrar desatualizados. Como grande vantagem, a velocidade da consulta é extremamente rápida.

Segundo (SALGADO e LÓSCIO, 2001), a integração de dados materializada é mais adequada quando são requisitadas porções específicas e previsíveis da informação disponível, quando usuários demandam alto desempenho de consulta, sem requerer que o estado da informação seja o mais atualizado, quando é necessário acessar copias privadas de informação e quando usuários querem guardar informações que não são mantidas nas fontes de dados, como informações históricas.

Compõem esse modelo de integração duas técnicas muito famosas: DW (*Data Warehouse*) e MDM.

2.3.3.1. *Data Warehouse*

Nessa arquitetura, os dados são recuperados, integrados e armazenados em um repositório de dados, obtendo-se assim uma visão materializada dos dados. Desta forma, as consultas podem ser avaliadas diretamente nesse repositório, sem haver necessidade de acessar as fontes de dados (ABITEBOUL; BUNEMAN; SUCIU, 2000).

Usar *Data Warehouse* é menos viável quando conjuntos de dados são atualizados constantemente, com certa frequência, pois exige que o processo de extração e transformação de dados seja executado continuamente, para sincronização.

Como as fontes de dados isoladas continuam ativas e sendo usadas separadamente, tais fontes continuam sendo atualizadas mesmo depois de integradas. Assim, sofrem alterações.

(SALGADO e LÓSCIO, 2001) enumeram duas opções para a manutenção de um sistema integrado com *Data Warehouse:*

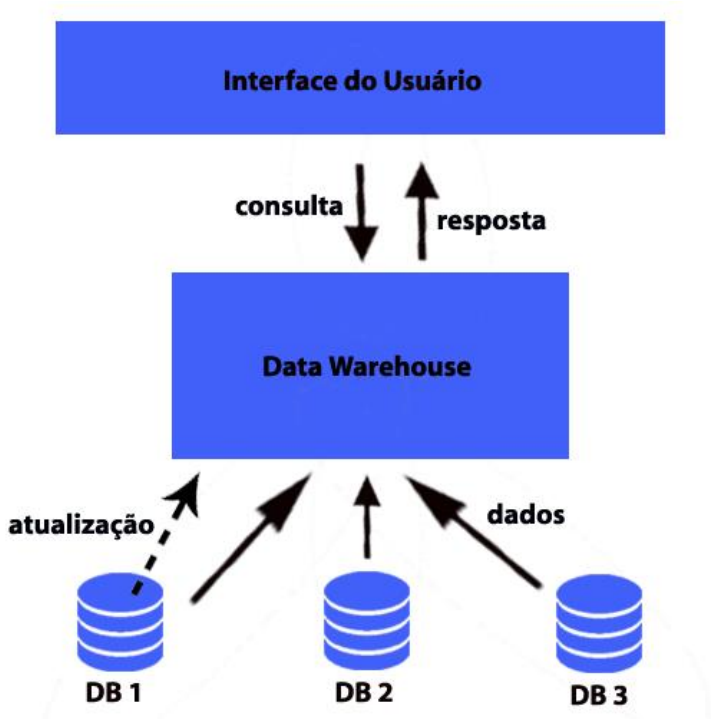
1. **Rematerialização da visão:** conteúdo é todo descartado e a visão é novamente materializada de acordo com os novos dados das fontes de dados.
2. **Manutenção incremental:** as alterações são propagadas incrementalmente, com a atualização acontecendo aos poucos.

O esquema de um *Data Warehouse* é apresentado na Figura 10, onde se observa as fontes de dados em um nível inferior e a camada de consultas do usuário no nível mais superior. O que temos na camada intermediaria é o *Data Warehouse,* um repositório central e com os dados integrados, disponíveis para consultas.

Assim, o usuário realiza uma consulta e o *Data Warehouse* retorna uma resposta de acordo com os dados que nele estão armazenados. Nesse processo, as bases de dados da camada inferior, isoladas, não são utilizadas.

Essas fontes de dados somente voltam a entrar em ação nos processos de atualização do *Data Warehouse*, conforme as setas da Figura 10, onde os novos dados são enviados ao repositório.

Figura 10 - Arquitetura de *Data Warehouse*



Fonte: Do Autor (2016)

2.3.3.2. *Master Data Management*

A arquitetura de *Master Data Management* é a segunda forma de integração de dados materializada, focada no gerenciamento dos dados mais importantes da organização.

De acordo com (HAKIMPOUR e GEPPERT, 2001), a integração de dados refere-se a combinar dados de fontes heterogêneas de forma que seja apresentada ao usuário uma única visão, totalmente uniforme. Isso leva à necessidade de uma ferramenta que integre todos esses sistemas distintos e garanta consistência e total integração das informações dentro da organização. Assim, o autor cita que ferramentas de MDM garantem a qualidade destes dados e proporcionam um significativo aumento de aproveitamento de informações úteis, muitas vezes, anteriormente ignoradas.

(LOSHIN, 2009) define que *Master Data Management* é mais do que apenas uma aplicação de integração de dados, é uma composição de pessoas, métodos, ferramentas e políticas que irão moldar o futuro de como as organizações buscam explorar o valor da informação que possuem.

Mas, para a utilização de técnicas MDM, deve-se primeiramente entender quais tipos de dados uma instituição possui. Os dados de uma empresa podem ser classificados como operacionais e não operacionais (CERVO e ALLEN, 2011):

- **Dados operacionais** são utilizados no suporte das atividades cotidianas da organização;
- **Dados não operacionais** são utilizados para suporte à decisão, normalmente capturados em DW.

Além desses dois principais tipos, as organizações lidam com pelo menos mais quatro tipos de dados considerados importantes (DREIBELBIS *et al.*, 2008):

- **Dados de referência:** são usados na categorização de outros dados, para assegurar a consistência de valores de dados mestres.
- **Dados históricos:** são usados para guardar alterações feitas nos dados mestres ao longo do tempo.
- **Metadados**: são informações descritivas úteis para ajudar a entender outros dados.
- ***Master Data*:** são então, os dados principais da organização, cruciais para sua sobrevivência.

Todos os dados considerados importantes dentro de um sistema podem ser chamados de *Master Data*, ou então dados mestres. Segundo (DREIBELBIS *et al.*, 2008), dados mestres são os dados que apresentam a informação mais valiosa que uma organização possui. Representam a informação fundamental acerca de um determinado negócio. Assim, são literalmente os dados que definem a empresa. Como exemplos de dados mestres, pode-se citar Clientes e Produtos.

A técnica de MDM é muito recente e ainda pouco explorada. A gestão de *Master Data* pode superar os problemas hoje existentes no âmbito da integração de dados, já que visam garantir também a qualidade dos dados.

(LOSHIN, 2009) ainda cita que uma aplicação MDM deve suportar as necessidades de negócio de uma organização, já que ferramentas MDM manipulam a informação mais importante da empresa, intrinsecamente ligada a decisões importantes.

Os quatro estilos de implementação de MDM mais comuns são: Consolidação, *Registry*, Coexistência e *Hub* Transacional (DREIBELBIS *et al.*, 2008).

A consolidação agrega os dados mestres em um mesmo repositório. O estilo *Registry* mantém um sistema de registro com *links* para dados existentes nos sistemas. A coexistência gere uma visão única de dados mestres, sincronizando alterações com outros sistemas. Com

Hub Transacional temos uma vista única de dados mestres assim como na técnica anterior, porém agora fornecendo acesso através de serviços. (SILVA, 2012).

Sendo assim, os dados mestres são os que definem uma instituição. Geralmente não sofrem alterações com o tempo, mantendo uma mesma estrutura, já que são o núcleo de um negócio. Podemos dividir os tipos de dados mestres em categorias: Pessoas (Cliente, Usuário, ...), Coisas (Produto, Serviço, ...), entre outras, considerando o ambiente de cada empresa.

A Figura 11 identifica um *Master Data* dentro de um ciclo de métodos e políticas de uma organização exemplo. Dentre diversas variáveis que compõem o processo de realização de uma compra ou a consulta do preço de determinado produto, tais ações somente seriam possíveis com a entidade Produto, que está diretamente relacionada às demais consultas e métodos. Fica assim evidente um dado mestre desta organização: o **Produto**.

Figura 11 - Identificação de dado mestre dentro de uma organização

Fonte: Do Autor (2016)

De acordo com (AUGUSTO, 2015), para a implementação de um sistema de MDM é fundamental a governança de dados. A GD estabelece as regras para a utilização dos dados numa organização. Exemplos de regras de governança de dados são: campos obrigatórios, valores padrão, requisitos de segurança, políticas de retenção de dados, etc.

Segundo (MOSLEY; BRACKETT; HENDERSON, 2009), a governança de dados corresponde a todas as políticas que se deve adotar na gestão do dado, conforme ilustrado na Figura 12, onde se observa todas as variáveis relacionadas à Governança de Dados.

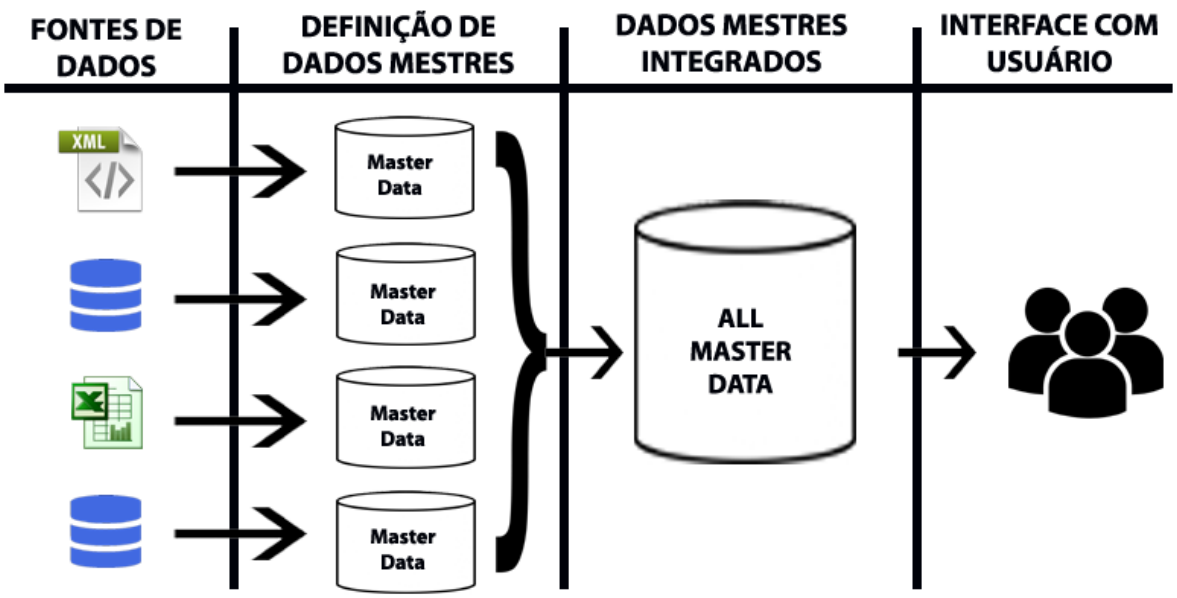
Figura 12 - Processos da Governança de Dados

Fonte: Adaptado de (MOSLEY; BRACKETT; HENDERSON, 2009)

A definição de dados mestres dentro de uma organização não acontece levando-se em consideração todas as fontes heterogêneas de dados simultaneamente. Através de processos de qualidade e governança de dados, são identificados os *Master Data* de cada fonte de dados, para, a partir disso, serem definidos os dados mestres globais, de toda a empresa/instituição.

A Figura 13 detalha esse processo, onde são definidas duas seções na camada intermediária de integração: a primeira onde são definidos os dados mestres de cada fonte de dados e a segunda onde tais dados são então finalmente integrados para um repositório único, contendo as informações mais importantes das fontes de dados iniciais em apenas um meio de acesso.

Figura 13 - Processo de extração e integração de dados mestres



Fonte: Adaptado de (AUGUSTO, 2015)

2.3.4. Comentários do Autor

O processo de integração de dados pode unir informações para tomadas de decisão a fim de obter vantagens competitivas no mercado e analisar e comparar informações de forma mais precisa.

Identificadas as diferentes formas de integração existentes, observam-se vantagens e desvantagens em ambas. A integração de dados materializada atende aos requisitos e especificações deste trabalho, através da técnica de MDM, descrita na seção anterior.

A utilização de dados mestres para integração de dados de fontes heterogêneas se mostra satisfatória. Sabe-se que desvantagens são apresentadas em todas as técnicas existentes, porém, devem ser contornadas.

As recentes pesquisas e trabalhos relacionados à MDM dão indícios de resultados satisfatórios, com as mais diversas finalidades às quais a utilização de dados mestres foi destinada.

Por fim, a conclusão à que se chega é a seguinte: integrar dados é uma tarefa extremamente complexa e minuciosa. Afinal, são as informações que definem uma instituição que estão sendo manipuladas. Por outro lado, tal processo de integração proporciona

significativos ganhos de resultados no que diz respeito às tomadas de decisão e na análise de resultados obtidos em consultas.

A identificação e definição de dados mestres é o ponto chave para o sucesso do processo de integração. Todas as demais fases e etapas isoladas que estão envolvidas nesse processo veem sua importância acentuada quando as fontes de dados estão finalmente integradas.

A identificação correta dos dados mestres é considerada imprescindível, pois senão, em caso contrário, não se teria MDM, apenas integração de dados.

2.4. Processamento das Consultas Realizadas

O objetivo mais importante da tecnologia de bancos de dados é a integração, não a centralização desses dados. É importante destacar que um desses termos não implica no outro.

