UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

UM MODELO PARA A IMPLANTAÇÃO DE UM DATA MART DE CLICKSTREAM PARA EMPRESAS PROVEDORAS DE ACESSO À INTERNET DE PEQUENO E MÉDIO PORTE

Dissertação submetida à Universidade Federal de Santa Catarina
para a obtenção do Grau de Mestre em Engenharia.

LUÍS ROBERTO ZART OLANYK

Florianópolis, 22 de novembro de 2002

LUÍS ROBERTO ZART OLANYK

UM MODELO PARA A IMPLANTAÇÃO DE UM DATA MART DE CLICKSTREAM PARA EMPRESAS PROVEDORAS DE ACESSO À INTERNET DE PEQUENO E MÉDIO PORTE

Esta Dissertação foi julgada adequada para obtenção do Título de "Mestre em Engenharia", Especialidade em Engenharia de Produção e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia de Produção.

_____

Prof. Edson Pacheco Paladini, Dr.

Coordenador do Curso

Banca Examinadora:

_____

Prof. Oscar Ciro López Vaca, Dr.

Orientador

_____

Prof. Aran Bey Tcholakian Morales, Dr.

_____

Prof. José Leomar Todesco, Dr.

_____

Prof. Osmar de Oliveira Bráz Junior, M. Eng.

DEDICATÓRIA

Dedico este trabalho à minha esposa Luciane Benatti, que soube compreender que as horas roubadas da nossa convivência foram necessárias para a realização deste objetivo.

AGRADECIMENTOS

Expresso meus agradecimentos a UFSC – Universidade Federal de Santa Catarina, ao Departamento de Engenharia de Produção e Sistemas e a UNIOESTE – Universidade Estadual do Oeste do Paraná que, ao viabilizarem o Programa Interinstitucional de Pós-Graduação em Engenharia de Produção, com os mestrados fora de sede, permitiram às pessoas que não residem em Florianópolis participarem deste curso.

Agradeço também aos meus colegas de turma e aos professores que ministraram as disciplinas e tornaram esta jornada tão agradável e prazerosa, compartilhado de suas amizades e participando de suas vidas, dentro e fora da sala de aula.

Em especial, gostaria de agradecer a minha família que soube me incentivar nas horas de dificuldade, trazendo sempre palavras de esperança e motivação.

Finalmente, agradeço aos meus colegas Mabel Pereira da Silva, Luciano Olanyk, Luís Fernando Enciso e João Antônio Grande Neto, pelo sempre pronto apoio as tarefas necessárias à realização deste trabalho.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS

ASP	-	Active Server Pages
CFL	-	Formato de Log Comum
CGI	-	Interface de Gateway Comum
CRM	-	Gerenciamento do Relacionamento com o Cliente
DBMS	-	Sistema de gerenciamento de banco de dados
DNS	-	Servidor de Nomes de Domínio
DSL	-	Linha Digital de Assinante
DSS	-	Sistemas de Suporte a Decisão
DW	-	Data Warehouse
ECFL	-	Formato de Log Estendido
EIS	-	Sistemas de Informações Executivos
E/R	-	Entidade/Relacionamento
eRM	-	Gerenciamento do Relacionamento eletrônico
ETL	-	Extrair, Transformar, Carregar
GMT	-	Horário do Meridiano Greenwich
HTML	-	Linguagem de Marcação de Hiper Texto
HTTP	-	Protocolo de Transferência de Hiper Texto
HTTP	-	Protocolo de Transferência de Hiper Texto
IIS	-	Servidor de Informações de Internet
IP	-	Protocolo Internet
KDD	-	Descoberta de Conhecimento em Banco de Dados
OLAP	-	Processamento Analítico On-line
OLTP	-	Processos de Transações On-line
PSI	-	Provedor de Serviços Internet
SQL	-	Structure Query Language
TI	-	Tecnologia da Informação
WWW	-	World Wide Web

RESUMO

Data warehousing é um dos campos de Sistemas de Apoio a Decisão (SAD) com mais rápida expansão na recente Tecnologia da Informação (TI). A Internet, apesar de sua juventude, mostra-se como um superpovoado ambiente de informações e com um alto grau de competitividade. Com o intuito de ampliar o relacionamento com clientes que utilizam sites da Web o presente trabalho busca formular as bases para construção de uma ferramenta SAD que auxilie neste relacionamento.

No trabalho são descritos os conceitos referenciados na literatura para construção de um data warehouse de clickstream, demonstrando os requisitos necessários e citando os principais pontos onde diferentes soluções se aplicam, para que, com bases sólidas se verifiquem quais as melhores opções podem ser empregadas na implantação do projeto.

De acordo com a estrutura física da organização em estudo, um modelo de implantação de um data mart de clickstream é proposto. Buscando solucionar problemas de navegação e com o foco na busca por uma melhora do serviço prestado para os clientes da organização é executada a implantação do protótipo, o qual mostrou-se importante para auxiliar estas tarefas.

Alguns dos resultados obtidos são apresentados, demonstrando assim o poder do protótipo construído. Por fim são discutidas algumas recomendações para trabalhos futuros.

Palavras-chave: Data Warehouse de Clickstream, Provedor de Acesso à Internet, Sistemas de Apoio à Decisão.

ABSTRACT

Data Warehousing is one of the fields of the Decision Support Systems (DSS), with the most quickly expansion at the recent Information Technology (IT) area. The Internet, in spite of your youth, reveal as a super populated information environment and in high degree of competitiveness. Intending to amplify the relationship with costumer users of web sites, the preset dissertation tend to formulate the basis aiming to structure a DSS tool that assists in this relationship.

In this dissertation are described concepts with regard to reference books for the construction of a data warehouse of clickstream, demonstrating the necessary requirements and mentioning the main subject matters where different solutions should be applied, so as to, with solid basis verifying whose best options can be engage at the implantation of the project.

According to a phisical structure of the organization in study, a model of implantation of a data mart of clickstream is proposed. Aiming to solve problems of navigation and focusing a search for improvement in services for costumers of the organization it is started the implantation of the model in wich itself display important as a auxiliary tool of this tasks.

Several results achieved are introduced so that it demonstrates the power of the constructed model. At the end some recommendations are discussed for future researches.

Keywords: Clickstream Data Warehouse, Internet Service Provider, Decision Support System.

CAPÍTULO 1 – INTRODUÇÃO

1.1 Apresentação do Tema

Neste capítulo são apresentados: o tema abordado, os objetivos que se pretende alcançar e a metodologia utilizada. Após uma breve justificativa da importância e pertinência da utilização da tecnologia de data warehouse para a organização é apresentada uma descrição da estrutura dos capítulos que compõem o presente trabalho.

Os processos de tomada de decisão caracterizam-se pela necessidade de estabelecer relacionamentos complexos entre variáveis e geralmente envolvem grandes volumes de dados. Sistemas de Suporte a Decisão (DSS) são sistemas computadorizados utilizados para apoiar processos de tomada de decisão na realização de análises sobre dados, estabelecendo relações entre os mesmos com o objetivo de obter indicadores.

Nos últimos anos a abordagem de data warehouse vem sendo amplamente utilizada como forma de prover um ambiente adequado a estas análises. Segundo Inmon (1997, p.33): "Um data warehouse é um conjunto de dados baseados em assuntos, integrado, não-volátil, e variável em relação ao tempo, de apoio às decisões gerenciais".

A utilização da tecnologia de data warehouse como ferramenta de suporte aos sistemas gerenciais de tomada de decisões para *e-business* (negócios eletrônicos), pode contribuir para descoberta de conhecimentos relevantes que podem ser importantes às organizações que lidam com esse novo mercado.

Na Internet e mais precisamente nos aplicativos de publicação para sites na Web, que tem a responsabilidade da produção de conteúdo, informações e processamento de transação, mapear o comportamento das ações de usuários é uma tarefa bastante complexa e caracteriza o que se pode chamar de suporte a decisão não convencional, envolvendo uma grande quantidade de fatores variáveis a serem tratados.

Com o intuito de esclarecer essa abordagem, este trabalho propõe um modelo para a

implantação de um data warehouse (DW) de clickstream (seqüência de cliques) em um provedor de acesso à Internet.

1.2 Objetivo Geral

O objetivo geral deste trabalho é propor um modelo para o desenvolvimento de um data mart de clickstream voltado para um provedor de serviços Internet.

1.3 Objetivos Específicos

- Identificar os principais problemas que podem ser resolvidos com a aplicação das técnicas de DW de clickstream;
- Coletar os elementos teóricos relacionados às técnicas de DW voltados para clickstream;
- Projetar o site da Web em questão para suportar DW de clickstream;
- Avaliar comparativamente as diversas ferramentas de extração de logs e transformação disponíveis;
- Validar a proposta através da construção de um protótipo de DW específico para análise de clickstream;
- Identificar problemas com relação à implementação da tecnologia de DW de clickstream;
- Promover ações a partir dos resultados obtidos com o modelo.

1.4 Justificativa

O serviço fornecido pelos provedores Internet em todo o mundo, ressalvando-se as

peculiaridades de cada mercado nacional, é semelhante em sua forma e operação. Estes serviços são oferecidos para clientes residenciais, comerciais e industriais.

A utilização de novas tecnologias e produtos tem que ser uma constante para que essas organizações tenham um diferencial competitivo diante da selvagem realidade do mercado, e medir os resultados da utilização desses recursos não é uma tarefa simples. Decisões empíricas devem ficar em segundo plano e a utilização de dados reais priorizada, para que, a tomada de decisão seja o mais objetiva e precisa quanto possível.

A implantação da tecnologia de DW de clickstream em um provedor de serviços Internet pode minimizar as deficiências das ferramentas de análise de sites da Web que não fornecem análises que envolvem seqüências temporais e podem auxiliar no aprendizado do comportamento de clientes em estudo. Aprender com os clientes é valioso e pode resolver uma grande quantidade de questões como, por exemplo: problemas técnicos envolvendo páginas da Web e o que é relevante para o usuário que navega no site.

Conforme esse conhecimento se consolida o relacionamento com o cliente pode gerar satisfação tanto para organização quanto para os usuários, contribuindo para o sucesso do serviço prestado.