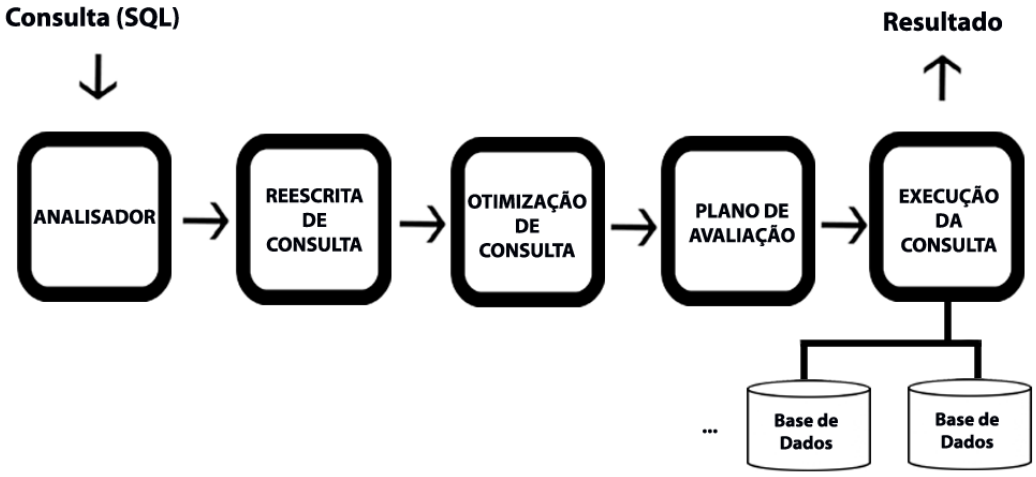
Quando, dentro de um sistema que possui fontes de dados integradas com MDM, o usuário realiza uma consulta, ocorre um processamento com essa consulta até a mesma chegar ao repositório de dados mestres. São as etapas pelas quais ela passa até a obtenção do seu resultado.

O processamento de uma consulta visa minimizar o tempo de resposta e fazer com que não ocorram falhas nesse processo. Por exemplo, se o usuário definir, na construção da consulta, o uso de algum elemento (tabela, atributo) que não esteja mapeado ou que não exista no dicionário de dados mestres criado, a consulta não realiza a tentativa de buscar algum resultando, já identificando, em uma das primeiras etapas de seu processamento, que existe um erro.

Segundo (FERRANDIN, 2002), o desempenho de uma consulta deve ser independente do local onde a mesma é executada. No seu processamento, a mesma é analisada e traduzida para o formato de uma expressão algébrica, onde é verificado se a consulta pode ser otimizada. Após a realização dessas etapas, ocorre a avaliação da consulta e a subsequente busca do resultado no repositório MDM.

A seguir, na Figura 14, a representação das fases pelas quais uma consulta passa desde a sua criação pelo usuário até a exibição do resultado, detalhando principalmente as etapas intermediárias, desconhecidas do usuário, porém fundamentais para o sucesso da mesma.

Figura 14 - Etapas do processamento de uma consulta



Fonte: Adaptado de (KOSSMANN, 2000)

Conforme a Figura, após a consulta SQL ser criada, ela primeiramente passa por um analisador, que a transforma em uma representação interna, para que possa ser facilmente utilizada pelas etapas posteriores. (BARBOSA, 2001) define que o analisador é responsável pela análise léxica e sintática da consulta global submetida, existindo a possibilidade de rejeição da consulta no caso de erros de sintaxe, de semântica ou de tipos incorretos. Quando acontece alguma rejeição, a consulta retorna ao usuário indicando o erro. Assim, uma consulta é dita incorreta se ela não apresenta todas as informações para a geração de um resultado.

(KOSSMANN, 2000) define que a segunda etapa, de reescrita da consulta, é onde acontecem típicas transformações como, por exemplo, eliminação de redundância e simplificação de expressões. Também, segundo (OZSU e VALDURIEZ, 1999) seleciona as partições de uma tabela que devem ser consideradas para responder à consulta.

Na etapa de otimização da consulta, o otimizador decide quais índices serão utilizados para o processamento da consulta e qual a quantidade de memória necessária para cada operação. Ainda de acordo com (BARBOSA, 2001), o objetivo do otimizador é encontrar uma boa estratégia ou, até mais importante que isso, evitar péssimas estratégias. A escolha de uma boa estratégia de otimização requer um cálculo de custos de execução entre as ordenações previamente definidas como candidatas.

Ao chegar ao plano de avaliação, também chamado de plano de refinamento, (KOSSMANN, 2000) define que este componente transforma o que até então foi produzido pelo otimizador em um plano executável.

Por fim, conforme indica a imagem, chega-se à etapa de execução da consulta, onde ela é finalmente executada, os dados são buscados na camada de MDM e o resultado obtido é enviado como retorno ao usuário.

2.4.1. Comentários do Autor

A forma como ocorre o processamento da consulta é extremamente importante na integração de dados, principalmente quando estamos falando de fontes de dados heterogêneas, com informações sendo buscadas em diferentes repositórios. É importante verificar as etapas intermediárias existentes nesse processo, antes do resultado ser gerado.

Mesmo o capitulo não tendo aprofundado as técnicas e algoritmos de processamento de consultas, bem como os diferentes modos de junção de subconsultas e as técnicas de otimização, sabe-se que qualquer consulta passa exatamente pelos passos apresentados anteriormente.

Isso leva à conclusão de que, para um processamento mais eficiente e assim uma resposta mais rápida, a forma como a fonte de dados é estruturada no momento da sua criação é um fator crucial. Uma fonte de dados organizada, construída através de uma boa análise de necessidades, faz com que a construção de consultas seja também, consequentemente, mais simples, acarretando em consultas mais simples e melhores estruturadas. Tal fator proporciona significativos ganhos de desempenho na aplicação de integração de dados, já que o sistema é composto basicamente pelo envio de consultas SQL e o retorno do resultado das mesmas, exigindo assim muito processamento de consultas.

3. TRABALHOS RELACIONADOS

Trabalhos relacionados à integração de dados de fontes heterogêneas e à utilização de MDM foram estudados a fim de avaliar as similaridades encontradas e as relações com o trabalho aqui desenvolvido.

Em (BARBOSA, 2001), a criação de um *middleware* para integrar fontes de dados heterogêneas utilizou a composição de *Frameworks*. O autor apresenta uma nova abordagem para a construção de sistemas de integração de dados, através da seleção, customização e integração de um conjunto adequado de componentes. Tal metodologia permitiu gerar sistemas *middleware* com configurações específicas para os requisitos de cada aplicação.

O *software* desenvolvido pelo autor se assemelha com a ferramenta desenvolvida neste trabalho em quesitos de flexibilidade e configuração. Ambos os trabalhos procuram trazer o máximo de dinamismo e customização do que está sendo integrado. Apesar de se tratar de um *framework* para o desenvolvimento de componentes, permitindo assim a integração de fontes de dados utilizando diferentes formas de comunicação ou diferentes linguagens de consulta, o trabalho de (BARBOSA, 2001) possui uma característica importante além de apenas realizar a integração, que muito se assemelha ao objetivo do MDM, sendo a seleção e integração apenas do conjunto adequado e necessário de dados ou componentes, integrando assim somente aquilo que é visto como necessário.

Em outro trabalho relacionado, (BUCHER, 2008) desenvolveu uma ferramenta para consulta sobre fontes de dados heterogêneas utilizando técnicas baseadas em mediadores. Assim, fez uso da abordagem virtual para garantir que a consulta seja sempre feita com as bases de dados atualizadas. A solução desenvolvida pelo autor realiza consultas sobre bancos de dados relacionais, arquivos XML e *Web Services*. A ferramenta é composta por uma interface para especificação dos esquemas de dados, de uma interface para escrita da consulta, um motor de processamento e a interface que exibe os resultados ao usuário.

O autor, em seu trabalho, explorou a integração de dados virtual, sendo as consultas assim, executadas diretamente nas fontes de origem, sem o mapeamento de um banco de dados que integre todas diferentes fontes definidas; diferente da integração de dados materializada descrita neste trabalho. Apesar de não fazer uso de dados mestres, a forma como ocorrem as consultas no nível mais superior da aplicação é a que mais se assemelha com o presente trabalho. Porém, o processamento dessa consulta no motor intermediário ocorre de forma totalmente diferente.

Em outro exemplo, (FERRANDIN, 2002) utilizou o padrão XML para integrar bancos de dados heterogêneos. Para tal, o autor fez uso da criação de visões materializadas dos dados presentes em cada uma das bases a serem integradas, agrupando posteriormente as diversas visões em uma única, no padrão XML.

Assemelha-se o fato de cada uma das fontes de dados definidas terem sua estrutura mapeada para serem integradas em uma única fonte de acesso. O autor faz uso de *wrappers* para exportar cada fonte de dados para o padrão XML e criar uma solução integrada no mesmo formato, para receber as consultas. Neste presente trabalho, é realizada a exportação somente dos dados mestres definidos pelo usuário e o padrão XML é aceito como entrada de fonte de dados a ser integrada, não como resultado da integração.

Já em (KAKUGAWA, 2010), o autor integrou bancos de dados heterogêneos utilizando grades computacionais, que procura possibilitar o compartilhamento e a coordenação de diversos recursos heterogêneos. No sistema desenvolvido, os dados ficam armazenados em seu local de origem, onde é realizado o compartilhamento de acesso. Assim, usuários acessam os dados integrados com a impressão de estarem armazenados localmente. Esse acesso é realizado com linguagem SQL.

Pode-se citar a similaridade da camada superior, correspondente à realização de consultas, onde o usuário, tanto na ferramenta desenvolvida em (KAKUGAWA, 2010) como na solução deste trabalho, precisa realizar somente uma consulta para pesquisar e encontrar dados que antes estavam dispersas em várias fontes de acesso.

Em (MORAES, 2009) temos um exemplo de ferramenta desenvolvida com o princípio de reescrever consultas utilizando *Web Services*. O autor desenvolveu uma ferramenta que cria um modelo global contendo a integração dos modelos locais de fontes de dados, através da utilização de um mediador. Esse mediador recebe as consultas do usuário e as reescreve conforme os esquemas locais. Como retorno, a ferramenta apresenta ao usuário o plano de execução da consulta, de acordo com os resultados da consulta. Como fontes de dados para integração, foi feito o uso de *Web Services* para consultas.

Assim, as consultas são reescritas de modo que possam ser direcionadas aos esquemas locais de cada uma das fontes de dados e encontrar o resultado desejado pelo usuário. Apesar de criar um modelo de dados global e de existir um mediador que recebe essa consulta, não existe a criação de um repositório que integre os dados, muito menos, mapeamento de dados mestres.

Em outro trabalho, (AUGUSTO, 2015) cita o desenvolvimento de uma ferramenta para gestão centralizada de dados, baseada no conceito de MDM. O autor cita a necessidade de consolidar registros de clientes e empresas, devido a problemas de redundância e má qualidade. O objetivo principal da ferramenta desenvolvida é limpar e eliminar registros redundantes de bancos de dados, implementando, assim, um sistema de garantia de qualidade dos dados mestres. Resultados do trabalho apontaram que com a implementação do MDM foi possível identificar e eliminar mais de 100.000 falsas entidades, a grande maioria resultante de processos de integração anteriores, realizados sem a utilização de MDM.

A principal semelhança com a ferramenta desenvolvida por (AUGUSTO, 2015) e o sistema de integração de fontes de dados heterogêneas através de dados mestres, é, sem dúvida, a utilização da técnica de MDM em ambos os trabalhos. O autor destaca a importância da normalização dos dados além do processo de apenas integrar.

Dentre os trabalhos analisados, notam-se várias características em comum. Entretanto, a não utilização de MDM se destaca na grande maioria, conforme ilustra a Tabela 2. Por ser uma técnica ainda pouco explorada e com pouquíssimas aplicações reais de implementação, o presente trabalho se diferencia por fazer a utilização de dados mestres para tentar assim obter resultados ainda mais satisfatórios nas consultas realizadas, ressaltando a importância de processos de normalização e eliminação de redundância sobre estes dados integrados.

Tabela 2 - Trabalhos relacionados estudados

Autor	Forma de Integração	Utilizou MDM
(BARBOSA, 2001)	Baseado em composição de *Frameworks*	Não
(BUCHER, 2008)	Utilizando técnicas baseadas em mediadores	Não
(FERRANDIN, 2002)	Através do Padrão XML	Não
(KAKUGAWA, 2010)	Utilizando Grades Computacionais	Não
(MORAES, 2009)	Reescrita de consultas usando *Web Services*	Não
(AUGUSTO, 2015)	Através do desenvolvimento de um sistema MDM	Sim

Fonte: Do Autor (2016)

4. SOLUÇÃO DESENVOLVIDA

Nas seções anteriores foram apresentados os principais conceitos relacionados aos objetivos deste trabalho, este capítulo descreve a aplicação destes conceitos ao apresentar a solução desenvolvida.

4.1. Visão Geral

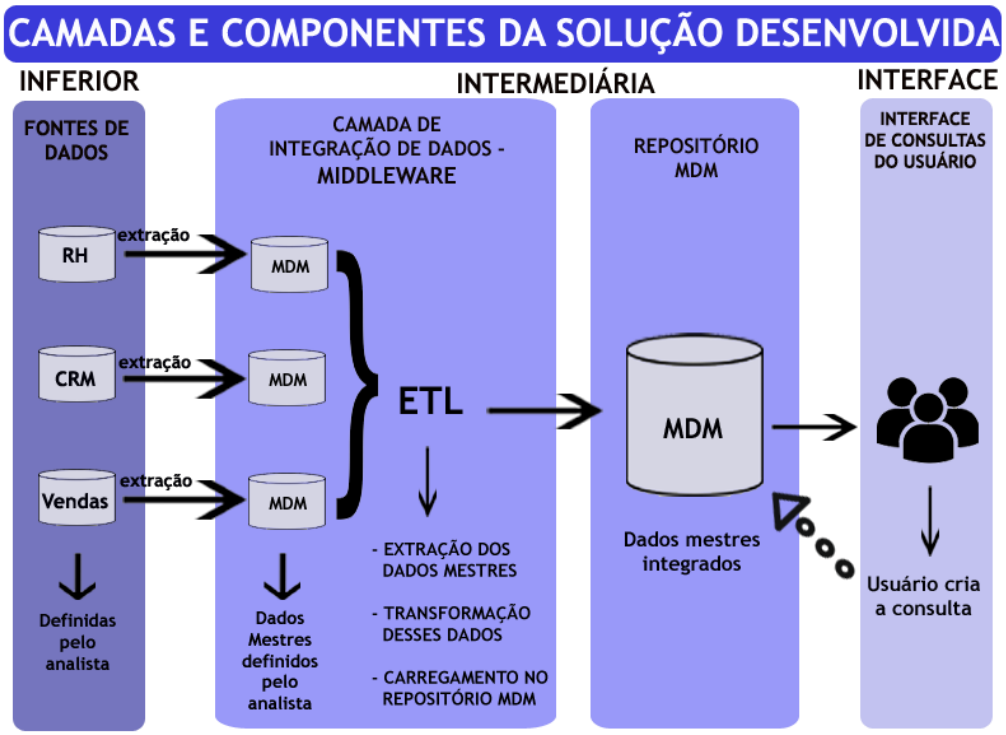
Organizar dados mestres como clientes, produtos, pessoas, fornecedores, serviços, é uma necessidade em empresas e organizações. Somente com estes dados em coerência é possível, a partir deles, extrair informação.

A solução desenvolvida neste trabalho consiste em uma ferramenta para integração de dados de fontes heterogêneas através do uso de dados mestres. Para atender a essa necessidade, o *software* permite a integração dos dados mestres, oriundos de diversas e distintas fontes de acesso, para uma visão global e única destes dados, aplicando não somente a integração sobre estes dados, como também regras de transformação específicas baseadas na qualidade dos dados apresentados.

Tal cruzamento de informações de fontes de dados isoladas pode auxiliar gestores e administradores em decisões sobre resultados obtidos, tomadas de decisões, previsões futuras, comparativos de crescimento, lucro ou ganhos, entre tantos outros aspectos.

A Figura 15 ilustra de forma global o funcionamento da aplicação, com os componentes que integram a ferramenta em todas as suas camadas, a forma como uma consulta é executada e a informação desejada é buscada e os relacionamentos da camada intermediária de *middleware* com a interface do usuário e as fontes heterogêneas. Pode-se observar que a criação do repositório MDM central possui várias etapas; já a consulta criada pelo usuário percorre um caminho muito mais curto, onde o resultado é buscado no MDM criado sem necessidade de ir às fontes heterogêneas definidas.

Figura 15 - Camadas, componentes e fluxo da solução desenvolvida.



Fonte: Do Autor (2016)

4.2. Desenvolvimento do *Software*

O *software* desenvolvido segue o fluxo da Figura 15 acima, sendo, portanto dividido em três grandes etapas ou camadas: **camada inferior**, de definição das fontes heterogêneas de dados; **camada intermediária**, de definição de dados mestres, integração e normalização desses dados; e **camada superior**, de utilização do repositório MDM criado através da criação de consultas SQL.

Seu desenvolvimento foi feito na linguagem PHP, utilizando banco de dados MYSQL para mapeamento das fontes de dados heterogêneas e criação do repositório MDM. O cuidado com a construção de um *layout* intuitivo para o usuário levou ao uso também de HTML5, *JavaScript* e CSS3. O sistema leva o usuário a cada etapa do mapeamento do MDM de forma detalhada e explicativa, contendo instruções de uso que auxiliam didaticamente sua utilização. Porém, não se descarta a necessidade de conhecimentos em Bancos de Dados para sua correta utilização.

O arquivo executável da ferramenta instala, automaticamente, o servidor APACHE e MYSQL necessários para a execução do *software*. Rodando através de qualquer navegador

Web, a ferramenta possui também *design* responsivo e não necessita de conexão com a Internet (salvo em casos onde se deseja integrar uma fonte de dados hospedada em servidor *online*).

Na primeira etapa de desenvolvimento teve-se a preocupação de permitir a integração de vários tipos diferentes de fontes de dados, fazendo com o que o usuário tenha a possibilidade de extrair seus dados mestres de fontes como arquivos CSV (gerados através de planilhas Excel), arquivos XML, *backups* gerados de bancos de dados, conexão com banco de dados *online* hospedado em outro servidor ou até mesmo a criação de uma fonte de dados através de comandos DDL (*Data Definition Language*).

A segunda etapa, vista como a mais importante por se tratar do motor de processamento interno, responsável pela realização dos mapeamentos e processos de normalização de dados, visa trazer o máximo de dinamicidade quanto à escolha do que será integrado e como será integrado.

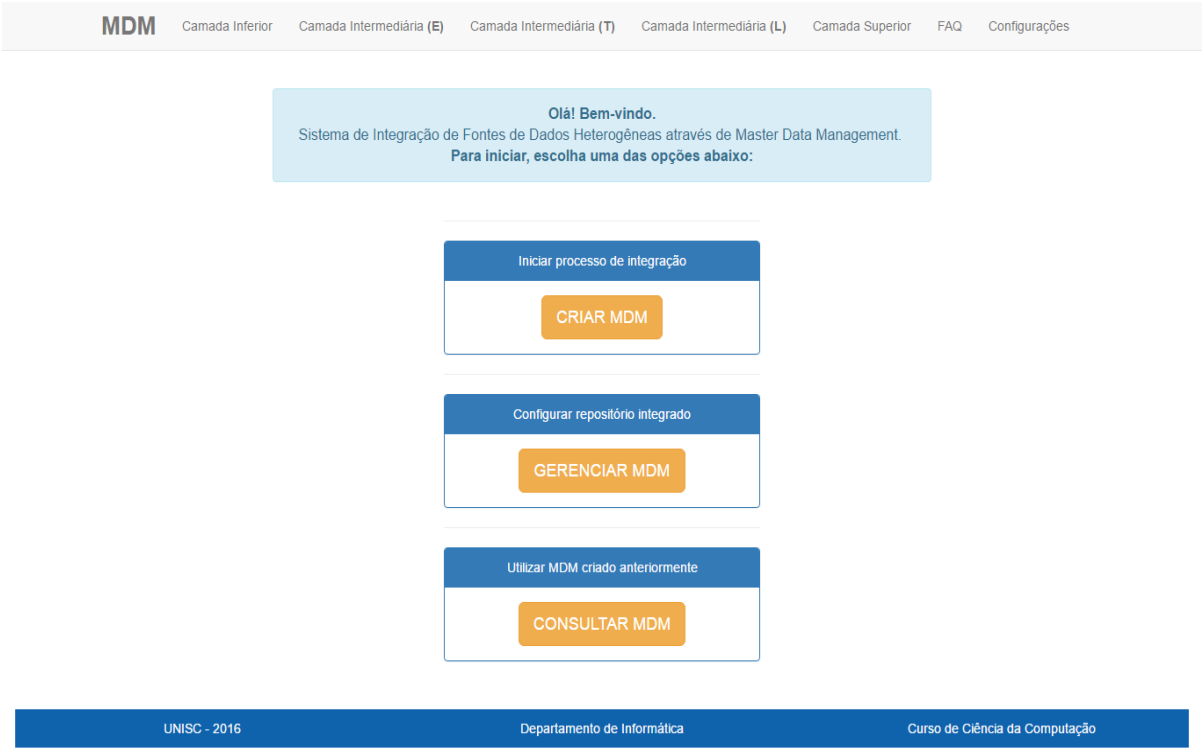
Já na terceira etapa ocorre a utilização do *software* para realização de consultas no MDM criado e a visualização da estrutura dessa fonte de dados integrados. Permite também, que consultas recorrentes e importantes sejam salvas e que o resultado da execução dessas consultas seja gerado em tabelas que permitem ordenação e filtro, para melhorar ainda mais a visualização do resultado final da *query* executada.

Para permitir tudo isso, o *software* contém um banco de dados interno, de controle e configuração, onde as definições do usuário, em todas as etapas de mapeamento, são salvas. Esse Banco de Dados é o que permite, também, que o repositório MDM criado fique salvo e assim possa ser utilizado em momentos futuros e que também tenha sua estrutura alterada a qualquer momento.

A Figura 16 mostra a interface inicial exibida ao executar a ferramenta, dando ao usuário, três opções:

- **Criação e mapeamento do MDM** pela primeira vez, seguindo as etapas sugeridas pelo *software*;
- **Configuração do MDM** criado anteriormente, a fim de alterar sua estrutura;
- **Utilização do MDM** criado, para consultas através da escrita de *queries*.

Figura 16 - Interface inicial da aplicação



Fonte: Do Autor (2016)

A Tabela 3 traz a descrição dessa interface, definida como tela inicial do sistema de integração de dados de fontes heterogêneas através de MDM.

Tabela 3 - Descrição da interface da tela inicial do sistema

Tela	Tela inicial do *software*
Objetivo	Escolher, entre as três opções, o que será feito.
Campos	-
Botões	
CRIAR MDM	Para a criação de um novo repositório MDM.
GERENCIAR MDM	Somente estará disponível caso exista um MDM para ser manipulado, permitindo a realização de alterações no MDM.
CONSULTAR MDM	Segue a mesma regra do botão anterior; caso exista um MDM criado, permite que sejam realizadas consultas no mesmo.
Parâmetros de Entrada	Escolha de uma das opções.
Parâmetros de Saída	Próxima tela (diferente para cada botão) responsável pela realização da ação escolhida.

Fonte: Do Autor (2016)

4.3. Principais Funcionalidades

A aplicação objetiva explorar a utilização de MDM em todas as etapas de integração, desde a definição dos dados mestres presentes em cada uma das fontes de dados até a utilização desses dados mestres nas consultas.

Na aplicação, o usuário não possui visão da estrutura do *middleware* e de como são aplicadas as normalizações bem como o repositório MDM é criado e alimentado. Isso proporciona uma interface menos complexa e de mais fácil utilização, sem o detalhamento de etapas intermediárias, que são realizadas automaticamente pelo componente responsável pela mediação (*middleware*).

O *software* desenvolvido neste trabalho possui duas importantes funcionalidades, onde ambas são de extrema importância para o cumprimento do objetivo de integração através de dados mestres: a primeira função é a **integração de fontes heterogêneas de dados em um único repositório**, criado dinamicamente a partir de definições do usuário. A segunda, porém não menos importante, é a característica da utilização de MDM, que visa, além de somente

integrar dados, também **aplicar regras de transformação para melhorar a qualidade dos dados integrados**.

Essas duas funcionalidades, combinadas, permitem ao usuário a criação de uma base de dados mestres confiável, que pode ser utilizada em processos de tomada de decisão causando um aumento de eficiência, redução de custos com operações baseadas em informações falhas e incoerentes e uma melhor governança.

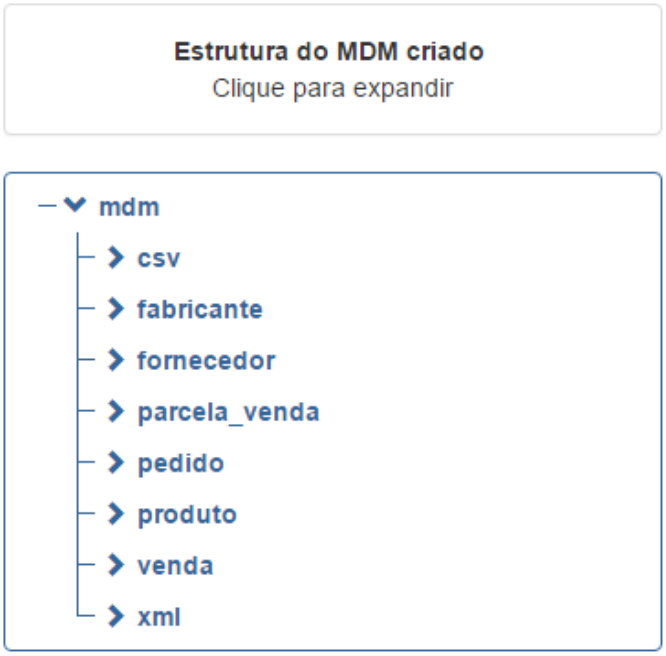
Além disso, a ferramenta possui funcionalidades específicas em cada uma das camadas da aplicação (explicadas nos tópicos **4.3.1, 4.3.2** e **4.3.3** deste trabalho), que objetivam trazer facilidades ao usuário, em cada uma dessas camadas da aplicação.

Conforme citado, já na camada inferior há a possibilidade de mapear diferentes fontes para serem integradas e, além disso, adicionar ou remover fontes de dados a qualquer momento. O *software* assume a responsabilidade de verificar cada arquivo carregado, verificando informações como tipo de arquivo, tamanho e estrutura.

Na integração das fontes de dados na camada intermediária, o desenvolvimento teve a preocupação não com somente realizar esse processo, mas também com opções de alteração da estrutura do MDM criado, permitindo a reconfiguração de tabelas, colunas e registros inseridos nessas tabelas, levando em consideração que os dados mestres tendem a permanecer os mesmos, mas a estrutura de integração destes dados mestres pode mudar com o tempo e com as necessidades da organização.

Todo esse processo acontece para que o MDM criado possa ser utilizado como fonte de consultas, gerando uma visão única global dos dados. Assim, na camada superior, o *software* traz uma árvore com a estrutura do MDM criado, conforme a Figura 17, onde são apresentadas as tabelas criadas e um clique sobre cada uma delas permite visualizar sua estrutura de colunas, funcionalidade que facilita a realização das consultas, pois dá ao usuário visão completa do repositório de dados mestres.

Figura 17 - Estrutura do MDM criado pelo usuário



Fonte: Do Autor (2016)

Além disso, existe também a possibilidade de salvar consultas que são recorrentes para o usuário, sendo possível acessá-las com maior rapidez e facilidade em uma listagem exibida na tela.

4.3.1. Camada Inferior: Carregamento das Fontes Heterogêneas de Dados

Tendo-se o conhecimento de que empresas muitas vezes possuem tabelas construídas em ferramentas como o Excel para persistirem informações e que também recebem arquivos em formato XML contendo, por exemplo, dados de compras realizadas, além de muitas vezes também separarem setores internos em bancos de dados diferentes, a aplicação de integração de dados prevê que possam ser integradas fontes de dados distintas e heterogêneas, como as citadas. A Tabela 4 define este ambiente de forma geral, com suas características principais.