1.5 Metodologia

Para alcançar os objetivos propostos as etapas seguintes que juntas integram o trabalho, foram enumeradas:

1. Levantamento de subsídios teóricos (livros, artigos, *papers* e pesquisa na Internet) acerca de data warehouse de clickstream e das principais metodologias empregadas nesta técnica;

2. Conhecimento das características de um provedor de serviços Internet onde o estudo será aplicado;

3. Conhecendo as técnicas mais utilizadas, aplicar as metodologias estudadas na

concepção de uma proposta de um modelo de DW;

4. Validação, mediante a construção de um protótipo de um data mart de clickstream;

5. Apresentar as conclusões obtidas da pesquisa, bem como sinalizar algumas recomendações para trabalhos futuros.

1.6 Estrutura do Trabalho

O primeiro capítulo traz uma visão geral do trabalho, seus objetivos, justificativas, metodologia e a estrutura do trabalho.

No segundo capítulo são apresentados os problemas que circundam os sistemas de suporte a decisão na administração de provedores de serviços Internet, além de relacionar algumas soluções possíveis para o problema.

No terceiro capítulo é apresentada a fundamentação teórica de projetos de sistemas de informações baseados em data warehousing de clickstream, com o objetivo de descrever sucintamente os principais conceitos necessários para a implantação da técnica.

No quarto capítulo são detalhadas as características da pesquisa com a definição dos termos, os pontos chaves, forma e delimitação do estudo. São apresentadas considerações em relação ao seu emprego, tendo em vista as metodologias de desenvolvimento utilizadas.

No quinto capítulo é apresentado o estudo de caso realizado a partir da arquitetura proposta para um provedor de acesso à Internet, com a análise dos resultados obtidos.

No sexto capítulo são apresentadas a conclusão e recomendação para trabalhos futuros.

CAPÍTULO 2 – O PROBLEMA

2.1 Introdução

Neste capítulo é apresentada a trajetória da Internet comercial no Brasil a partir do fornecedor desse produto – o provedor de serviços Internet (PSI), desde a implantação dos provedores até os dias de hoje, descrevendo os principais serviços oferecidos e ressalvando as principais dificuldades encontradas nessas organizações. Também é proposta uma ferramenta de suporte a decisões, justificando sua necessidade e os benefícios que poderá trazer para o PSI.

2.2 Identificação do Problema

No primeiro momento da Internet comercial no Brasil, distinguiram-se dois tipos de empresas, a Embratel, na época uma estatal que trazia ao Brasil a tecnologia de interconexão de redes operando tanto no atacado como no varejo, e os pequenos empreendedores, como por exemplo as empresas DGLNet em Campinas-SP e a Netcertto Informática em Cascavel-PR, que formavam a grande maioria e foram responsáveis pela rápida formação capilar da rede Internet no Brasil. Essas empresas possuíam baixo capital e poucas pessoas envolvidas e, em geral, foram criadas para serem exclusivamente provedores de acesso à Internet.

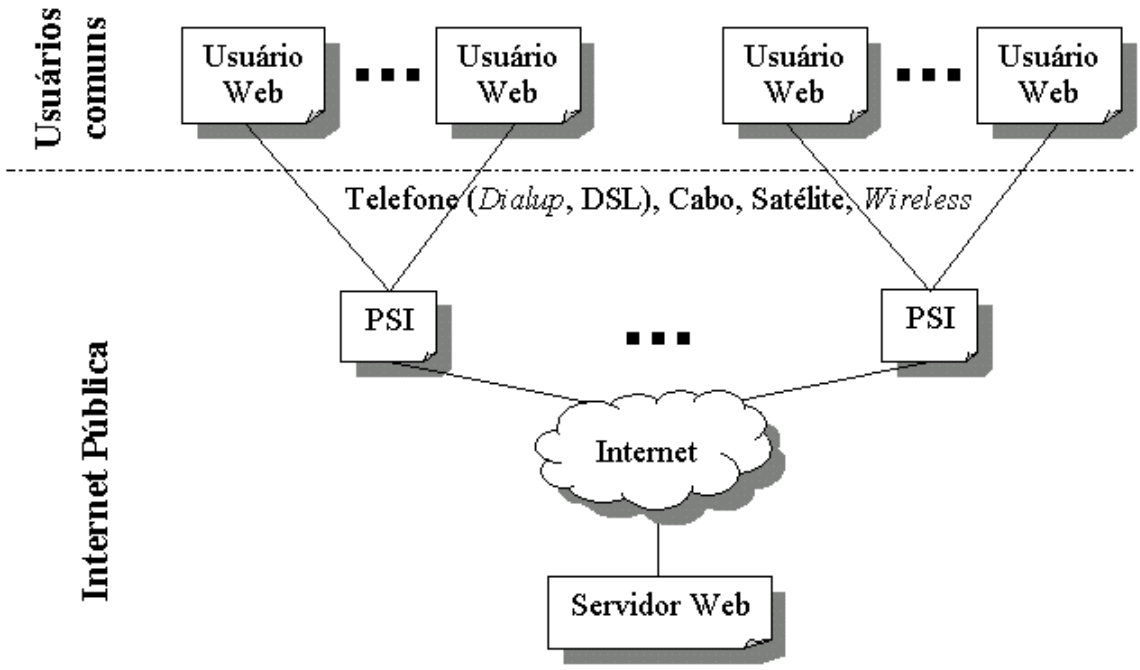
O setor de acesso comercial à Internet no Brasil é bastante recente. Desde sua implantação, em meados de 1995 tem passado por transformações ambientais profundas. As empresas estatais de telecomunicações foram proibidas de comercializar os serviços de Internet no varejo, com o intuito de promover competitividade entre a iniciativa privada. Num segundo momento, entraram em um processo de privatização.

Estas transformações sempre ocorrem de forma muito rápida e exigem que os executivos possuam capacidade para fazer com que suas organizações consigam se adaptar a tais mudanças no ambiente.

Enquanto que a maioria dos provedores de conteúdo público na Internet se encarregava de disponibilizar um mar de informações na rede, os últimos metros entre o usuário dos sistemas e a Internet eram providos pelos provedores de acesso à Internet.

A figura 1 mostra o modelo de funcionamento dos provedores de acesso à Internet, através de diferentes meios como: *dialup* (acesso discado) e DSL (Linha digital de assinante) fornecido por companhias telefônicas, a cabo fornecido por redes de TV, satélite e conexões *wireless*.

Figura 1 - Configurações Típicas dos Provedores de Acesso à Internet



Fonte: Adaptado de Sweiger, 2002.

Com a entrada dos grandes grupos no mercado nacional, como o Universo On-Line (UOL), uma empresa de capital aberto e a América On-Line (AOL), uma multinacional do ramo, percebeu-se a necessidade de oferecer algo mais do que simplesmente o acesso à Internet aos clientes. Oferecer algum valor agregado, tal como conteúdo exclusivo ou especializado ou ainda um suporte técnico personalizado aos clientes. Esses serviços passaram a ser uma forma importante de fidelização para as pequenas e médias organizações.

Os pacotes tradicionais de serviços oferecidos por um PSI podem ser classificados em:

- Estrutura diversificada de provimento de acesso à rede Internet;

- Serviços aplicativos locais: *e-mail* (correio eletrônico), *chat* (salas de conversa), site (portal), entre vários outros;

- Hospedagem de domínios e sites comerciais;

- Hospedagem de *e-commerce* (comércio eletrônico).

O que antes era um território virgem e pouco habitado hoje se mostra como um superpovoado ambiente de informações, com um alto grau de competitividade. Além do que, as empresas precisam dar ênfase ao crescimento da rentabilidade de forma consistente.

Um dos caminhos para a prosperidade pode estar em construir ferramentas para cultivar a Internet, e isso é possível através da construção de um data warehouse de clickstream.

2.3 Porque Utilizar Data Warehouse

Durante a década de 90, antes da revolução da Web realmente tomar força, as empresas de Tecnologia da Informação – TI aprenderam como publicar os ativos da empresa, que estão na forma de dados, para analistas internos e para a gerência. Essa publicação é a tarefa central do data warehouse, cumprindo a promessa de "extrair os dados" depois que os sistemas operacionais baseados em OLTP (Processos de Transações On-Line) "inserem os dados" (Kimball e Merz, 2000).

Em um ambiente de *e-business*, a TI não simplesmente serve o negócio, mas deve executar o negócio – este é o negócio. Está é a grande mudança comparada com os modelos mais antigos de negócios (Sweiger, 2002).

No ambiente da Internet, todas organizações têm acesso às mesmas ferramentas e a mesma infra-estrutura, tudo em um custo relativamente baixo. Embora muitos *e-business*

consigam uma vantagem inicial em negócios, dificilmente conseguem manter essa vantagem ao longo do tempo não fornecendo assim um diferencial competitivo consistente.

Existe uma riqueza de informações acumuladas por servidores da Web. Utilizar estes registros para melhorar a performance e usabilidade de um site da Web, carregando esses registros em um banco de dados especial para aprender como pessoas navegam no site e como a performance pode ser melhorada é um dos benefícios do DW de clickstream (LURIE, 2001).

A grande ironia é que os dados do clickstream são relativamente fáceis de se colecionar. Diferentemente dos dados de sistemas legados, clickstream são automaticamente registrados por todos servidores populares da Web em vários formatos padronizados por seus fornecedores.

Um passo intermediário no caminho do DW de clickstream é a utilização de ferramentas de analise de arquivos de log registrados por servidores da Web, produtos comerciais ou de uso livre como: WebTrends®, Analog® ou NetTracker®. Essas ferramentas são boas especialmente para fornecer informações estatísticas como, entrada do site e páginas de saída do site, e outros agregados estatísticos brutos. Mas normalmente não conseguem fornecer analises que envolvem seqüência de tempo, mostrando o que acontece durante o curso de uma visita ao site ou quais as tendências globais dessas estatísticas através do tempo.

2.4 O Objetivo é a Satisfação do Cliente

Um PSI pode utilizar o data warehouse de clickstream para analisar o escopo de uma completa sessão de browser, ao qual envolvem todos os endereços visitados de seu domínio por um usuário, durante uma sessão. Essa análise do conjunto de sessões de clickstream no nível do PSI pode fornecer uma visão do comportamento e das preferências desta população de usuários no ambiente Web.

Patrícia Seybold (1998), em seu clássico livro “Customer.com”, identifica oito fatores de sucesso para qualquer negócio, mas especialmente para negócios alavancados na Web. Esses fatores incluem: Ter como alvo o cliente certo, possuir a experiência total do cliente, simplificar processos de negócio que têm impacto sobre o cliente, fornecer uma visão de 360

graus do relacionamento com o cliente, deixar que os clientes se sirvam, ajudar os clientes a fazer seus trabalhos, entregar serviços personalizados e promover comunidade.