Tabela 4 - Ambiente de especificação das fontes de dados a serem integradas.

Ambiente	Especificação da base de dados
Camada	Inferior
Descrição	Local para definição das fontes de dados, que podem ser arquivos em formato CSV, XML, .MYSQL (*backups* gerados a partir de bancos de dados existentes), conexão com Banco de Dados existente ou criação de BD através de comandos DDL.

Fonte: Do Autor (2016)

A Figura 18 mostra uma parte da tela do *software* onde são adicionadas as fontes de dados, com cinco diferentes opções disponíveis.

Figura 18 - Opções para upload de fontes de dados para o software desenvolvido.

Fonte: Do Autor (2016)

A ferramenta realiza uma **carga de cada uma das fontes de dados** para o servidor de Banco de Dados MYSQL local, criando assim um banco de dados interno para cada uma das fontes definidas. Isso permite que tabelas possam assim ser integradas e associadas nas etapas seguintes.

Para isso, primeiramente os arquivos carregados pelo usuário passam por uma **validação** a fim de identificar suas características. A Figura 19 traz um pequeno trecho do código que verifica alguns parâmetros do arquivo, nesta imagem, exclusivamente para arquivos CSV (porém, as validações são realizadas para todas as diferentes fontes de dados definidas), como erros no arquivo, sua extensão e tamanho.

Figura 19 - Validação de arquivo CSV.

```php
// Tipo de arq. permitido, somente arquivos oriundos de planilhas, como o CSV
$tiposPermitidos= array('application/vnd.ms-excel');
// Tamanho máximo (em bytes)
$tamanhoPermitido = 1024 * 10000; // 10 Mb
// O nome original do arquivo no computador do usuário
$arquivo_nome = $_FILES['csv']['name'];
// O tipo  do arquivo.
$arquivo_tipo = $_FILES['csv']['type'];
// O tamanho, em bytes, do arquivo
$arquivo_tamanho = $_FILES['csv']['size'];
// O nome temporário do arquivo, como foi guardado no servidor
$arquivo_temp = $_FILES['csv']['tmp_name'];
// O código de erro associado a este upload de arquivo
$arquivo_erro = $_FILES['csv']['error'];

// Verifica se foi realizado o upload de um arquivo
if(!file_exists($_FILES['csv']['tmp_name']) || !is_uploaded_file($_FILES['csv']['tmp_name']))
{
    $erro = TRUE;
}else{

    // verifica tamanho do arquivo
    if($arquivo_tamanho > $tamanhoPermitido)
    {
        $erro = TRUE;
    }

    // Verifica o tipo de arquivo enviado
    elseif(array_search($arquivo_tipo, $tiposPermitidos) === false)
    {
        $erro = TRUE;
    }else{
```

...

// continua com o upload do arquivo

Fonte: Do Autor (2016)

Esse processo garante maior segurança, uma vez que não permite tentativas de *upload* de arquivos em formatos desconhecidos pelo *software* ou o carregamento de arquivos corrompidos.

Caso a fonte de dados heterogênea definida esteja dentro das permissões para utilização, a ferramenta cria um banco de dados para esta fonte de dados. Com isso, se o processo de criação da estrutura do banco de dados (CREATE DATABASE ...) ocorre sem a

apresentação de erros, o *software* popula a base de dados recém criada com as informações contidas no arquivo da camada inferior.

Como exemplo, o trecho de código da Figura 20 a seguir é responsável unicamente por realizar **o carregamento de um DUMP** para o servidor MYSQL de Banco de Dados local, identificando automaticamente todos os comandos DDL, como CREATE, ALTER e TRUNCATE, e DML (*Data Definition Language*), como INSERT, UPDATE e DELETE presentes no arquivo para realizar a importação.

Figura 20 - Trecho do código-fonte que realiza o carregamento de arquivo DUMP que será utilizado na integração de dados.

```
// conexao ao servidor e criação do banco de dados com o mesmo nome do arquivo adicionado
$link = mysql_connect('localhost', 'root', '');
$sql = " CREATE DATABASE ".$nome_mdm; // cria banco de dados

if(mysql_query($sql, $link)) // se nao houve erro
{
    $ok = TRUE; // apenas verificação
    $db_selected = mysql_select_db($nome_mdm, $link); // seleciona banco de dados para importar dump

    // Realiza upload do dump para o mysql, pegando o conteudo do arquivo
    if($fp = file_get_contents($pasta . $arquivo_nome)) // caminho do arquivo (pasta e nome)
    {
      $var_array = explode(';',$fp); // separa em um array as consultas do arq. dump

      foreach($var_array as $value) // percorre o array
      {
        mysql_query($value.';', $link); //executa a query (DDL ou DML)
      } // fim foreach

    }else{
        $erro = TRUE;
    } // fim else
```

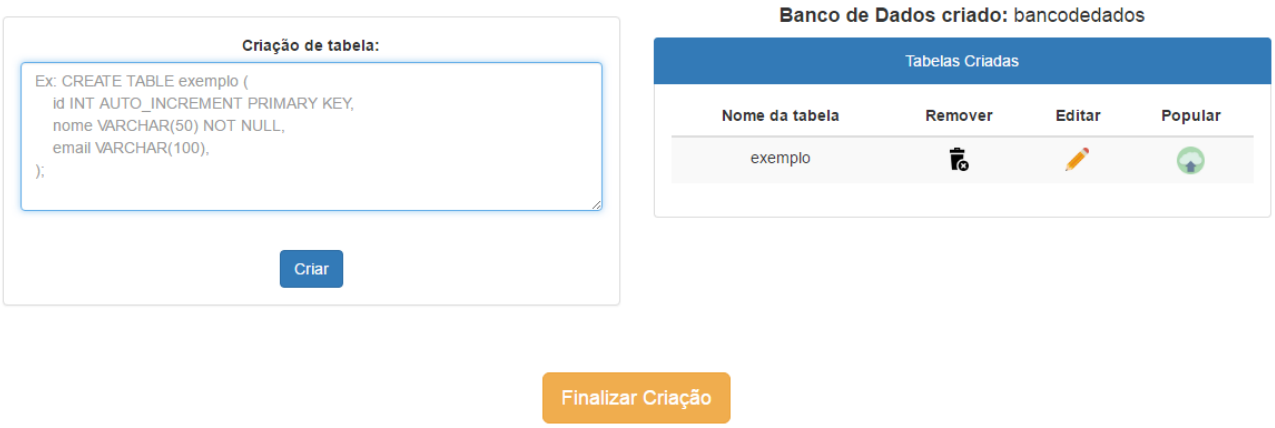
Fonte: Do Autor (2016)

O sucesso dessa operação mostra na tela, ao usuário que está realizando o processo de integração (ainda em etapa inicial), a fonte de dados pronta para uso, exibindo também informações de quantidade de tabelas criadas, data de criação e tipo de arquivo de origem. Tudo isso, sem a necessidade de nenhuma interação após a seleção da fonte de dados. Todas as verificações de erros e garantias de sucesso são realizadas pelo *software.*

A camada inferior, com o objetivo de oferecer mais possibilidades, também permite que uma base de dados seja **criada a partir de instruções SQL**. Assim, podem ser criadas e populadas tabelas que serão realmente necessárias para a integração, já que o carregamento de uma fonte de dados pronta pode conter inúmeras tabelas e colunas desnecessárias.

A Figura 21 exemplifica essa opção, mostrando a criação de um banco de dados (denominado **bancodedados**) e a criação de uma tabela (denominada **exemplo**).

Figura 21 - Criação de estrutura de banco de dados através de comandos SQL (DDL)



Fonte: Do Autor (2016)

As tabelas depois de criadas podem ainda ser **removidas**, terem sua estrutura **editada** (nome da tabela e suas colunas) e ainda receberem dados de entrada para serem, então, populadas com dados.

A camada inferior desenvolvida no *software* assume a responsabilidade de garantir o sucesso de cada uma das operações dentro do fluxo de definição de uma fonte de dados; de informar a ocorrência de erros, caso estes sejam encontrados; de permitir adição e remoção de fontes de dados; de tornar fácil e intuitiva a realização de cada uma dessas etapas citadas tornando o processo, de forma geral, simples e fácil.

Concluídas as definições das fontes heterogêneas de dados, o sistema segue ao *middleware*, que irá integrá-las na camada intermediária da aplicação, através da manipulação dos dados mestres, dentro do motor de processamento do *software*, para a então criação do repositório MDM.

4.3.2. Camada Intermediária: Integração e Normalização

Nesse ambiente, a aplicação já trabalha na camada intermediária e também na etapa mais importante para o objetivo de integração de dados: o **mapeamento de dados mestres**. Assim, a camada intermediária do software é o local onde são definidos os dados mestres das

fontes de dados reconhecidas e é feita a montagem de um modelo de dados global, baseado nesses dados mestres selecionados pelo analista responsável. Em uma segunda etapa, após o término da construção do repositório de dados mestres, são aplicadas as correções sobre os dados através de **regras de normalização e limpeza**.

A Tabela 5 especifica esse ambiente de mapeamento do MDM, considerado o mais importante da aplicação, por ser o responsável de realizar a integração de dados mestres e assim criar o repositório MDM.

Tabela 5 - Ambiente da camada intermediária do software

Ambiente	Integração de dados mestres e normalização destes dados
Camada	Intermediária
Descrição	Responsável pelo mapeamento dos dados mestres presentes nas fontes heterogêneas de dados definidas para uma nova fonte de acesso global a esses dados, além da aplicação de regras de limpeza sobre estes, visando uma melhor governança de dados.

Fonte: Do Autor (2016)

Muitas vezes o processo de **mapeamento de equivalências de esquemas** é a parte mais difícil, já que muitas fontes de dados possuem estruturas totalmente distintas umas das outras, e definir quais tabelas e atributos irão compor o modelo global de MDM é uma tarefa minuciosa, mas é também uma das mais importantes, pois impacta diretamente no resultado final do uso da aplicação.

A saída gerada na conclusão dessa etapa de mapeamento de MDM é uma estrutura nova de dados, composta por tabelas e atributos das fontes heterogêneas. Nessa etapa os dados ainda não foram transformados e padronizados, somente definidos. Tal processo ocorre, também, nessa mesma camada intermediária, porém em etapa posterior à definição destes dados mestres e a subsequente criação do modelo global.

Portanto, nesta etapa, o analista tem ao seu alcance a estrutura de todas as fontes de dados que foram especificadas para que a partir da visão de cada modelo de dados possa determinar quais estruturas representam os dados mais importantes das fontes de dados adicionadas, ou seja, **os dados mestres que serão gerenciados**.

A ferramenta tenta trabalhar da forma mais clara e intuitiva possível de modo a exigir o mínimo de trabalho e interação do usuário. Porém, deve-se levar em consideração o fato dessa etapa ser minuciosa, portanto, repleta de passos até a conclusão do mapeamento.

4.3.2.1. Integração de dados mestres

Com as fontes heterogêneas de dados definidas, o início do processo de integração das mesmas exige que seja definida a **fonte de dados *Master***, que corresponde à base de dados com maior prioridade sobre as demais. Consequentemente a isso, todas as demais fontes de dados atuam de forma secundária à primeira, sendo denominadas ***Slave***.

A escolha do Banco de Dados *Master* feita pelo usuário é armazenada no Banco de Dados interno de controle do *software*, fazendo com que a escolha do analista persista até o final de todo o processo e a preferência à BD *Master* prevaleça.

Os dados mestres, correspondentes às informações mais preciosas da empresa ou organização, estão espalhados nas várias fontes de dados. Para que o repositório MDM comece a ser modelado, esses dados precisam ser encontrados e definidos, para que os demais, insignificantes para a integração, sejam descartados.

A ferramenta já contém, desde sua instalação, um Banco de Dados (oculto, por estar sem tabelas e registros) chamado de **mdm**, que no final de todo o processo de integração possui a estrutura de dados mestres de forma completa, correspondendo assim, ao resultado dos mapeamentos realizados. Esse Banco de Dados é fixo (sem a possibilidade de alteração de nome ou remoção) para que possa assim ser sempre encontrado pelas regras internas do *software* na medida em que os dados começam a ser enviados para esta base central de acesso.

Iniciando essa definição, o sistema apresenta ao usuário que está realizando a integração, uma listagem contendo todas as tabelas, de cada uma das fontes de dados definidas na camada inferior, para que seja realizada a **seleção das tabelas correspondentes aos dados mestres**. A Figura 22 ilustra esse processo para uma fonte de dados adicionada (**fornecedores**), onde se definiu, neste exemplo, três tabelas como importantes, contendo dados mestres desejáveis de integração: **fabricante, fornecedor** e **produto.**

Figura 22 - Fonte de Dados da camada inferior com listagem de suas tabelas para identificação dos dados mestres nela presentes

Fonte: Do Autor (2016)

Seguindo esse processo para cada uma das fontes de dados mapeadas na camada anterior, **são definidos os dados mestres de cada uma das fontes existentes**. As tabelas selecionadas são extraídas de seu modelo local de dados e exportadas para um modelo global, através do processo de exportação de dados da ferramenta.

Antes de o sistema começar a popular o MDM, o *software* precisa ainda da definição, por parte do analista, dos **Identificadores Universais (IU)** para cada uma das tabelas definidas. Os IU são usados pelo algoritmo interno para mapear os dados no MDM evitando duplicidade de registros. Assim, a cada associação realizada, os identificadores universais são comparados a fim de evitar, já na primeira fase de migração dos dados, que registros duplicados sejam mapeados no MDM. A preocupação da ferramenta com problemas de heterogeneidade acompanha cada uma das etapas, e os IU são peças chave fundamentais no auxílio à identificação das duplicidades. Exemplos de IU podem ser: CPF, CNPJ, ISBN, Email, entre outros.

As tabelas escolhidas para serem integradas, a partir da premissa de serem dados mestres, quando correspondentes à fonte de dados definida como *Master*, são instantaneamente carregadas no banco de dados MDM. O *software* cumpre internamente, com essa etapa, através de duas instruções, exibidas na Figura 23.

Figura 23 - Carregamento de tabela do Banco de Dados Master para o Banco de Dados MDM

```
// cria a tabela
$sql = " CREATE TABLE IF NOT EXISTS mdm.".$master['nome_tabela']." LIKE mdm_".$master['nome_base']."."
    .$master['nome_tabela']." ";
$query = mysql_query($sql, $link);

// popula a tabela
$sql = " INSERT INTO mdm.".$master['nome_tabela']." SELECT * FROM mdm_".$master['nome_base']."."
    .$master['nome_tabela']." ";
$query = mysql_query($sql, $link);
```

Fonte: Do Autor (2016)

Eliminando todo o código fonte restante que estaria antes e depois das consultas exibidas na Figura, pode ser verificado que, pelo fato do Banco de Dados *Master* ter prioridade sobre os demais, o mapeamento das tabelas dessa fonte é bastante simples, através da execução de duas *queries*: a primeira, a criação de uma tabela, no MDM, contendo a mesma estrutura da tabela que está sendo integrada. Mesma estrutura significa mesmo nome de tabela, mesmas colunas, chaves primárias existentes, tipos de dados.

Quebrando a *query* acima e identificando as diferentes partes que a compõem, pode-se identificar:

... **CREATE TABLE IF NOT EXISTS bancodedadosmdm.tabela** ...

Onde o comando *create table if not exist* cria a tabela, *bancodedadosmdm* corresponde ao banco de dados onde está sendo criada a tabela, neste caso, o **mdm**, e *tabela* é o nome da tabela que será criada, pelo padrão do *software*, mesmo nome da tabela original na fonte de dados mapeada na camada inferior.

A segunda parte desta primeira *query* é a responsável por criar exatamente a mesma estrutura da tabela que está sendo integrada:

... **LIKE bancodedadosinferior.tabela** ...

O que garante que a tabela criada no MDM terá a mesma estrutura que a original. Terá, também, os mesmos dados, carregados através da segunda *query* da Figura, que também pode ser dividida em duas partes principais:

... **INSERT INTO bancodedadosmdm.tabela** ...

A primeira parte define em qual Banco de Dados e tabela será realizada a inserção dos registros (no caso, Banco de Dados MDM e a tabela criada na *query* anterior).

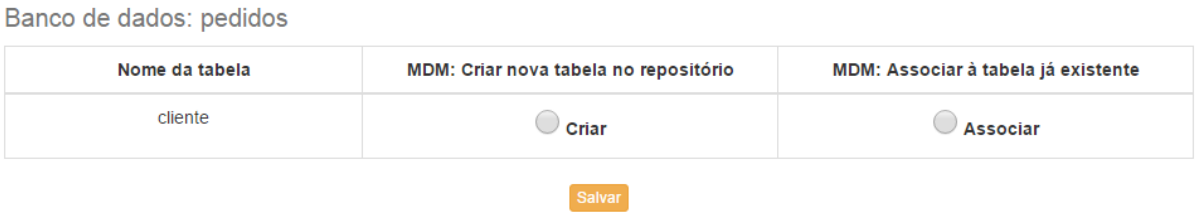
A segunda parte desta *query* define quais dados serão copiados para a nova tabela no MDM e qual o Banco de Dados de origem e a respectiva tabela:

... **SELECT * FROM bancodedadosinferior.tabela** ...

O resultado é a migração de uma tabela, com estrutura de colunas e registros, de uma fonte de dados para outra.

Para as tabelas restantes, que não foram mapeadas no MDM automaticamente, por serem originais de fontes de dados *Slave*, a ferramenta oferece duas opções, conforme ilustrado na Figura 24, para uma tabela de nome **cliente,** de um Banco de Dados de nome **pedidos.**

Figura 24 - Mapeamento, no MDM, das tabelas provenientes de Fontes de Dados *Slave*



Fonte: Do Autor (2016)

A escolha pela opção de **Criar** leva o *software* a executar as mesmas duas consultas explicadas anteriormente, pois se trata da operação de criação de novas tabelas no MDM, exatamente com a mesma estrutura e registros, vindas das fontes de dados heterogêneas.

A opção de **Associar** é a mais complexa, pois trata de associar uma tabela de uma fonte de dados definida na camada inferior, à outra tabela já existente no MDM, considerando-se que:

- A associação deve ocorrer coluna a coluna, para que os dados possam ser corretamente relacionados, já que as duas tabelas em processo de associação podem ter estrutura totalmente diferente e número de colunas também distinto;
- Algumas colunas da tabela que vem da camada inferior podem não ser de interesse do analista e assim desejáveis de serem descartadas;
- Outras colunas desta tabela, por não terem nenhuma correspondente de mesma similaridade na tabela do MDM, serão adicionadas como novas colunas;
- Colunas em comum, nas duas tabelas, possuem somente seus dados integrados, sendo realizada a verificação do IU para evitar a duplicidade de registros que já existirem no MDM.

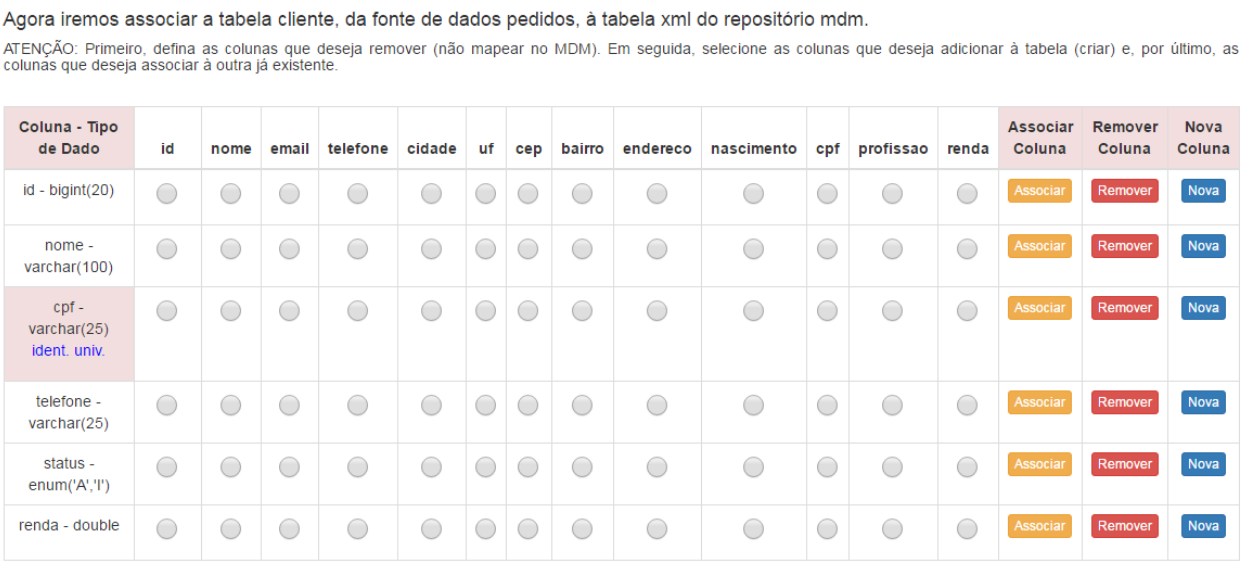
O item **4.4** deste capítulo, que trata da validação da ferramenta, detalha, passo a passo, como esse processo de associatividade de tabelas funciona, através da simulação e apresentação dos resultados obtidos com o processo.

De forma geral, a opção de associação de uma tabela proveniente da fonte de dados definida na camada inferior com outra tabela já existente no MDM, ocorre através da definição, por parte do usuário, do que irá acontecer com **cada uma das colunas** da tabela da fonte de dados da camada inferior, tendo como opção: **Remover** a coluna (não mapear essa coluna no MDM), **Nova** coluna (criar, na tabela selecionada do MDM, uma nova coluna com a mesma estrutura e dados daquela que está sendo mapeada) ou **Associar** (adicionar os dados da coluna à mesma coluna da tabela no MDM, integrando assim somente os dados, e não a estrutura).

Esse processo pode ser melhor identificado na Figura 25, que mostra um exemplo de associação, utilizando uma tabela de nome **cliente**, de uma base de dados de nome **pedidos,** mapeada na camada inferior. Essa tabela está sendo integrada ao MDM, na tabela **xml** existente.

Identifica-se, à esquerda, o nome e o tipo de dado de cada uma das colunas da tabela que está sendo integrada. No topo estão as opções de colunas da tabela do MDM, onde pode ser selecionada a coluna em que ocorrerá a associação (quando escolhida a opção de *associar*). À direita, conforme se observa, três botões referentes às três opções citadas anteriormente, disponíveis, conforme mencionado, para cada uma das colunas da tabela, permitindo assim grande dinamicidade para a realização do mapeamento, no MDM, somente do que será realmente necessário, definido pelo analista.

Figura 25 - Associação de tabela de fonte de dados da camada inferior com tabela do mdm



Fonte: Do Autor (2016)

Assim, a ferramenta orienta o usuário, dando auxílio através de telas intuitivas e simples, a realizar todo o processo de integração sem esquecer-se de nenhuma etapa.

Com todas as tabelas de dados mestres mapeadas, coluna a coluna, o repositório central de dados MDM já está criado e alimentado. Neste ponto em diante, as fontes de dados heterogêneas deixam de ser utilizadas e a ferramenta trabalha apenas com o Banco de Dados **mdm.**

Cumpre-se com isso o **objetivo de integração de dados**. Através disso, o *software* conduz o usuário à aplicação de regras e comandos para normalização da estrutura de tabela e colunas e verificação de registros do MDM criado, com o real objetivo de melhorar a governança de dados, garantindo assim maior eficiência, coerência e garantia de qualidade.

A Figura 26 ilustra que a integração está concluída, mas que, para o MDM ser utilizado para a escrita de consultas e a consequente busca de resultados, é necessário, antes disso, passar pela etapa de transformação, devido ao fato de o *software* não ser uma ferramenta apenas de integração de dados, mas sim, integração de **dados mestres** provenientes de fontes heterogêneas. E de nada adianta extrair dados mestres de diversas fontes de acesso se no novo repositório, integrados, estes dados continuarem bagunçados, não relacionados e duplicados.

Figura 26 - Conclusão da etapa de integração dos dados, restando a normalização dos mesmos para que seja possível realizar consultas

Fonte: Do Autor (2016)

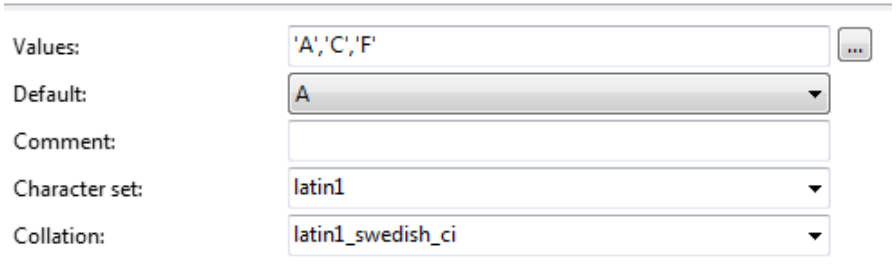
4.3.2.2. Normalização dos dados mestres integrados

A próxima etapa refere-se a duas características importantes do uso de MDM para integração de dados: **a limpeza e normalização dos dados**.

Com isso, o *software*, através do motor de processamento da camada intermediária, tem também o objetivo de, além de armazenar uma estrutura de dados mestres, padronizar tais valores, como por exemplo, converter preços para uma mesma moeda, medidas para um mesmo sistema métrico, mantendo assim um mesmo padrão de dados em cada nova tabela do MDM.

Também consiste em substituir valores faltantes, preenchendo-os com seus valores padrão (quando existirem) definidos na especificação da fonte de dados. A Figura 27 mostra que, para o exemplo da tabela **pedido**, a coluna **status** indica se o pedido realizado por determinado cliente encontra-se entre uma de três possíveis opções (*Values*): **A** (Aberto), **C** (Cancelado) ou **F** (Finalizado).

Figura 27 - Exemplo de opção padrão (*Default*) para coluna de tabela



Fonte: Do Autor (2016)

Assim, pode-se observar que, caso ocorra alguma operação onde o **status** do pedido não seja definido, a coluna será automaticamente preenchida com o valor padrão escolhido, neste caso, **A** (Aberto). A opção de transformação dos dados no MDM dá ao analista a opção de realizar essa operação, de preencher as colunas que tenham um valor *Default* pelo seu respectivo valor *Default*, quando estiverem em brancas ou como nulas (*NULL*).

A ferramenta possui também outra opção importante de normalização, que se trata de definir uma máscara padrão para determinado tipo de dado, deixando assim os dados em um mesmo formato. Por exemplo, a Figura 28 mostra alguns registros de uma tabela no MDM logo após terem sido integrados das diversas fontes heterogêneas.

Figura 28 - Coluna telefone de tabela com registros sem padrão de formatação

nome	email	telefone	cidade	uf
Joao Gomes	gomes.joao@outlook.com	(51) 9743 8229	Santa Cruz do Sul	RS
Maria do Rosario	rosariomaria.1980@gmail.com	51 99653823	Rio Pardo	RS
Fernando Fernandes	ff1989.fernando@hotmail.com	55 3734 8800	Estrela	RS
Murilo Nascimento	murilonascimento@yahoo.com.br	51 9889-7000	Santa Cruz do SUl	RS
Alexssandra Menezes	menezes.alexssanra70@gmail.com	(51) 85401234	Lajeado	RS
Guilherme Ramos	guiramos@gmail.com	5180295566	Santa Cruz do Sul	RS
Aline Regina Nascar	nascar.ar@outlook.com	5136041200	Santa Cruz do Sul	RS
Ana Clara Fagundes	anaclarafagundes@bol.com.br	51 3719 9000	Santa Cruz do Sul	RS
Carlos Eduardo	leonidas.escobar@hotmail.com	519876-6789	Santa Cruz do Sul	RS
Liane Kartz	kartz.1975liane@gmail.com	5137706612	Cachoeira do Sul	RS

Fonte: Do Autor (2016)

Para fins de **normalização**, a ferramenta permite que seja definido um formato padrão de dado para colunas que apresentam este problema. Isso acontece, primeiramente, pela definição da coluna com problema (Como a coluna *telefone* do exemplo da Figura acima), para então a definição, pelo usuário, do formato padrão desejado. Para a Figura, por exemplo, pode ser definido que o *telefone* esteja sempre no seguinte formato: **XXYYYYYYYY**, ou seja, somente com números, sem utilização de caracteres especiais como separadores nem utilização de espaços para separar, por exemplo, DDD e telefone.