Os fatores de sucesso de (Seybold, 1998) são um tipo de fundação para o gerenciamento do relacionamento com o cliente – CRM. Os objetivos de CRM são construir lealdade de clientes, aumentar o lucro e integrar transparentemente cada função de negócio do cliente.

Toda empresa terá de aprender a adequar seus produtos sob medida ao cliente, porque é a vantagem mais importante do marketing um-a-um. Se a empresa apenas identificar, diferenciar o cliente e interagir com ele, o cliente receberá o serviço, mas ainda estará à mercê, estrategicamente, dos concorrentes de custo menor. Mas se adaptar sob medida ao cliente poderá dar uma vantagem real, graças ao que sabe dele. Talvez não modifique o produto, mas terá de adequar algum aspecto de seu negócio (ROGERS, 2000).

Construir um bom ambiente de *front-end* na Web é possível com pouco investimento e em pouco tempo. Sistemas de *Back-end* semelhantes a servidores de transação e os DW de clickstream são construções bem mais complexas, além de ter o custo elevado de implantação. A boa notícia é que formam o diferencial das TI, dando a chave de um sucesso duradouro em um *e-business* (Sweiger, 2002).

Os DW de clickstream devem ter objetivo maior do que analisar tráfego. Ele deve servir de ferramenta para o relacionamento do *e-bussines* com o cliente, uma nova concepção chamada de Administração do Relacionamento eletrônico, ou simplesmente eRM. O ambiente da Web é rico em dados eRM – o que se precisa é explora-los (Sweiger, 2002).

Todo o data warehouse deve se esforçar para ser um verdadeiro sistema de apoio à decisão (*decision-support system* – DSS). De nada adianta ser sábio sem ação. O resultado tangível, final, real de qualquer data warehouse deve ser igual a decisões tomadas como resultado do conhecimento adquirido (Kimball, 1998a).

2.5 Análise dos Dados que vem da Web

A *World Wide Web* – WWW, é proveniente de experiências que foram conduzidas nos laboratórios da CERN (*European Organization for Nuclear Research*), na Suíça (CERN, 2002). Uma equipe de programadores dedicou seu tempo a investigar e criar um modo de transmitir informações em um formato que se tornou conhecido como *Hypertext Transport Protocol* – HTTP (Homer, 2000).

Usando uma linguagem de marcação chamada de *Hypertext Markup Language* – HTML, projetada com simplicidade e com uma estrutura flexível se permitiu que textos e imagens gráficas fossem exibidos em um navegador da Web ou em outra aplicação adequadamente habilitada. A exibição de informações na Web utiliza um meio para proporcionar a localização, chamado de *Universal Resourse Locator* – URL.

Um modelo de aplicação básico da Web é completamente sem estado. Esse modelo não usual de aplicação usa um novo paradigma de programação que é estranho ao ambiente cliente/servidor, e precisa ser entendido antes dos procedimentos de captura de dados de clickstream em ambientes de aplicativos Web serem realizados (Kimball e Merz, 2000).

Quadro 1 - Principais Métodos de Requisito do Protocolo HTTP

MÉTODO	DESCRIÇÃO	SINTAXE
GET	Obtém uma página Web.	GET URL HTTP- versão.
HEAD	Devolve somente o cabeçalho HTTP e não o conteúdo da página Web.	HEAD URL HTTP- versão.
POST	Um formato de envio de campos de um formulário HTML de um cliente para um servidor Web.	POST URL HTTP- versão.

Fonte: Apache, 2002.

Existem diferentes métodos que são da terminologia de orientação a objetos para tipos de transações implementadas no protocolo HTTP, três destes são relevantes para análise de clickstream (Sweiger, 2002), e são apresentados no quadro 1.

Um cliente de um site da Web utilizando um navegador pode acessar uma página, desconectar da rede, reconectar, e o seu acesso à próxima página no site não perde a continuidade. A única informação requerida para acessar a página é a URL, e nenhum estado é requerido para mostrar essa URL.

Existem vários caminhos para passar parâmetros entre transações HTTP, estas técnicas auxiliam na avaliação de estado de uma sessão. Ferramentas como cookies ou *strings* de consultas podem ser utilizadas para ajudar na identificação de usuários.

Uma *string* de consulta passa seus parâmetros anexados a URL, começando depois de um ponto de interrogação (?) e são separados pelo caractere (&) para identificar espaçamento. Por exemplo, em uma busca através do site especializado Google, a URL teria a seguinte sintaxe: http://www.google.com/search?hl=br&lr=lang_br&safe=off&q=certto+internet. Note que os valores de busca passados pela variável "q" são "certto+internet" e as outras variáveis "hl, lr e safe" são utilizadas para setar a *string*.

Outras técnicas que utilizam programas *scripts* podem ser utilizadas. Para que um *script* funcione com um servidor da Web é necessário que haja algum tipo de aplicação intermediária. Ele precisa ser capaz de aceitar uma solicitação do usuário, ler e interpretar o arquivo de *script* no servidor apropriado, criar a página de saída e comunicar para o servidor da Web para onde ela foi enviada como resposta para o cliente (Homer, 2000).

No início da Web o mecanismo utilizado para executar outra entidade através do HTML foi o Common Gateway Interface – CGI. Hoje, CGI não é o único caminho para executar códigos em resposta a requisição de páginas. Isto também é possível utilizando-se linguagens de *script* executadas por um pré-processador HTML, semelhantes ao *Microsoft Active Server Pages* – ASP ou ao UNIX/Linux-*oriented Hypertext Pré-processor* – PHP. Outra opção é o Java, que pode ser usado em parceria com o *Java Server Pages* – JSP (Sweiger, 2002).

O ambiente de aplicação da Web é composto por um grande conjunto de tópicos, entre eles estão o HTTP, *strings* de consulta, cookies, registro de logs, CGI, linguagens de *script*, servidores Web e servidores de aplicação. Alguns tópicos devem ser analisados mais profundamente, pois formam a base para a fonte de dados que habitará o DW de clickstream.

2.5.1 Identificando Fontes de Dados Potenciais da Web

Existem basicamente três fontes primárias de informações dos sites da Web para o povoamento do data warehouse de clickstream:

1. Conjunto de arquivos de logs;

2. Ferramentas de análise;

3. Servidores de conteúdo externos.

A primeira delas é o conjunto de arquivos de logs gerados por servidores de site da Web, estes logs foram projetados originalmente para prover estatísticas administrativas para sistemas Web. Outras duas fontes de dados de logs incluem dados que podem ser capturados da rede local – dados internos e também da Internet – dados externos, através de ferramentas específicas.

A extensão dos dados armazenados em arquivos de logs determina a qualidade e integridade da informação avaliada no warehouse de clickstream. A maioria dos softwares servidores da Web permite perfeitamente controlar quais elementos de dados devem ser logados (Kimball e Merz, 2000).

Os softwares servidores da Web mais comuns implementam pelo menos um dos três padrões de formatos mais importantes da Web: NCSA – Formato de log comum (CLF), NCSA – Formato de log estendido (ECLF) ou W3C – Formato de arquivo de log estendido (ExLF).

O primeiro desses padrões é o mais antigo e contém uma quantia mínima de dados. Servidores como Apache, NCSA, e Netscape utilizam o formato CLF como default, mas a Microsoft é uma exceção. Ela decidiu criar um formato proprietário para seu servidor, o *Internet Information Server* – IIS (Sweiger, 2002).

Um exemplo do formato de log – CLF é ilustrado no quadro 2 – linha 1, e se utilizou um servidor Web Apache (Apache, 2002). A interpretação para essa inserção é a seguinte: Nós

sabemos que o usuário “Aníbal” utilizado-se do endereço IP 200.250.43.26, fez uma consulta ao site e obteve sucesso (200) na busca do arquivo index.htm em 23 de Março de 2002 às 23h34min, e que 423 bytes de código foram transferidos. Neste caso, o usuário é identificado porque fez uma autenticação anterior, não sendo um fato comum na navegação.

Quadro 2 - Armazenamento de Log no Formato CLF

200.250.43.26	-	Aníbal	[23/Mar/2002:23:34:08 –0300]	“GET /index.htm HTTP/1.1”	200	423

Fonte: Servidor Apache da Web – Empresa Netcertto Informática Ltda.

Estes são apenas alguns dos elementos de dados que podem ser armazenados em arquivos de logs de servidores do site da Web. O anexo A trás um conjunto de elementos de dados que podem ser relevantes na construção de um DW de clickstream utilizando o formato CLF ou seu formato estendido ECLF.

A segunda fonte de dados para o warehouse é interna a uma empresa e relata, na sua maioria, a usabilidade dos dados da rede. Estas ferramentas podem ser classificadas como analisadores de arquivos log e site Web, analisadores de rede utilizando técnicas *sniffer* – de escuta, servidores *proxy* e *firewalls*, servidores de *cache*, servidores de mídia e servidores de aplicação Web.

A terceira e última fonte de dados para o DW é externa e pode ser classificada como tendo serviços de conteúdo de *cache*, aplicativos de parceiros da empresa, dados *on-line*, dados de negócio ou associação de consumidores (Sweiger, 2002).

2.5.2 Mecanismos de Identificação de Usuários

O histórico de clickstream contido em um warehouse expressa as ações naturais que os usuários fazem quando visitam um site da Web. Para que se estabeleça uma identificação durante uma sessão para estes usuários, necessária para uma análise temporal, é preciso utilizar alguns mecanismos que podem criar uma identidade. Existem basicamente três técnicas para

avaliar o estado de uma sessão HTTP (Sweiger, 2002).

A primeira dessas técnicas é o uso de cookies. O propósito é manter uma identidade para cada visitante, independentemente de seu ponto de entrada no complexo de sites da Web de uma organização. É sugerido o uso de um único cookie com ID (identificador) para cada usuário codificado e que contenha somas de verificação ou talvez, até mesmo, códigos para correção de erros para que se possa detectar e descartar cookies que foram alterados pelo usuário ou por software de proteção (Kimball e Merz, 2000).

Quadro 3 - Parâmetros do Protocolo Cookie

Name	O nome usado para referenciar o cookie
Value	O valor da assinatura para o cookie no Web site
Domain	O domínio da Internet do servidor que tem acesso ao cookie
Path	O caminho URL para o qual o cookie é válido
Expires	Data/hora de expiração do cookie (GMT)
Secure	Um flag que indica que um cookie é usado somente com uma conexão segura

Fonte: Adaptado de IETF, 2002.