Com isso a ferramenta verifica, internamente, campo a campo da coluna, eliminando caracteres indesejados (que devem ser informados pelo usuário, separados por vírgula, em um campo específico para isso). Pelo exemplo da Figura, podem ser informados para remoção, na coluna *telefone*, a seguinte sequência de caracteres: (,), -. Após esse processo, os dados da coluna estarão em um mesmo formato, sem a presença dos mesmos.

A Figura 29 apresenta o resultado dessa operação, onde os caracteres indesejados, responsáveis por aplicarem diferentes formatos de número de telefone à coluna (muitas vezes devido ao sistema utilizado pela empresa permitir que cada usuário cadastre o telefone da forma que desejar), foram removidos.

Figura 29 - Coluna telefone de tabela com registros normalizados

nome	email	telefone	cidade	uf
Joao Gomes	gomes.joao@outlook.com	5197438229	Santa Cruz do Sul	RS
Maria do Rosario	rosariomaria.1980@gmail.com	5199653823	Rio Pardo	RS
Fernando Fernandes	ff1989.fernando@hotmail.com	5537348800	Estrela	RS
Murilo Nascimento	murilonascimento@yahoo.com.br	5198897000	Santa Cruz do SUl	RS
Alexssandra Menezes	menezes.alexssanra70@gmail.com	5185401234	Lajeado	RS
Guilherme Ramos	guiramos@gmail.com	5180295566	Santa Cruz do Sul	RS
Aline Regina Nascar	nascar.ar@outlook.com	5136041200	Santa Cruz do Sul	RS
Ana Clara Fagundes	anaclarafagundes@bol.com.br	5137199000	Santa Cruz do Sul	RS
Carlos Eduardo	leonidas.escobar@hotmail.com	5198766789	Santa Cruz do Sul	RS
Liane Kartz	kartz.1975liane@gmail.com	5137706612	Cachoeira do Sul	RS

Fonte: Do Autor (2016)

Algumas situações podem exigir que determinadas colunas estejam formatadas de acordo com determinado padrão. Pelo exemplo da coluna telefone utilizada, a integração com algum possível sistema de envio de SMS pode, por exemplo, exigir que todos os telefones de clientes estejam no formato *XXYYYYYYYY* para que o envio possa ser realizado.

Após o processo de limpeza e remoção dos caracteres, a ferramenta oferece ainda ao usuário opção para inserir, em determinada posição, algum caractere, para que este seja aplicado então, da mesma forma, em todas as colunas.

Como exemplo, é possível definir que um campo **cpf** formatado de diferentes formas, após passar pelo processo de limpeza e conter apenas números (*11122233344*) seja corrigido para o formato correto de CPF (111.222.333-44). Isso, através de um campo onde o analista informa o formato da máscara que será aplicada, utilizando o caractere #. Assim, é informada a seguinte sequência para formatar um número de CPF: ###.###.###-##, sendo o *software* responsável pela aplicação do padrão em todos os valores da coluna.

Além disso, a ferramenta conta com opção para alteração de nomes de tabelas e colunas, a fim de resolver também problemas de **heterogeneidade sintática**, mantendo assim um mesmo padrão de nomenclatura.

Por fim, com os dados integrados e normalizados, o *software* permite que consultas sejam então realizadas sobre estes dados, na camada superior da aplicação, podendo-se assim, a partir dos dados, extrair informação. Esse processo, ilustrado na Figura 30, é fundamental para uma boa governança de dados, com a garantia de informações coerentes e organizadas.

Figura 41 - Obtenção de informações a partir dos dados integrados e normalizados

Fonte: Do Autor (2016)

4.3.3. Camada Superior: Realização de Consultas

A camada superior da aplicação corresponde ao local onde os dados mestres mapeados podem ser buscados através da **construção de *queries***. A Tabela 6 descreve este ambiente, o nível mais superior da ferramenta de integração de fontes heterogêneas de dados a partir de MDM.

Tabela 6 - Ambiente da camada superior

Ambiente	Realização de consultas
Camada	Superior
Descrição	Nível mais superior, onde as consultas são digitadas e especificadas para que a informação desejada possa ser buscada no MDM criado, exibindo o resultado em tabelas que permitem filtro e ordenação.

Fonte: Do Autor (2016)

O *software* possui uma interface gráfica intuitiva, que pode ser vista na Figura 31, para a realização das consultas sobre o repositório central criado com MDM. Este banco de dados recebe as consultas realizadas pelo usuário nessa camada superior da aplicação e retorna o resultado baseado nos dados mestres analisados pela consulta.

Figura 31 - Interface de realização de consultas sobre o repositório MDM criado

Fonte: Do Autor (2016)

Para que a consulta possa ser realizada corretamente e com facilidade pelo usuário, o sistema proposto prevê que, além de uma interface para entrada da consulta desejada, tenha-se a opção de **visualização do modelo de dados mestres** definido pelo analista. Assim, mesmo sem possuir conhecimento sobre cada fonte de dados heterogênea mapeada, possa-se criar uma consulta baseada apenas nos dados mestres definidos e mapeados. A estrutura do Banco de Dados central criado é exibida ao lado do campo para inserção da consulta, através de uma estrutura de árvore que possui as tabelas e colunas integradas.

Através da premissa de que o repositório MDM pode ser utilizado recorrentemente para a busca de determinadas informações, a ferramenta permite que consultas possam ser salvas, evitando assim a necessidade de digitar grandes *queries* repetidas vezes, ficando acessíveis através da opção de *Consultas salvas* na tela da camada superior.

O *software* analisa as consultas executadas pelo usuário, verificando as tabelas e campos informados na *query*. O desenvolvimento dessa verificação é importante para

contornar possíveis erros no retorno da consulta e impedir avanços desnecessários no processamento, já que em caso de sintaxe incorreta, não há necessidade de prosseguir para a próxima etapa.

Por se tratar de um ambiente para realização de consultas, comandos DDL (como *CREATE*, *DROP*, *RENAME*) não são aceitos, pois impactariam na estrutura do MDM criado, característica da camada intermediária da aplicação. A Figura 32 mostra a tentativa de remoção de uma tabela do repositório central e o resultado da validação realizado pela ferramenta.

Figura 32 - Validação de consultas

Fonte: Do Autor (2016)

O *software* valida os comandos DDL não aceitos através da verificação da presença de palavras reservadas do MYSQL, não permitidas na consulta, na *query* digitada pelo usuário, conforme ilustra a Figura 33, com trecho do código-fonte responsável pela verificação.

Figura 33 - Verificação da utilização de palavras reservadas não permitidas

```php
$sql = $_POST['consulta']; // consulta digitada pelo usuario

$permite_consulta = TRUE; // inicia permissao de realização da consulta como TRUE

// Verifica se a consulta do usuário possui algum comando MYSQL não permitido
if(preg_match('(create|alter|truncate|comment|rename|grant|revoke|database|drop)', $sql) === 1) {
    $permite_consulta = FALSE; // se possuir, não permite a consulta (FALSE)

}
```

Fonte: Do Autor (2016)

Se o **motor de processamento interno** verificar que a estrutura da consulta realizada está correta e sem erros de sintaxe, o resultado é retorno através de uma tabela. Essa tabela permite que cada coluna possa ser ordenada de forma ascendente ou descendente e que receba filtros de acordo com os diferentes valores de cada uma dessas colunas. Assim, o resultado gerado pode ser visualizado da maneira que cada usuário preferir, organizando o resultado na tela. A Figura 34 mostra uma simples consulta de duas colunas (*cliente* e *cpf*) sobre determinada tabela, onde o resultado foi ordenado pelo nome do cliente de forma ascendente, apenas clicando sobre o nome da coluna (*cliente*).

Figura 34 - Resultado de consulta realizada e ordenação por coluna específica

Resultado da consulta:

cliente Filtrar: Todos	cpf Filtrar: Todos
ALBERTO SOUZA	385.383.574-09
AMANDA MORAES	318.285.847-54
GIOVANA AZAMBUJA	382.902.309-07
JONAS HENRIQUE FUELTERN	662.316.578-92
JOSE CARVALHO NETO	883.266.664-24
MATEUS FERREIRA	390.264.929-10
PATRICIA DUARTE	341.765.685-00
PAULO VIANA	581.206.340-04
RAFAEL BRITO	432 323 802-98
REGIS MONTARRI	566.143.062-00

Fonte: Do Autor (2016)

O tamanho da tabela gerada (resultado) é totalmente **dinâmico**, de acordo com os campos informados pelo usuário na construção da consulta (*SELECT campo1, campo2* ...), onde o *software* identifica o número de colunas buscadas para gerar a tabela com o tamanho correspondente.

Para *queries* que retornam um grande número de registros, a aplicação cria também, dinamicamente, paginação para facilitar a exibição dos resultados.

Com isso, a camada superior da ferramenta procura trabalhar de forma a proporcionar ao usuário **boa navegabilidade**, de acordo com boas técnicas de organização de conteúdo na tela como também validar as consultas executadas, evitando que problemas ocorram ou ainda

que o Banco de Dados central criado com os dados mestres sofra alterações através de consultas inválidas criadas no campo para digitação destas consultas.

4.4. Validação

A validação do *software* desenvolvido neste trabalho objetiva **verificar a eficiência do sistema**, a fim de certificar se este atende aos requisitos funcionais e não funcionais. Para o cumprimento deste objetivo, foi realizada a construção do repositório MDM seguindo-se a sequência de passos do *software*, passando-se pela definição e criação das fontes de dados da camada inferior, associação de tabelas e colunas ao MDM, limpeza e normalização das tabelas do banco de dados já integrado e por fim a realização de consultas para então garantir a coerência de todo o processo através da análise dos resultados da camada superior da aplicação.

Inicialmente, para tornar possível a integração de diferentes fontes heterogêneas de dados e validar todos os tipos de associações permitidas pelo *software*, criou-se fontes de dados fictícias, distintas uma da outra, independentes, cada uma possuindo dados mestres fictícios (como clientes e produtos), mas passíveis de integração, simulando a utilização em uma mesma empresa.

Para comprovar o fato de que não apenas bancos de dados podem ser integrados, mas também outras fontes de armazenamento de dados, criou-se, para esse processo de simulação e validação, as seguintes fontes heterogêneas de dados:

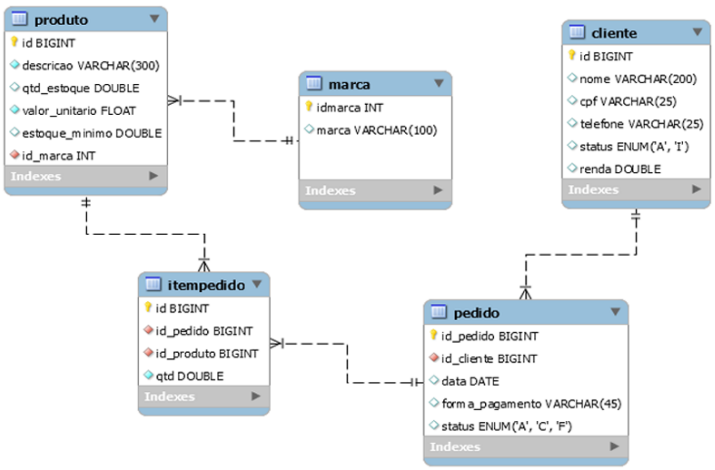
1. Arquivo em formato CSV simulando **controle de compras** realizadas por clientes em loja;
2. Arquivo em formato XML simulando **cadastro de pessoas**;
3. Banco de Dados MYSQL (posteriormente exportado como DUMP) contendo dados de **fornecedores de produtos** para loja;
4. Banco de Dados MYSQL (posteriormente exportado como DUMP) contendo **lista de pedidos** (compras ainda não finalizadas) realizados por clientes;
5. Banco de Dados MYSQL (posteriormente exportado como DUMP) com **dados de pagamentos e transferências** realizados pelo departamento financeiro.

As fontes de dados criadas são o ponto de partida para a utilização e validação da ferramenta, tendo-se assim o necessário para a criação de um repositório homogêneo de acesso a dados. Estes, mapeados através do gerenciamento de dados mestres.

Na Figura 35, que apresenta parcialmente a estrutura de algumas dessas fontes de dados, tem-se alguns registros do arquivo CSV, o modelo entidade-relacionamento (ER) do banco de dados de pedidos e um registro do arquivo XML de usuários. Notam-se as **diferenças tanto na forma de armazenamento e organização dos dados**, como também a duplicidade de registros e, inclusive, tabelas comuns que aparecem repetidas nestas fontes isoladas de dados.

Figura 35 - Fontes Heterogêneas de Dados prontas para integração

	A	B	C	D	E	F
1	cliente	cpf	cod_produto	nome_produto	preco_produto	qtd_comprada
2	ALBERTO SOUZA	385.383.574-09	3	Smartphone Dual Chip	1200	1
3	ALBERTO SOUZA	385.383.574-09	6	Pendrive 16GB	30	1
4	ALBERTO SOUZA	385.383.574-09	12	Cadeira de Praia	50	1
5	ALBERTO SOUZA	385.383.574-09	23	Garrafa Térmica 2L	60	1
6	MATEUS FERREIRA	390.264.929-10	14	TV 46 polegadas SMARTV	2000	1
7	MATEUS FERREIRA	390.264.929-10	15	cabo HDMI	50	1
8	MATEUS FERREIRA	390.264.929-10	31	Máquina de Lavar 12Kg	1100	1
9	PATRICIA DUARTE	341.765.685-00	2	Sofá 2 e 3 lugares	1500	1




```
<cadastro>
        <usuario>
                <nome>João Gomes</nome>
                <email>gomes.joao@outlook.com</email>
                <telefone>51974382291</telefone>
                <cidade>Santa Cruz do Sul</cidade>
                <uf>RS</uf>
                <cep>96810002</cep>
                <bairro>Centro</bairro>
                <endereco>Rua Marechal Floriano, 870</enderec
                <nascimento>20/04/1960</nascimento>
                <cpf>93282745398</cpf>
                <profissao>pintor</profissao>
        </usuario>
```

Fonte: Do Autor (2016)

Considerando-se que existem semelhanças **e dados mestres em comum** dentro de cada uma das fontes heterogêneas de dados, tem-se a premissa necessária para realizar a integração através da técnica de MDM.