O protocolo cookie foi desenvolvido pela Netscape como uma solução para problemas de estado. O quadro 3 mostra os parâmetros que podem ser programados nesse protocolo (IETF, 2002).

Segundo Sweiger (2002), armazenar logs de cookies separadamente dos logs de servidores de sites da Web não é uma boa prática porque pode fazer o processo de extração de dados mais complexo. Os analistas terão que ficar atentos as dificuldades do trabalho de sincronização de ambos logs quando tiverem que fazer a extração para o warehouse.

Um dos problemas da utilização de cookies é que ele pode ser desabilitado. Um mecanismo alternativo é utilizar uma *string* de consulta, com informações passadas de volta para o servidor através de um método de submissão. Por exemplo, se o usuário Aníbal, citado no quadro 2, já efetuou um login no site da Web e esta navegando, mas com parâmetros de

cookie desabilitados em seu browser, é possível utilizar a *string* para link de página apresentada no quadro 4.

Quando o usuário clicar no link, ele enviará ao servidor da Web, as informações úteis para o gerenciamento de seu estado (Sweiger, 2002).

Quadro 4 – Exemplo de *String* de Consulta com Passagem de Parâmetros

```
<A HREF=”próxima_pg.htm?username=Aníbal&logged_in=true”>Vá para próxima página</A>
```

A última técnica para passar informações de estado é utilizar campos de formulário oculto. O usuário não consegue visualizar esses campos na página da Web e esses valores são enviados ao servidor da Web assim que um usuário submeta um formulário, como se fosse um procedimento normal de envio de formulário.

2.5.3 Definindo o Projeto do Web Site

É importante tratar como objetivo monitorar o usuário e suas ações do momento que ele entra no site da Web até o momento em que ele sai, atribuindo um identificador (ID) de sessão para o registro dessas ações e, se desejável, identificando o usuário. O comportamento de usuário durante uma visita a um site da Web pode fornecer pistas valiosas sobre a eficácia do site, bem como sobre os hábitos de navegação do usuário.

Tendo compreendido o mapa de sessão de um usuário e atribuído a ele um ID, pode-se associar a sessão outras ações relacionadas ao usuário que permitirão enriquecer o perfil da sessão e completar toda a coleta de informações necessárias para a futura personalização do site. As técnicas de monitoramento de site da Web fornecem a base para a personalização de site.

O webmaster da empresa fornece ferramentas de navegação que abarcam servidores departamentais e fornecem sistemas de pesquisa e sumário da empresa que podem ser integrados. Mas provavelmente, não será o arbitro final do conteúdo. Suas responsabilidades e

deveres não podem incluir o incentivo ou a autoridade para assegurar a captura de dados completamente adaptados para os data marts e warehouses. É preciso estabelecer data marts para divisões e departamentos individuais. Depois, assegurar que os dados possam ser adaptados em um DW central com o acordo e o auxílio do webmaster corporativo e de outros elementos-chave dentro da empresa (Kimball e Merz, 2000).

O fato de simplesmente monitorar a venda de um produto ou o foco de trabalho de uma empresa na Web pode mascarar muita das atividades auxiliares de coletar informações, "olhar vitrines" e descarregar amostras são parte de uma experiência de agradar o cliente. E, naturalmente, a implicação é que essas atividades auxiliares também iluminam consideravelmente as preferências do cliente. O clickstream contém detalhes nunca vistos sobre os "corredores" percorridos pelo cliente antes da verificação final – *check out* (Kimball e Merz, 2000).

2.5.4 Precauções na Web para Suportar DW de Clickstream

Existem algumas questões que, se não tratadas desde o começo, farão da análise de clickstream algo difícil, até mesmo sem sentido, para aplicações de DW. É preciso descrever as precauções que devem ser tomadas para tornar o site da Web algo muito eficiente para suportar o DW de clickstream. Segundo Kimball e Merz (2000), os seguintes tópicos devem ser considerados:

- Sincronização de servidores: É preciso considerar a probabilidade de que, ao longo do tempo, a empresa esteja servindo a Web a partir de vários servidores distribuídos. O DW capacitado para a Web coleciona as informações de eventos com registro de data/hora de muitas fontes diferentes, tanto de dentro quanto de fora da organização. Analisar o comportamento dos usuários depende de saber exatamente quando um evento ocorreu em relação ao outros eventos. Isso requer manter um padrão preciso de tempo através de todas as fontes de dados. A exatidão que se precisa não é obtenível pelo ajuste manual de um relógio. Esse nível de exatidão requer ferramentas sofisticadas de sincronização.

- Rótulos de conteúdo para páginas: É essencial que os rótulos de conteúdo sejam

desenvolvidos para permitir que os eventos de página sejam classificados e codificados para serem analisados posteriormente. Os atributos para uma página precisam ser projetados por uma equipe conjunta de desenvolvedores de sites e de DW e, então, disponibilizados para o sistema de extração de log da Web de modo que as páginas possam sempre ser descritas precisa e completamente.

- Cookies consistentes: Existem diferentes maneiras de utilizar o servidor de cookie. Um método de obter um ID de usuário encapsulado no cookie é colocar em cada potencial página de entrada departamental ou divisional um campo que chame o servidor de cookie para ler o cookie de nível corporativo.

Todos os tópicos abordados anteriormente devem ser planejados e implementados antes de começar o processo de extração dos dados para o data warehouse.

2.6 Conclusão

Neste capítulo, apresentou-se a organização envolvida na pesquisa, classificando suas principais dificuldades e onde a tecnologia para DW de clickstream pode ser utilizada no auxiliar da tomada de decisão.

Devido a complexidade dos dados que serão utilizados no povoamento do data warehouse, a análise dos dados que vem da Web mostra os pré-requisitos necessários a estruturação do site da Web como HTTP, *strings* de consulta, cookies, registros de log e CGI, identificando as fontes de dados que povoarão o DW.

São enumeradas as precauções para construção do site da Web a fim de suportar a implementação do DW de clickstream com o objetivo de formação de um banco de dados que possa ser analisado e os resultados da consulta sejam condizentes com a realidade.

O objetivo foi fundamentar o problema de pesquisa, de modo a orientar quais os principais pontos que deverão ser abordados na revisão da bibliografia no próximo capítulo para que se possa propor um modelo condizente com as necessidades da organização.

CAPÍTULO 3 – DATA WAREHOUSING DE CLICKSTREAM

3.1 Introdução

O impacto da Web transformou a missão da Tecnologia da Informação (TI), passando de mero suporte a aplicativos de legado para as capacidades de produção de conteúdo, informações e processamento de transações, tudo isso através de interfaces de navegador. A Web é muito mais que uma tecnologia para conectar dispositivos de processamento distribuídos, sendo uma nova e mais barata forma de comunicação.

Neste capítulo é abordada a técnica de construção de um data warehouse de clickstream, seus principais elementos, fases da construção e aspectos de implementação, a partir da visão de diversos autores. O objetivo é fundamentar a apresentação da arquitetura para solução do problema no capítulo 4 e a discussão da aplicação do protótipo no capítulo 5.

3.2 Definição

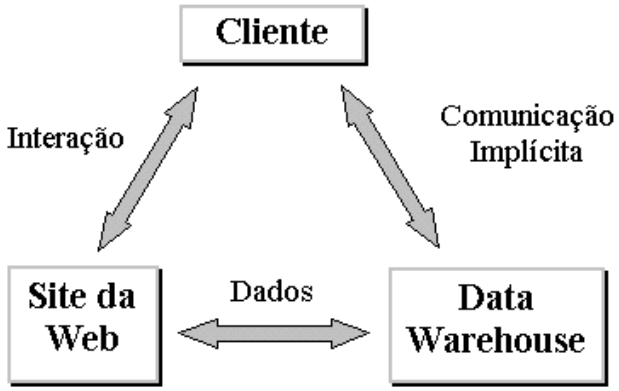
A audiência para dados de data warehouse cresceu do gerenciamento interno para a inclusão de clientes, parceiros e um grupo muito maior de funcionários internos. O foco da Web na "experiência do cliente" fez com que muitas empresas se tornassem muito mais conscientes da necessidade de aprender sobre o cliente e de fornecer informações úteis a este cliente. A fim de estar emparelhado com esta responsabilidade, o data warehouse deve ser ajustado para suportar os dados que vem da Web.

A figura 2 mostra a interação do cliente com o site da Web relacionando-se diretamente. Com o acumulo de dados no warehouse essa interação pode evoluir para um melhor relacionamento com o cliente, a partir da utilização de técnicas para a busca de conhecimento como, por exemplo, pesquisas estatísticas ou a utilização da técnica de Data Mining.

A Web é uma imensa fonte de dados comportamentais, indivíduos interagem com

informações por meio de seus navegadores aleatoriamente, por qualquer site que esteja hospedado na Internet armazenando um rastro nos servidores que disponibilizam este serviço. Embora esses dados de seqüência de cliques (clickstream), em muitos casos, estejam em um estado bruto e não tenham uma aparência adequada, eles têm o potencial de fornecer detalhes nunca imaginados sobre cada gesto efetuado por cada ser humano que esteja utilizando a mídia da Web.

Figura 2 - O Cliente, o Site da Web e o Data Warehouse



Fonte: Adaptado de Kimball e Merz, 2000.

O data warehouse não pode ser centralizado, assim com à Internet não pode sê-lo. De alguma forma deve-se adotar uma filosofia de projeto que permita que ilhas separadas de data warehouse através da Web vejam e se comuniquem umas com as outras de forma eficiente evitando problemas de inconsistência de informações.

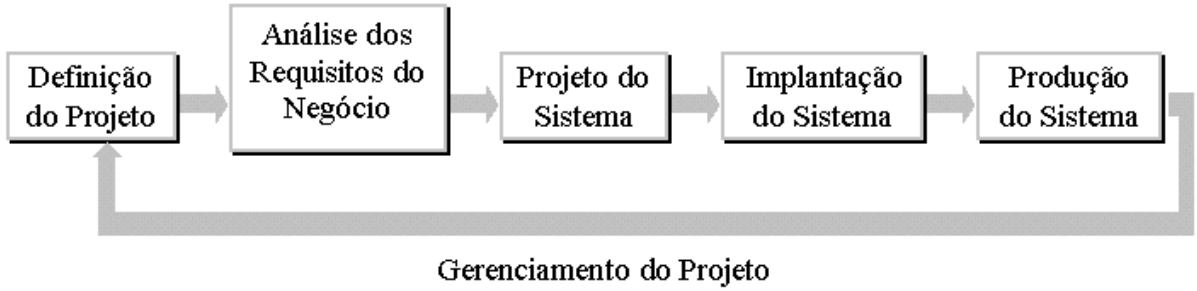
3.3 Fases Típicas do Desenvolvimento em um DW de Clickstream

O desenvolvimento do DW de clickstream pode ser decomposto em fases distintas que são ilustradas na figura 3 e são citadas em Kimball, (1998b) e também em Sweiger, (2002).

Pode se perceber que, uma vez que o sistema está em produção, o ciclo de

desenvolvimento começa novamente. O ciclo de desenvolvimento de um DW de clickstream nunca termina e sim se renova buscando sempre estar atualizado com as necessidades de quem utilizará o sistema.

Figura 3 - Fases Típicas do Desenvolvimento de um DW



Fonte: Adaptado de Kimball (1998b) e Sweiger (2002).

A visão global do desenvolvimento do DW de clickstream, ilustrado na figura 3, mostra as principais fases para a administração do projeto e a equipe de desenvolvimento. É preciso decompor essas fases detalhadamente, para apresentar de uma maneira clara algumas das dificuldades encontradas na implementação do data warehouse. Esses tópicos são apresentados nas sessões seguintes.

3.3.1 Necessidade do Data Warehouse

As fases mais importantes do projeto estão na definição do projeto, que envolvem o planejamento e a análise dos requisitos do negócio, porque elas dão a direção para todo o desenvolvimento que se segue (Sweiger, 2002).

A declaração da missão da empresa também é considerada um passo importante no processo de planejamento, ela deve definir as metas da empresa em termos comerciais e, dessa forma, definir também um claro propósito de investir em um data warehouse. Além disso, documentar a declaração da missão de uma organização é essencial para garantir que o

processo de implementação se ajuste aos propósitos gerais do plano, assim como aos objetivos da organização (Harrison, 1998).

No processo de planejamento do data warehouse é importante equilibrar todas as tarefas individuais a serem realizadas na construção e suporte do sistema. É preciso compreender as pessoas e os papéis envolvidos com perspectiva, discernimento e flexibilidade, pois nenhum projeto jamais seguirá completamente o plano original (Kimball, 1998b).

Os componentes principais do estágio inicial de desenvolvimento do DW são: A definição do projeto, a identificação dos papéis, o planejamento das atividades e os requisitos do negócio e os dados da auditoria. Cada um desses tópicos é abordado a seguir.

3.3.1.1 Definição do Projeto

O primeiro passo em um projeto é entender por que a organização precisa de um DW de clickstream e o quanto essa necessidade é sentida na organização. Os interessados no projeto deveriam responder algumas questões como, "O que eu procuro para analisar, e porque eu preciso analisar isso?" Para responder o porque dessas questões, é necessário um entendimento dos requisitos que devem ser endereçados (IBM, 1998).

É preciso fazer uma lista dos diferentes departamentos e executivos interessados que podem reconhecer o potencial desses dados. Mais tarde no projeto, é preciso entrevistar esses executivos e departamentos para entender detalhadamente suas necessidades, e para certificar que eles realmente estão interessados no projeto (Kimball, 1998a).

Segundo Kimball e Merz (2000), para que o projeto de DW possa ser agilizado, é preciso examinar cinco indicadores chaves que são citados em seguida.

- Forte envolvimento da gerência;

- Uma grande motivação para o negócio, fazendo com que os usuários finais enxerguem as pérolas do conhecimento que os dados provavelmente conterão;

- Uma boa parceria entre a TI e o negócio. A TI compreende e envolve-se com os

usuários do negócio e esses mesmos usuários finais respeitam e apreciam a competência que a TI traz para o negócio;

- Uma cultura analítica de suporte, onde exista uma tradição de gerenciamento baseado em números e confiança em dados para revelar tendências e desenvolvimentos importantes no negócio;

- A existência de dados reais em um sistema operacional real, contendo conteúdo de qualidade suficiente para servir como base para o DW.

Uma das principais razões para a definição do escopo do projeto é prevenir-se de constantes mudanças através do ciclo de vida como o surgimento de novos requisitos. No data warehousing, a definição do escopo dos requisitos deve ter atenção especial. Entretanto, duas das chaves mais valiosas do DW são sua flexibilidade e sua habilidade para manusear consultas desconhecidas. Por conseguinte, é essencial que o escopo seja definido para reconhecer que, o que o DW deverá entregar estará incluso nas definições iniciais do projeto (IBM, 1998).

3.3.1.2 Identificando Papéis e Planejando as Atividades

Em qualquer projeto de data warehouse sua construção requer o envolvimento de vários papéis separados. Alguns são permanentes e centrais para o projeto e outros somente são necessários durante um período determinado de tempo (Kimball, 1998b).

É imensamente útil listar todos os papéis no início do projeto de um DW, pensando em quando e como preencher esses papéis. Os principais papéis relacionados são apresentados no anexo B, de acordo com a ordem em que entram em cena e no decorrer do projeto.

No planejamento do projeto serão detalhadas as tarefas, prazos, status e responsáveis. Na especificação das tarefas deve-se atentar para o maior desmembramento possível em sub-tarefas, a fim de ter um maior controle e na estimativa de tempo (Kimball, 1998b) e (Inmon, 1997).

3.3.1.3 Análise dos Requisitos do Negócio e os Dados de Auditoria

Os tipos de requisitos necessários para um DW são diferentes daqueles utilizados em sistemas transacionais. Quase todo o foco das informações necessárias em um DW está no processo de decisão de marketing para o usuário final, contrário aos sistemas transacionais, onde o foco está nas funções de performance do processo (Sweiger, 2002).

O centro da metodologia da busca de requisitos para um DW de clickstream apóia-se em entrevistas com usuários do negócio para descobrir informações especificas, necessárias para o projeto dos modelos apropriados de dados dimensionais.

As entrevistas com os usuários finais e a condução da auditoria dos dados devem ser consideradas como um único passo, porque os requisitos do negócio e os dados disponíveis interagem entre si tão profundamente que precisam ser entendidos ao mesmo tempo. A essência do trabalho do líder de projeto de DW é escolher os dados corretos para atender corretamente às necessidades do negócio (Kimball e Merz, 2000).

Um modelo de estrutura abrangente que reúne os requisitos de negócio e a realização da auditoria de dados é dividido em quatro fases (Kimball, 1998b). Uma consulta a essas fases, que estão disponíveis no anexo C, é recomendada para uma melhor formação da postura da pessoa que procederá as entrevistas.

Uma forma de analise dos requisitos para o ambiente de DW é através de sessões de *brainstorming* (tempestade de idéias) entre os analistas de negócios envolvidos no projeto, para desenvolver questões OLAP (*OnLine Analytic Processing*). Pode-se visitar sites Web atuais para adquirir experiência e simular alguns cenários de negócio para desenvolver estas questões OLAP. Depois de capturar as questões do negócio e as questões OLAP, é preciso dividir por categorias. A divisão deve ser de acordo com o assunto por área que um projeto de DW atuará (Song, 1999).

Algumas categorias para questões OLAP, citadas a seguir, demonstram áreas em que se pode formular questões que os executivos do negócio e a equipe de desenvolvimento do site Web gostariam de ver respondidas em suas análises.

- Projeto de sites Web e análise da navegação;
- Serviços ao cliente;
- Promoções.

No projeto de sites Web e análise da navegação informações referentes a pico de tráfego em relação ao tempo, existência de algum padrão de navegação por usuário, páginas mais ou menos visitadas, a origem do navegante, a freqüência de utilização do site de Web por determinado usuário, as buscas mais populares formam algumas das questões iniciais de consulta ao DW de clickstream. Na área de serviços ao cliente questões como as principais reclamações sobre serviços disponibilizados no site Web ou o perfil do cliente que utiliza bastante o site Web é adequado ao serviço oferecido. Na área de promoções é possível identificar os *banners* de propaganda mais acessados.

3.3.2 Projeto do Data Warehouse

As diferenças entre o projeto de um DW de clickstream e um projeto de DW tradicional aparecem neste estágio. A análise dos dados de sites da Web adiciona um novo nível de complexidade e deverá ser feita através de ferramentas específicas (Sweiger, 2002), que atendem as necessidades do site da Web que será o foco do trabalho.

O repositório das informações constitui o núcleo do ambiente do DW, onde estarão contidos todos os dados extraídos dos sistemas transacionais, formando assim a base necessária para o processo de tomada de decisão. A primeira fase no projeto do repositório é a definição do modelo de dados. A modelagem de dados do DW consiste na metodologia pela qual as informações do DW são representadas e armazenadas (Sell, 2001).

A tradicional abordagem entidade-relacionamento (E/R), utilizada com sucesso no projeto de sistemas para o ambiente operacional, não é adequada para a construção de um DW. No modelo E/R, a estrutura é otimizada para recuperar, criar e atualizar registros individuais em tempo real e preservar a integridade dos dados. No DW, a utilização dos dados é diferente. As consultas realizadas recuperam um grande número de registros e os resumem,

durante o processamento. Neste caso, a estrutura normalmente não é normalizada para evitar junções de muitas tabelas, obtendo-se assim uma performance superior (Kimball, 1998a).

Quadro 5 - Diferenças entre (MD) e (E/R)

USO	MODELO DIMENSIONAL	MODELO ENTIDADE/RELACIONAMENTO
Processamento de transação	Não	Sim
Alvo de limpeza de dados	Não	Sim
Pré-consolidação de dados finais	Sim	Não
Consulta *ad-hot*	Sim	Não
Relatórios	Sim	Não
Exploração de dados	Sim	Não
Previsão	Sim	Não
Investigar através de dados distribuídos	Sim	Não
Armazenamento ciente agregado de dados	Sim	Não
Estrutura previsível, padrão	Sim	Não
Arquitetura e administração distribuída	Sim	Não
Ferramenta de suporte explícita para usuário final	Sim	Não
Compatibilidade com OLAP	Sim	Não
Dimensões lentamente em alteração	Sim	Não

Fonte: Adaptado de Kimball e Merz, 2000.

Como um DW é um banco de dados orientado somente para consulta de seus dados, a orientação da técnica criou os denominados modelos estrela. Está formação é de fundamental importância para que um projeto de DW resulte em um armazém de dados organizado e acessível, com as informações necessárias à gestão de negócios, e não, ao controle do negócio (Machado, 2000).