As fontes definidas foram, primeiramente, carregadas dentro do *software* desenvolvido, cumprindo assim a primeira etapa do processo de integração. A Figura 36 ilustra o resultado dessa operação, em que as fontes heterogêneas já aparecem carregadas dentro da aplicação, preparadas para o próximo passo, correspondente à definição dos dados mestres e o início da integração dos mesmos.

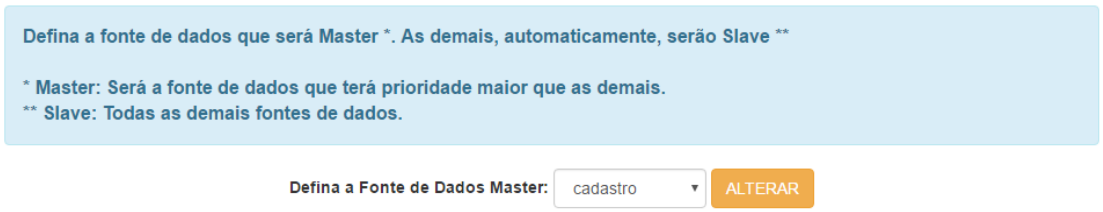
Figura 36 - Fontes de Dados carregadas no software e prontas para serem integradas

Fonte: Do Autor (2016)

Com isso, o processo de validação **segue à camada intermediária**, onde está localizado o motor de processamento e o *middleware* responsável pela integração, onde ocorrem todas as demais etapas até que o MDM esteja totalmente mapeado e construído.

Assim, seguindo a ordem sucessória das telas interativas, fez-se a definição da **fonte de dados *Master***, escolhendo-se, para esta validação, a fonte de dados **cadastro** oriunda do arquivo **xml**, ilustrado na Figura 37.

Figura 37 - Definição da Fonte de Dados *Master*



Fonte: Do Autor (2016)

Em seguida, é realizada a **definição dos dados mestres**, etapa muito importante, pois é sobre estes dados que se realiza a integração e consequentemente são estas as tabelas que fornecem os resultados das consultas na utilização do MDM criado.

Foram definidas, para cada uma das fontes de dados adicionadas, as tabelas (dados mestres) consideradas importantes e desejáveis para serem integradas. Na Figura 38, tem-se uma visão da estrutura interna que controla e armazena as definições do usuário.

Figura 38 - Definição de Dados Mestres

id	nome_base	nome_tabela	master	identificador_universal
46	financeiro	parcela_venda	N	(Null)
47	financeiro	venda	N	(Null)
48	fornecedores	fabricante	N	(Null)
49	fornecedores	fornecedor	N	(Null)
50	fornecedores	produto	N	(Null)
51	pedidos	cliente	N	(Null)
52	pedidos	pedido	N	(Null)
53	pedidos	produto	N	(Null)
54	cadastro	xml	S	(Null)
55	compras	csv	N	(Null)

Fonte: Do Autor (2016)

Nessa tabela de configuração do *software* (interna, não vista pelo usuário), pode-se observar o nome de cada tabela escolhida assim como a fonte de dados à qual cada uma pertence. Também se identifica, dentre as tabelas selecionadas, as que correspondem à fonte de dados definida como *Master* (com o identificador **S** na coluna ***master***).

Os identificadores universais, responsáveis por evitar a duplicidade de registros já no primeiro momento em que os dados são integrados, são definidos a seguir, selecionando-se uma coluna (que representará, então, o identificador universal), para cada uma das tabelas.

Na Figura 39, a tabela de controle interno apresenta, agora, a coluna de mapeamento dos IU preenchida, associando assim, um IU para cada tabela definida como dado mestre.

Figura 39 - Definição dos Identificadores Universais

id	nome_base	nome_tabela	master	identificador_universal
46	financeiro	parcela_venda	N	id_venda
47	financeiro	venda	N	id_venda
48	fornecedores	fabricante	N	descricao
49	fornecedores	fornecedor	N	email
50	fornecedores	produto	N	nome
51	pedidos	cliente	N	cpf
52	pedidos	pedido	N	id_pedido
53	pedidos	produto	N	id
54	cadastro	xml	S	cpf
55	compras	csv	N	cpf

Fonte: Do Autor (2016)

Através desses simples passos, tem-se o necessário para a integração ocorrer, por se tratar da informação necessária para o componente interno de processamento identificar as relações e definições para mapear corretamente cada um dos registros. As tabelas referentes à fonte de dados definida anteriormente como *Master*, por terem prioridade sobre todas as demais, são automaticamente carregadas no novo repositório central. Nesta validação, a tabela **xml** da fonte de dados **cadastro**, que possui o IU **cpf.**

A ferramenta permite que, caso desejado pelo usuário, colunas vistas como desnecessárias, possam ser removidas, a fim de manter **somente as informações importantes**.

Com isso se encerra no repositório integrado, o mapeamento das tabelas, colunas e registros provenientes da fonte de dados *Master*.

Para todas as demais tabelas (*Slave*), definidas nas etapas iniciais, que são também mapeadas no repositório MDM, cabe escolher uma entre duas opções:

- **Criar** uma nova tabela no repositório de MDM contendo a estrutura e os dados da tabela escolhida (de forma bastante simples, um *CREATE*);
- **Associar** a tabela escolhida a alguma das tabelas já existentes no MDM (de forma bastante simples, um *UPDATE*).

Sendo assim, os dados mestres (tabelas definidas nas fontes de dados *Slave*) que ainda não possuíam alguma referência passível de integração no MDM, como tabelas com uma estrutura similar, foram enviados ao repositório central de integração com o comando de criação de nova tabela.

Para as demais, desejáveis de associação à alguma tabela já criada no MDM, o *software* permite definir em qual dessas tabelas ocorrerá a associação e, em seguida, quais colunas serão associadas entre si, quais colunas serão adicionadas na tabela escolhida e quais não serão integradas. Na validação realizada, por exemplo, escolheu-se por associar a tabela **cliente,** da fonte de dados **pedidos** (na camada inferior), à outra tabela já existente no MDM, correspondente à tabela **xml** (mapeada antes das demais por vir da fonte de dados *Master*). Somente migrar a tabela para o MDM não seria uma boa escolha, pela quantidade de colunas e registros em comum em ambas as tabelas.

A **dinamicidade da ferramenta** permite que esse mapeamento seja realizado coluna a coluna. A Figura 40 mostra que, para cada uma das colunas da tabela **cliente,** pode-se definir a coluna correspondente na tabela **xml.** Assim, por exemplo, nomes de clientes que antes se encontravam dispersos em dois locais diferentes (alguns clientes na tabela **cliente** no banco de dados **pedidos** e outros clientes no arquivo **XML**), agora estarão em uma única coluna na mesma tabela do MDM (coluna **nome**, tabela **xml**).

Figura 40 - Associação de colunas com o mesmo tipo de dado entre duas tabelas diferentes de fonte de dados diferentes

Fonte: Do Autor (2016)

Para validação dessa integração, a Figura 41 apresenta a tabela do MDM escolhida (tabela **xml**) antes do processo de associatividade acontecer.

Figura 41 - Tabela do MDM antes de receber dados de outra tabela da camada inferior

id	nome	cpf	email	telefone	cidade	uf	cep	bairro	endereco	nascimento
1	João Gomes	93282745398	gomes.joao@outlook.com	51974382291	Santa Cruz do Sul	RS	96810002	Centro	Rua Marechal Floria	20/04/1960
2	Maria do Rosário	84839240012	rosariomaria.1980@gmail.com	5199653823	Rio Pardo	RS	96640000	Barro Vermelho	Rua 10, 454	20/04/1980
3	Fernando Fernandes	57383297611	ff1989.fernando@hotmail.com	5537348800	Estrela	RS	95880970	Centro	Rua Coronel Flores,	12/09/1989
4	Murilo Nascimento	11029343377	murilonascimento@yahoo.com.br	5198897000	Santa Cruz do SUl	RS	96810124	Centro	Rua Venâncio Aires,	10/10/1050
5	Alexssandra Menezes	99928165710	menezes.alexssanra70@gmail.com	5185401234	Lajeado	RS	95900970	Florestal	Avenida dos Quinze	20/02/1970
6	Guilherme Ramos	03429186544	guiramos@gmail.com	5180295566	Santa Cruz do Sul	RS	96810656	Goiás	Rua Professor Carlo	20/11/1992
7	Aline Regina Nascar	08099012456	nascar.ar@outlook.com	5136041200	Santa Cruz do Sul	RS	96810908	Centro	Rua Marechal Deod	15/05/1978
8	Ana Clara Fagundes	87045410045	anaclarafagundes@bol.com.br	5137199000	Santa Cruz do Sul	RS	96810042	Centro	Rua Vinte e Oito de	25/12/1965
9	Escobar Leonidas	91065634999	leonidas.escobar@hotmail.com	5198766789	Santa Cruz do Sul	RS	96810068	Centro	Rua Thomaz Flores,	20/01/1969
10	Liane Kartz	01034004429	kartz.1975liane@gmail.com	5137706612	Venâncio Aires	RS	96810182	Centro	Assis Brasil, 18	14/08/1965

Fonte: Do Autor (2016)

Em seguida, a Figura 42 traz a mesma tabela após a integração citada anteriormente, onde colunas em comum foram associadas (no exemplo, colunas **nome** e **cpf**), além da adição de novas colunas da tabela cliente (coluna **renda**) e a remoção de colunas desnecessárias da tabela cliente, que não foram "levadas" para o MDM.

Figura 42 - Tabela do MDM depois de receber dados de outra tabela da camada inferior

id	nome	cpf	email	telefone	cidade	uf	cep	bairro	endereco	nascimento	renda
1	João Gomes	93282745398	gomes.joao@outlook.com	51974382291	Santa Cruz do Sul	RS	96810002	Centro	Rua Marechal Floria	20/04/1960	(Null)
2	Maria do Rosário	84839240012	rosariomaria.1980@gmail.com	5199653823	Rio Pardo	RS	96640000	Barro Vermelho	Rua 10, 454	20/04/1980	(Null)
3	Fernando Fernandes	57383297611	ff1989.fernando@hotmail.com	5537348800	Estrela	RS	95880970	Centro	Rua Coronel Flores,	12/09/1989	(Null)
4	Murilo Nascimento	11029343377	murilonascimento@yahoo.com.br	5198897000	Santa Cruz do SUl	RS	96810124	Centro	Rua Venâncio Aires,	10/10/1050	(Null)
5	Alexssandra Menezes	99928165710	menezes.alexssanra70@gmail.com	5185401234	Lajeado	RS	95900970	Florestal	Avenida dos Quinze	20/02/1970	(Null)
6	Guilherme Ramos	03429186544	guiramos@gmail.com	5180295566	Santa Cruz do Sul	RS	96810656	Goiás	Rua Professor Carlo	20/11/1992	(Null)
7	Aline Regina Nascar	08099012456	nascar.ar@outlook.com	5136041200	Santa Cruz do Sul	RS	96810908	Centro	Rua Marechal Deod	15/05/1978	(Null)
8	Ana Clara Fagundes	87045410045	anaclarafagundes@bol.com.br	5137199000	Santa Cruz do Sul	RS	96810042	Centro	Rua Vinte e Oito de	25/12/1965	(Null)
9	Escobar Leonidas	91065634999	leonidas.escobar@hotmail.com	5198766789	Santa Cruz do Sul	RS	96810068	Centro	Rua Thomaz Flores,	20/01/1969	(Null)
10	Liane Kartz	01034004429	kartz.1975liane@gmail.com	5137706612	Venâncio Aires	RS	96810182	Centro	Assis Brasil, 18	14/08/1965	(Null)
11	Pedro Ernesto	10466969644	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
12	Maria Eduarda	59756705256	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
13	Doris Elisa Jahl	31778095954	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
14	Fernando Deres	63779743622	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
15	Elisa Lemos	87407990033	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
16	Ismael França	35088077967	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
17	Marcos Palmeiras	34736605502	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
18	Carlos Eduardo	77882940837	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
19	Daniella Monica Cardo	75095238214	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
20	Sebastião Dell Veno	90508158108	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)

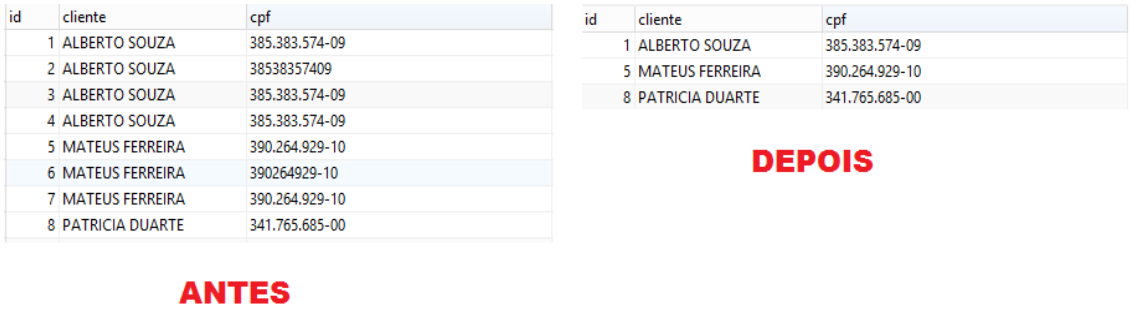
Fonte: Do Autor (2016)

É validado, com isso, principalmente, a utilização dos **identificadores universais** definidos nas primeiras etapas, onde os IU das duas tabelas são comparados para cada registro inserido, não permitindo assim a inserção de algum dado repetido. Eliminando assim, a duplicidade de registros que acontecia quando as fontes de dados se encontravam organizadas heterogeneamente.