A modelagem dimensional (MD) busca modelar dados para aprimorar o entendimento

e o desempenho. Tanto os modelos dimensionais como os E/R têm seu lugar adequado no DW (Kimball e Merz, 2000). O quadro 5 resume as diferenças entre as duas técnicas.

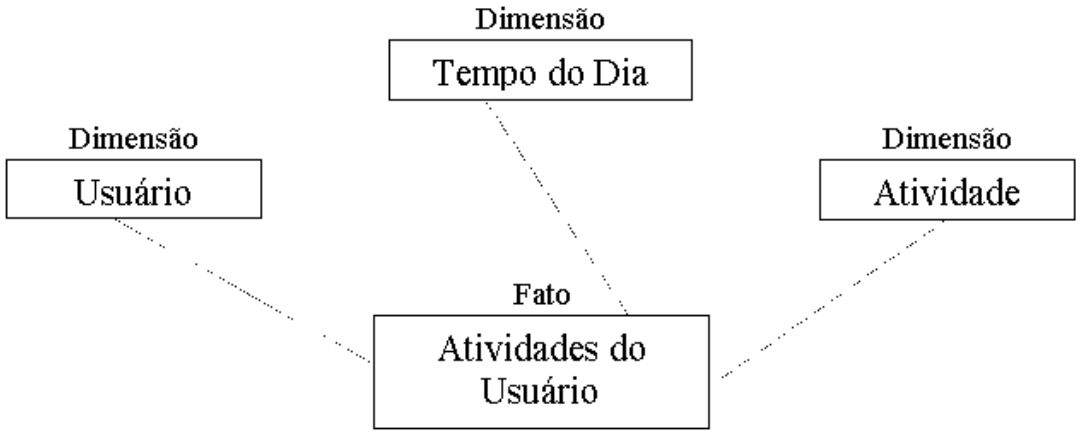
3.3.2.1 Modelagem Dimensional

Todos os modelos dimensionais são construídos em torno do conceito de fatos medidos. Todos esses fatos são numéricos, sendo que, alguns deles são verdadeiros em um instante específico de tempo, e outros representam uma medida acumulada sobre um período de tempo (Kimball e Merz, 2000).

Um fato é um dado elementar básico que é usado no processo de decisão. Fatos são o que o usuário precisa saber para tomada de decisão. Todos os fatos relatados em um simples evento são armazenados em uma tabela de fatos, como uma fila (Sweiger, 2002).

O objetivo na modelagem dimensional é cercar os fatos medidos com o maior número possível de dados contextuais. Cada um dos itens na lista de contexto é chamado de dimensão. As dimensões são, cada uma delas, descrições textuais ricas de algo que existe no momento em que o registro da tabela de fatos é definido (Kimball e Merz, 2000).

Figura 4 - Um Exemplo de Modelo Dimensional



Como exemplo, ilustrado na figura 4, tomam-se as dimensões de análise de dados de um clickstream: Usuário, Tempo do Dia, Atividade, e pode-se localizar um determinado fato,

como as atividades de um determinado usuário no site.

Os atributos da dimensão usuário podem ser nome do usuário, seu e-mail, seu identificador para o site da Web, entre outros. Os atributos do tempo são a hora do dia, o período e os atributos da atividade podem ser um evento identificado, o tipo de atividade, a qual grupo pertence, entre outros.

Além da junção das tabelas dimensionais o fato Atividades do usuário pode conter os atributos tempo de inicio e fim de uma atividade, o número de páginas visitadas, entre outros.

O modelo dimensional produz um projeto de banco de dados consistente com o modo como o usuário entra e navega no DW. O modelo dimensional é referido com freqüência como protocolo em estrela, devido à aparência do projeto lógico do banco de dados (Harrison, 1998).

Além de agilizar o processamento das consultas, o modelo dimensional permite uma melhor visualização dos dados, devido à forma simples de organiza-los. Esta forma ainda propicia a flexibilidade necessária para eventuais ajustes que se façam necessários no modelo (Kimball, 1998b), como:

- Adicionar novos fatos à tabela de fatos, desde que correspondam ao mesmo nível de detalhes;
- Adicionar novas dimensões;
- Adicionar novos atributos às dimensões existentes;
- Redefinir o nível de detalhes de dados.

Na segunda fase da modelagem é constituído o modelo físico, onde são incluídas características físicas e as chaves.

As chaves, que serão responsáveis pelas junções das tabelas lógicas, podem ser definidas como: composta, estrangeira, generalizada, primária e substituta (*surrogate*). Segundo Kimball (1998b), a definição para cada uma dessas chaves é a seguinte.

- ✓ Chave composta: Uma chave em uma tabela do banco de dados, composta de vários campos. Por exemplo, a chave geral em uma típica tabela de fatos, é um subconjunto de chaves estrangeiras na tabela de fatos.

- ✓ Chave estrangeira: Um campo em uma tabela de banco de dados relacional cujos valores são aproveitados a partir de valores de uma chave primária em outra tabela.

- ✓ Chave generalizada: Uma chave primária da tabela de dimensões que foi criada ao generalizar uma chave original de produção como um número de produto ou um número de cliente, normalmente ao criar uma chave substituta.

- ✓ Chave primária: Um campo, em uma tabela de banco de dados, que é único e diferente para cada registro na tabela.

- ✓ Chave substituta: Uma chave artificial, normalmente um inteiro consecutivamente atribuído, que é utilizado em um modelo dimensional para concatenar uma tabela de dimensões a uma tabela de fatos. Na tabela de dimensões, a chave substituta é a chave primária. Na tabela de fatos, a chave substituta é uma chave estrangeira e pode ser parte da chave primária da tabela de fatos.

O ponto a favor de uma chave substituta é que não há absolutamente nenhuma semântica no valor da chave. É somente um veículo para unir a tabela de dimensão a uma tabela de fatos (Kimball e Merz, 2000). Se um objeto sendo referido é desconhecido, não medido, inaplicável, corrompido ou ainda não acontecido, uma chave semântica significativa de produção não pode ser aplicada. Contudo, o DW tem que fornecer uma chave. As chaves substitutas isolam o DW de alterações na administração de chaves nos dados de produção, além do que podem tratar com facilidade dimensões que se alteram lentamente.

Dimensões que se alteram lentamente são dimensões que tem seus atributos em alteração ao longo do tempo. São divididas em três versões: Quando existe uma nova descrição e a antiga é sobrescrita, sem adicionar um novo registro ou alterar a chave; Quando uma alteração física verdadeira ocorreu em um ponto específico no tempo e se cria um novo registro de dimensão, com uma nova chave de DW; E quando uma descrição alternativa,

simultânea de alguma coisa está disponível, neste caso um campo extra de "valor antigo" é adicionada a dimensão afetada (Kimball e Merz, 2000).

Tendo definido os atributos que figurarão nas tabelas de fatos e dimensões com suas características físicas, ainda resta decidir sobre os aspectos de organização dos dados e performance, representados inicialmente pela definição da granularidade, e da definição de agregados.

3.3.2.2 Granularidade

Kimball e Merz (2000, p.124) citam que: "O grão é a definição formal do que é um único registro de tabela de fatos. Declarar o grão é extremamente importante para modeladores dimensionais. Todos os projetos dimensionais devem se iniciar pela declaração do grão, e então, partir para decidir quais fatos e dimensões se ajustam a esse grão. Uma vez que o grão é declarado, todos os fatos devem se ajustar a esse grão".

Existem diferentes níveis de detalhes no DW e a sua definição afeta diretamente o volume de dados e a qualidade das consultas que poderão ser feitas. Uma granularidade de alto nível garante maior rapidez nas consultas feitas pelos usuários, mas, em contrapartida, há uma diminuição da riqueza das informações. Por outro lado, uma granularidade de baixo nível possibilita a obtenção de respostas ricas em detalhes, mas haverá um aumento do volume de dados, o que conseqüentemente fará com que o tempo de resposta seja maior e que o investimento em hardware seja maior (Inmon, 1997).

Por exemplo, em um DW de clickstream, o grão da tabela de fatos poderia ser de um hit (consulta) de página, o qual significaria que o conteúdo dimensional necessário para ir abaixo do componente da página individual não existiria porque cada ponto visualizado de um hit de página contém um componente de página (Sweiger, 2002).

É necessário encontrar um ponto de equilíbrio. O nível adequado de granularidade deve ser definido de tal forma que atenda as necessidades do usuário, tendo como limitação os recursos disponíveis.

3.3.2.3 Agregados

Criar agregados está em resumir e armazenar dados que estão disponíveis na tabela de fatos com o objetivo de melhorar a performance de respostas às perguntas dos usuários do sistema. Existem três aproximações para agregações: nenhuma agregação, agregação seletiva, ou agregação exaustiva. Em alguns casos, o volume de dados na tabela de fatos é pequeno e a performance é aceitável sem agregados. Normalmente o volume de dados é bastante grande e é necessária a implementação de agregados (Mark, 2002).

Existem duas abordagens para armazenar agregados: definindo novas tabelas de fatos ou definindo campos níveis. A utilização da primeira abordagem torna mais simples a manutenção, a carga e a utilização dos dados (Kimball, 1998a).

Por exemplo, suponha que um grande DW de vendas no varejo tenha 400 milhões de entradas em sua tabela de fatos de transações de venda. Além disso, suponha uma consulta comum de um usuário que relacione vendas, distrito e categoria de produtos. Essa consulta pode ser satisfeita pelo processamento de todas as 400 milhões de entradas cada vez que é executada. O administrador poderia criar um nível dimensional agregado envolvendo distrito e categoria de produtos e então essa consulta de usuário poderia ser executada através de uma tabela bem mais enxuta (Sweiger, 2002).