Realizada a integração das fontes de dados heterogêneas, são aplicadas, na próxima etapa, ainda dentro da camada intermediária, **regras de transformação** sobre os dados integrados, a fim de normalizar e organizar ao máximo o MDM.

A seguir, a Figura 43 mostra um exemplo de **transformação aplicada** em uma tabela do MDM, onde a coluna **cpf** foi formatada para um mesmo padrão (XXX.XXX.XXX-XX) e registros que ainda estavam duplicados, pelo motivo da tabela ter sido copiada por inteiro da camada inferior, eliminados.

Figura 43 - Transformações sobre dados em tabela do MDM



id	cliente	cpf
1	ALBERTO SOUZA	385.383.574-09
2	ALBERTO SOUZA	38538357409
3	ALBERTO SOUZA	385.383.574-09
4	ALBERTO SOUZA	385.383.574-09
5	MATEUS FERREIRA	390.264.929-10
6	MATEUS FERREIRA	390264929-10
7	MATEUS FERREIRA	390.264.929-10
8	PATRICIA DUARTE	341.765.685-00

ANTES

id	cliente	cpf
1	ALBERTO SOUZA	385.383.574-09
5	MATEUS FERREIRA	390.264.929-10
8	PATRICIA DUARTE	341.765.685-00

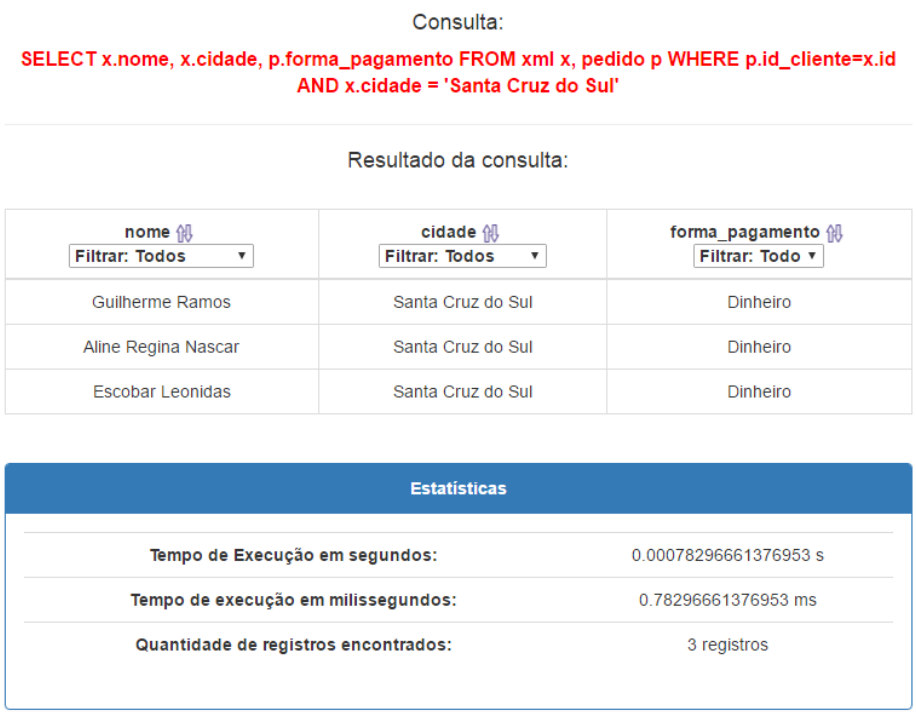
DEPOIS

Fonte: Do Autor (2016)

Por fim, após aplicadas todas as transformações sobre os dados, a fim de corrigir problemas de heterogeneidade e sintaxe, a **validação** da ferramenta de integração de fontes de dados heterogêneas através de *Master Data Management* se conclui com a realização de consultas na **camada superior**, que fornece uma visão global da estrutura do MDM mapeado.

A consulta criada com tal intuito, ilustrada na Figura 44, utiliza somente o repositório integrado de dados e sua estrutura de tabelas e colunas, exibindo como resultado **dados coerentes, sem redundância, sem erros e padronizados**.

Figura 44 - Realização de consulta no MDM mapeado



Consulta:

```
SELECT x.nome, x.cidade, p.forma_pagamento FROM xml x, pedido p WHERE p.id_cliente=x.id
AND x.cidade = 'Santa Cruz do Sul'
```

Resultado da consulta:

nome (Filtrar: Todos)	cidade (Filtrar: Todos)	forma_pagamento (Filtrar: Todo)
Guilherme Ramos	Santa Cruz do Sul	Dinheiro
Aline Regina Nascar	Santa Cruz do Sul	Dinheiro
Escobar Leonidas	Santa Cruz do Sul	Dinheiro

Estatísticas

Tempo de Execução em segundos:	0.00078296661376953 s
Tempo de execução em milissegundos:	0.78296661376953 ms
Quantidade de registros encontrados:	3 registros

Fonte: Do Autor (2016)

Conclui-se, através dos testes realizados com a utilização do *software* desenvolvido para a integração das diferentes fontes de dados, que os resultados validados, etapa a etapa, foram satisfatórios e de acordo com os objetivos da ferramenta.

Define-se o processo de validação como de suma importância para a garantia de funcionamento do sistema criado. Através dele, verificou-se o **atendimento aos requisitos funcionais** em cada uma das camadas de criação do MDM, onde a correta realização dos mapeamentos foi também validada diretamente no repositório de dados integrados.

5. TRABALHOS FUTUROS

O processo de integração de dados através de MDM é um tema abrangente principalmente pelas inúmeras possibilidades de integração e pela quantidade de recursos e tecnologias envolvidas. Trabalhar com integração de dados sempre abre um leque de possibilidades quanto às tecnologias que podem ser utilizadas para realizar a criação de um repositório homogêneo. A escolha do MDM para tal processo define algumas características e regras que acompanham o desenvolvimento.

Este trabalho propôs a criação de uma ferramenta que permite a integração de diferentes fontes de dados heterogêneas através da técnica de MDM. O processo de definição das diferentes fontes de dados na camada inferior e o mapeamento de seus dados mestres no repositório central, homogêneo, ocorre de forma dinâmica, de acordo com as necessidades pontuais de cada usuário, podendo ter sua estrutura alterada a qualquer momento.

Porém, o *software* desenvolvido não implementa o processo de atualização do MDM criado, a partir das atualizações que acontecem nas fontes de dados isoladas (que continuam funcionando e recebendo novos dados).

A integração de dados materializada, característica deste trabalho, prevê essa desvantagem quanto à sua utilização. Sabe-se, entretanto, que dados sempre atualizados garantem resultados com um nível maior de coerência e satisfação.

Assim, como sugestão para trabalhos futuros cita-se o desenvolvimento, dentro da ferramenta desenvolvida, do processo de atualização do MDM, que ocorre depois do mesmo ter sido criado, com base nas atualizações recebidas nas tabelas das fontes de dados da camada inferior, que podem continuar sendo manipuladas separadamente. Tendo em vista, principalmente, alterações de dados mestres ou alterações em valores de dados mestres já mapeados no MDM.

6. CONCLUSÃO

Este trabalho propôs apresentar uma solução para a integração de dados provenientes de fontes heterogêneas, utilizando a técnica de MDM, ainda pouco utilizada em aplicações e ferramentas de integração de dados.

A conceituação de diferentes componentes que fazem parte das várias camadas do processo de integração mostra que criar um repositório de dados confiável, homogêneo, de fácil utilização para o usuário e construído com o uso de dados mestres, consiste, primeiramente, de uma elaboração detalhada dos objetivos finais da aplicação e de um planejamento passo-a-passo para um desenvolvimento sem retrabalho.

Com isso, o presente trabalho objetivou a especificação das estruturas, componentes e conceitos que foram importantes e necessários para possibilitar o desenvolvimento da aplicação. Também, a descrição da solução desenvolvida com abrangência tanto em como se dá a sua utilização como também na forma em que foi construída, apresentando detalhes da estrutura interna do *software* de integração de fontes de dados.

Conclui-se, através do levantamento bibliográfico realizado, do desenvolvimento do *software* proposto e do processo de validação realizado, que a técnica de Gerenciamento de Dados Mestres é uma opção de integração válida e confiável, desde que a integração seja realizada seguindo todas as regras de transformação de dados e a definição dos dados mestres de cada fonte de dados seja feita por um analista capacitado e de forma avaliativa e minuciosa. Com isso, pode-se afirmar que a escolha correta dos dados mestres que compõem o repositório MDM, na camada intermediária, interfere diretamente na qualidade dos dados retornados nas consultas realizadas na ferramenta de integração. Dentre tantas formas e diferentes possíveis caminhos a seguir quando se deseja integrar dados de fontes heterogêneas, MDM aparece como, além de uma técnica que permite tal integração, também como auxiliador na governança e qualidade de dados, trazendo assim, maior confiança no resultado final.

7. REFERÊNCIAS

ABITEBOUL, S.; BUNEMAN, P; SUCIU, D. *Data on the Web*. Morgan Kaufmann Publishers. First Edition. 2000.

AUGUSTO, N. J. *Implementação de um sistema de MDM como meio de integração de aplicações*. Mestrado em Ciências da Computação, Faculdade de Ciências da Economia e da Empresa, Universidade Lusíada de Lisboa, Portugal. 2015. 83 p.

BARBOSA, A. C. P. *Middleware para Integração de Dados Heterogêneos Baseado em Composição de Frameworks*. Tese de Doutorado, Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro. 2001. 154 p.

BUCHER, S. *Uma Ferramenta para Consulta sobre Fontes de Dados Heterogêneas*. Departamento de Informática, Curso de Ciência da Computação, Universidade de Santa Cruz do Sul, Santa Cruz do Sul, RS, Brasil. 2008.

CERVO, D.; ALLEN, M. *Master Data Management in Practice Achieving True Customer MDM*. New Jersey: Wiley Corporate F & A.2011.

DEGAN, J. O. C. *Integração de dados corporativos: uma proposta de arquitetura baseada em serviços de dados*. Instituto de Computação. Universidade Estadual de Campinas. Tese de Mestrado. 2005.

DREIBELBIS, A.; MILMAN I.; RUN, P. V.; HECHLER, E.; OBERHOFER, M.; WOLFSON, D. *Entreprise Master Data Management: Na SOA Approach to Managing Core Information*. New Jersey: IBM Press. (1ª ed.). 2008.

FERRANDIN, M. *Integrando Bancos de Dados Heterogêneos através do padrão XML*. Universidade Federal de Santa Catarina. Programa de pós-graduação em Ciências da Computação. 2002.

GOIS, T. S.; ZAUPA, A. P. *Integração de Esquemas Relacionais e XML com Realização de Consultas na Base de Dados Integrada*. Faculdade de Informática, Universidade do Oeste Paulista, Presidente Prudente, São Paulo, Brasil. 2010.

HAKIMPOUR, F.; GEPPERT, A. *Resolving semantic heterogeneity in schema integration: an ontology based approach.* Proceedings of the International Conference on Formal Ontology in Information Systems. Vol. 2001, p. 297-308. 2001.

JOSIFOVSKI, V.; RISCH, T. *Query Decomposition for a Distributed Object-Oriented Mediator System*. Uppsala Database Laboratory, Computing Science Department, Uppsala University, Uppsala, Sweden. 2002.

KAKUGAWA, F. R. *Integração de Bancos de Dados Heterogêneos utilizando Grades Computacionais*. Universidade de São Paulo. Dissertação de Mestrado. 2010.

KALINICHENKO, L. A. *Compositional Specification Calculus for Information Systems Development*. In Proc. of the East-West Symposium on Advances in Databases and Information Systems. Maribor, Slovenia. 1999. 16 p.

KANTORSKI, G. Z. *Interoperabilidade de Bancos de Dados Heterogêneos Através da WWW*. Porto Alegre: PPGC da UFRGS, Tese de Mestrado. 1999.

KOSSMANN, D. *The state of the Art in Distributed Query Processing*. ACM Computing Surveys. Vol. 32. Nº 4. 2000. P. 422-469.

LOSHIN, David. *Master data management*. Elsevier/Morgan Kaufmann, 2009. 274 p.

MORAES, L. F. *Reescrita de Consultas Utilizando Web Services*. Departamento de Informática, Curso de Ciência da Computação, Universidade de Santa Cruz do Sul, Santa Cruz do Sul, RS, Brasil. 2009.

MOSLEY, M.; BRACKETT, M.; EARLEY, S.; HENDERSON, D. *The DAMA Guide to The Data Management Body of Knowledge*. First Edition. 2009. 430 p.

OZSU, M. T.; VALDURIEZ, P. *Principles of Distributed Database Systems*. 2nd edition. Ed. Englewood Cliffs: Prentice-Hall. 1999. 845 p.

RAM, S.; RAMESH, V.; *Schema Integration: Past, Present and Future. Management of Heterogeneous and Autonomous Databases Systems,* Morgan Kaufmann Publishers. 1999. P. 119-155.

RIBEIRO, C. H. F. P. *Banco de Dados Heterogêneos: Mapeamento dos Esquemas Conceituais em um Modelo Orientado a Objetos*. Porto Alegre: CPGCC da UFRGS, Tese de Doutorado. 1995. 165 p.

SALGADO, A. C.; LÓSCIO, B. F. *Integração de Dados na Web*. Centro de Informática, Universidade Federal de Pernambuco (UFPE), Recipe, Pernambuco, Brasil. 2001.41 p.

SHETH, A. P. *Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics*. Wright State University, Ohio, United States.1999. P. 5-30.

SILVA, R. J. R. *Framework para a avaliação de ferramentas de Master Data Management*. Instituto Superior de Economia e Gestão. Universidade Técnica de Lisboa. Dissertação de Mestrado. 2012.

UCHÔA, E. M. A., MELO, R. N. *Integração de Sistemas de Bancos de Dados Heterogêneos Usando Frameworks*. Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro, Brasil. Tese de Doutorado. 1999.