3.3.3 Implementando o Pós-processador de Clickstream

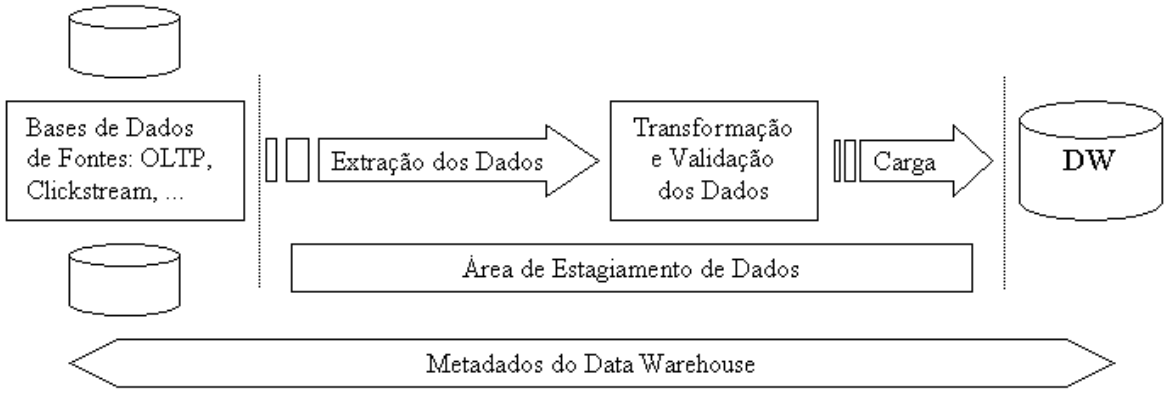
"Um sistema de suporte à decisão deve obter dados de um ou mais sistemas fontes. Essa tarefa é realizada para implementação dos processos de extração, transformação e carga dos dados" (CRAIG, 1999), e também é conhecido como processo ETL.

A figura 5 ilustra de forma geral os processos ETL. Após a extração dos dados, sobre eles são aplicadas transformações. Quando os dados estiverem no formato apropriado, o processo ETL carrega os dados para as tabelas de destino no data warehouse para que possa ser acessado por ferramentas de análise.

Uma coisa que esses componentes tem em comum é que são construídos em uma área

comum de estagiamento de dados. O esquema de estagiamento deve ser especificado para suportar todos os passos do processamento ETL, da extração à carga ao DW (Sweiger, 2002). O estagiamento de dados é o ambiente onde dados vindos de diversas fontes com diversos formatos deverão ser padronizados, podendo ser considerado um trabalho minucioso e exaustivo.

Figura 5 - Processos ETL



Fonte: Adaptado de Sweiger, 2002.

A seguir serão abordadas separadamente as etapas de extração, transformação e carga de dados.

3.3.3.1 Processo de Extração dos Dados

A tarefa de extração de dados está relacionada com as seguintes etapas do ciclo de vida de DW: definição do escopo do projeto de DW, análise dos sistemas fontes e a especificação dos programas de extração (Pereira, 2000).

- No escopo do projeto são identificados os grupos de usuários que irão interagir direta ou indiretamente com o sistema e definidos os requisitos de informação para os processos de negócios a serem suportados por uma abordagem DW.

- A análise dos sistemas fontes tem por objetivo compreender os dados distribuídos pela organização e integrá-los de forma a refletir a perspectiva histórica de interesse às análises do ambiente de decisão.

- Os programas de extração devem dar suporte a captura incremental dos dados que equivale a uma replicação baseada em dados modificados para posterior distribuição ao DW.

3.3.3.2 Processo de Transformação dos Dados

Uma vez que os dados tenham sido extraídos dos sistemas fontes, um conjunto de transformações deve ser processado sobre esses dados. A transformação dos dados pode ser simples ou complexa, dependendo da natureza dos sistemas fontes. Em algumas situações, múltiplos estágios de transformações são necessários (Pereira, 2000).

As rotinas de limpeza e integração atuam sobre os dados extraídos. A execução dessas rotinas de limpeza sobre os dados coletados permite assegurar sua consistência. Dados migrados para o DW, sem estarem integrados, não podem ser empregados no suporte a uma visão corporativa dos dados (INMON, 1997a).

A inexistência de um código único atribuído ao mesmo elemento de dados; o emprego de vários formatos de campo para um mesmo dado; ou o atributo identificador de dados que pode diferir entre os sistemas transacionais, mostra três exemplos de situações onde a transformação deverá operar para integrar os dados.

3.3.3.3 Processo de Carga no DW

O processo de carga é freqüentemente um ponto complexo. Antes que os dados possam ser carregados para o data warehouse, novidades ou mudanças nos dados dimensionais devem ser processados. Uma vez que todos os dados para as dimensões foram estagiados, o dados devem ser destinados para as tabelas de fatos para serem validados e todas as chaves geradas e checadas através de tabelas dimensionais (Sweiger, 2002).

Uma vez que se tenha adequado as mudanças no DW, no processo de ETL para suportar clickstream, preservando a integridade referencial e a qualidade dos dados, é preciso assegurar que diferentes usuários possam acessar os dados com rapidez e facilidade. Como os usuários navegam através de uma montanha de dados de clickstream, eles esperam ferramentas analíticas poderosas para habilitar análises *ad hoc* e também velocidade e consistência nas respostas.

3.3.4 Analisando os Dados em um DW de Clickstream

É importante definir as principais categorias de exploração de dados e mostrar quais as transformações que precisam ser feitas nos dados do warehouse para deixa-los prontos para a exploração de dados. A seguir é apresentado um conjunto de características para análise dos dados (Kimball e Merz, 2000).

- A exploração de dados é um conjunto de técnicas de análise poderosas para dar sentido a conjuntos de dados muito grandes;

- Não há uma abordagem de exploração de dados, mas antes, um conjunto de técnicas que podem ser freqüentemente utilizadas em combinação umas com as outras para extrair mais *insights* ou idéias dos dados explorados;

- Cada ferramenta de exploração de dados pode ser visualizada logicamente como um aplicativo que é um cliente do data warehouse;

- O objetivo do warehouse é fornecer "conjuntos de observação prontos para utilização" para a exploração de dados.

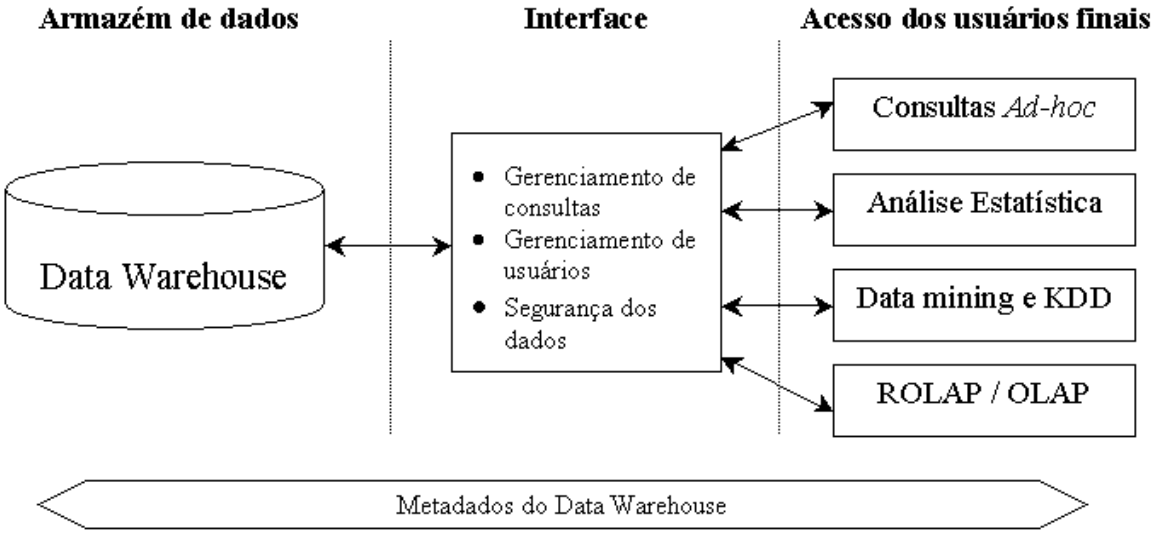
O propósito do warehouse é coletar, armazenar e apresentar dados da melhor maneira possível a ferramentas de exploração de dados. Não é seu propósito realizar realmente a exploração de dados. Essa função cabe mais para um aplicativo analítico do que para um banco de dados.

Na figura 6 apresenta-se a visão da fase final na construção do DW, mostrando a interface entre o DW e os aplicativos de consulta dos usuários finais. Essa interface é

responsável por gerenciar consultas, usuários e a segurança de acesso do DW.

A consulta *ad-hoc* é um tipo especifico de ferramenta de acesso a dados do DW por usuários finais que solicita ao usuário para formar suas próprias questões por manipulação direta de tabelas relacionais e suas junções (Kimball, 1998b).

Figura 6 - Visão da Fase Final do DW de Clickstream



Fonte: Adaptado de Sweiger, 2002.

A análise estatística é projetada para reduzir uma grande quantidade de dados a uma simples relação ou fórmula, com cálculos de média.

Segundo Harrison (1998), data mining é a exploração e análise, por meios analíticos ou semi-analíticos de grandes quantidades de dados para descobrir modelos e regras significativas. É particularmente útil para problemas de modelagem não-linear com grande número de variáveis.

KDD ou *Knowledge Discovery Database* é uma técnica que completa a análise do data mining verificando o que realmente é conhecimento e o que é agregado de informações, ou seja, extração de conhecimento.

Ferramentas OLAP provem análises multidimensionais, com operações dinâmicas com as seguintes características que são necessárias para capacitar à análise de clickstream: *slice-and-dice* (operações para realizar navegação por meio dos dados na visualização de um cubo), *drill-down* (aumento do nível de detalhamento) e *drill-up* (saindo do nível de detalhamento).

Existem três arquiteturas básicas OLAP: A Multidimensional pura (MOLAP), a Relacional pura (ROLAP) e uma híbrida (HOLAP) que combina as características de MOLAP e ROLAP. Cada arquitetura tem seu conjunto de benefícios e desvantagens. O quadro 6 trás um mapa de comparação das características dessas arquiteturas. Além destas três arquiteturas básicas OLAP, existem outras duas extensões que apareceram posteriormente: A DOLAP e a WOLAP.

Quadro 6 - Mapa de Comparação das Características da Arquitetura OLAP

CARACTERÍSTICAS	TIPO DE SOFTWARE OLAP		
	MOLAP	**HOLAP**	**ROLAP**
Tamanho potencial do DW de clickstream	Pequeno (Cubo de dados somente)	Grande (Cubos de dados e base de dados relacionais)	Grande (Base de dados relacionais)
Performance de consultas iniciais	O mais rápido	Moderado	Sem prognóstico
Tempo de resposta em consultas Drill	Consistente	Relativamente consistente	Inconsistente (depende da consulta)
Capacidade analítica avançada (análise de séries temporais, ranking, exceções estatísticas, etc)	Alto	Alto	Baixo (limitado pelo SQL)
Eficiência no tempo de carga	Muito baixo (Simples carga de dados do cubo)	Médio (Impacto pelos dados do cubo)	Muito alto (Usa paralelismo SQL)

Fonte: Adaptado de Sweiger, 2002.

A DOLAP é uma OLAP que se baseia numa arquitetura DESKTOP, ou seja, é uma ferramenta para usuários que possuam uma cópia de sua base multidimensional ou de um subconjunto dela ou que queiram acessar um repositório de dados central. Basicamente acessa

os cubos já existentes do banco de dados ou um conjunto de cubos selecionados pelo usuário.

A WOLAP ou WebOLAP, ferramenta OLAP que permite algum nível de acesso via *browser* Web. As facilidades dessa arquitetura são: a possibilidade de plataformas independentes para dar suporte a usuários distantes e aplicações de *groupware*.

Usando a modelagem de dados, mecanismos de busca em banco de dados, técnicas analíticas em conjunto com ferramentas OLAP que tenham uma grande quantidade de características, a análise de um data warehouse de clickstream pode melhorar enormemente na performance das consultas de usuários (Sweiger, 2002).

3.3.5 Metadados

Em termos simples, metadado é definido como sendo "dado sobre o dado". Ou seja, o metadado descreve ou qualifica outro dado, incorporando, a este, significado. Sem metadado, a informação se restringe a um conjunto de dados sem significado.

Através de uma solução eficaz de metadados é possível avaliar o impacto das mudanças nos sistemas transacionais e, portanto, a tarefa de manutenção desses sistemas torna-se menos complexa. Da mesma forma, os metadados auxiliam o processo de construção e manutenção do DW (Pereira, 2000).

Tipicamente, a construção de um DW envolve a tarefa de extração de dados. A integração de todos esses dados exige conhecimento dos seus significados, estruturas, locais de armazenamento e sistemas que os mantêm atualizados, todo esse conhecimento deve fazer parte dos metadados.

A inexistência de metadados integrados capazes de descrever completamente os dados dificulta ainda mais a integração e o compartilhamento dos dados nas organizações (Brackett, 1996).

Os metadados assumem papel importante no processo de transformação dos dados, pois através deles, serão armazenadas as lógicas das transformações e os mapeamentos entre os dados dos sistemas transacionais e aqueles mantidos no DW (Pereira, 2000).

Existem basicamente duas classificações de metadados: metadados técnicos e de negócios. Em (Moncla, 1999), é apresentada uma possível classificação acerca do emprego dos metadados técnicos e de negócios nos processos típicos de um ambiente de DW: definição, transformação e derivação, gerenciamento e administração. O quadro 7 exibe exemplos de metadados seguindo essa classificação.

- Os metadados para definição são responsáveis por esclarecer a semântica, o contexto e descrever formalmente os elementos de dados.

- Os metadados para transformação e derivação são empregados para descrever a lógica das transformações aplicadas sobre os dados e definir sua procedência.

- Os metadados do contexto de gerenciamento e administração permitem otimizar o acesso aos dados e melhorar o processo de aquisição de informação.

Quadro 7 - Classificação de Metadados

METADADOS	DEFINIÇÃO	TRANSFORMAÇÃO/ DERIVAÇÃO	GERENCIAMENTO/ ADMINISTRAÇÃO
Negócio (Geralmente não estruturado)	O que significa? Onde posso encontra-lo?	Como foi calculado? Quais foram às fontes? Que regras de negócio foram aplicadas?	Que treinamento está disponível? Quem está à frente da equipe? Qual a forma mais fácil de obter a informação? Quão recente é a informação?
Técnico (Geralmente estruturado)	Formato Tamanho Domínio Banco de Dados Catálogo	Filtros Agregados Cálculos Expressões	Capacidade de planejamento Alocação de espaço Indexação e reindexação Utilização do disco Escalonamento de tarefas

Fonte: Adaptado de Moncla, 1999.

O motivo central de um sistema de suporte à decisão é apresentar informações de cunho prático aos analistas. Para tal, os metadados devem traduzir a terminologia técnica para os termos dos negócios. O suporte proporcionado deve ser enfocado sob o ponto de vista do usuário final e não apenas sob a ótica dos desenvolvedores.

3.4 Segurança dos Dados no DW

Uma das ironias do trabalho da gerência de DW é a tensão entre o dever de publicação e o dever de proteção. A gerência de DW tem sido confiada com os dados da organização, mas pode ser responsabilizada se eles forem perdidos ou roubados (Kimball e Merz, 2000).

Ninguém em uma organização é mais bem equipado para qualificar usuários e conceder direitos de acesso ao DW do que a própria gerência do data warehouse, logo ela deve gerenciar ativamente a segurança do DW, dominando o assunto para que possa orientar especialistas em segurança (Kimball, 1998b).

Existem muitas vulnerabilidades e uma pessoa não pode pensar nelas todas simultaneamente. A visualização dos problemas é difusa e complexa e está é a dificuldade de se saber por onde começar, além do que segurança é um tópico fundamentalmente negativo, e isto não tem um lado bom porque as pessoas se afastam diretamente do assunto (Segurança Máxima, 2000).

Algumas ferramentas de ambiente de rede como: roteadores, *firewalls*, servidores de diretório e criptografia de dados podem auxiliar na estruturação da segurança para o pleno funcionamento do DW. A Figura 7 ilustra uma possível arquitetura de rede, levando-se em consideração fatores de segurança.

Segundo Kimball (1998b), alguns elementos de compromisso devem ser seguidos, para que se aumente o nível de segurança dos dados:

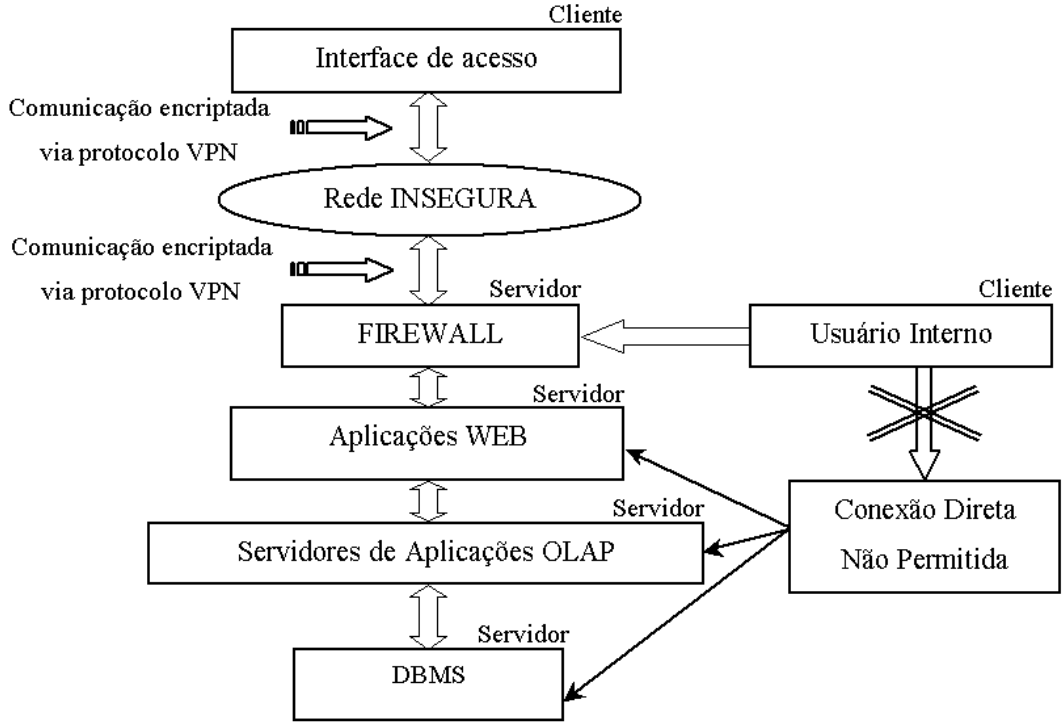
- Consciência;

- Suporte a executivos;

- Policiamento;
- Vigilância;
- Desconfiança;
- Renovação.

O trabalho dos gerentes de DW é profissionalmente publicar e proteger as informações da organização. As primeiras medidas são tarefas táticas que deveriam ser feitas imediatamente para sanar problemas pré-existentes, já um segundo grupo de medidas é estratégico e deve ser planejado e atualizado ao longo do tempo.

Figura 7 - Uma Típica Arquitetura de um DW Moderno



Fonte: Adaptado de Kimball, 1998b

Segundo Kimball e Merz (2000), a estrutura de segurança do Data Warehouse para

clickstream deve-se basear em quatro elementos:

- Dois fatores para autenticação;
- Uma conexão segura;
- Definição forte dos papéis do usuário;
- Acesso a todos os objetos do Warehouse controlados por papéis.

Um fator de segurança pode ser representado por uma senha de acesso. Infelizmente se outra pessoa conhece essa senha, ele torna-se você, com todos os direitos e privilégios. Com dois fatores de segurança pode-se agregar ao primeiro fator um segundo que aperfeiçoe a autenticação, como um cartão magnético ou um equipamento biométrico.

A conexão segura deve garantir que nenhum aplicativo consiga monitorar os dados que trafegam pela rede (seja pelas LANs ou WANs). Para conseguir esse fato pode-se utilizar técnicas de VPN (Rede Privada Virtual) para redes WANs e comunicações locais criptografadas nas redes LANs. O modelo proposto na figura 7 ilustra esses pontos.

A definição dos papéis de cada usuário fornece os direitos para acessar recursos de informações no DW e deve ser organizada por grupos de interesse.

3.5 Arquitetura do Data Warehouse

A escolha da arquitetura é uma decisão gerencial do projeto, e está normalmente baseada nos fatores relativos à infra-estrutura disponível, ao ambiente de negócios, juntamente com o escopo de abrangência desejado.

A abordagem de implementação escolhida é uma decisão que pode causar impactos quanto ao sucesso de um DW. Muitas variáveis afetam a escolha da implementação e arquitetura, entre elas o tempo para a execução do projeto, o retorno de investimento a ser realizado, a velocidade dos benefícios da utilização das informações, a satisfação dos usuários executivos e os recursos necessários à implementação de uma arquitetura (Machado, 2000).

Antes de apresentar as arquiteturas de DW, é preciso definir um componente importante nesse contexto, o Data Mart. Um Data Mart é um conjunto de tabelas de fatos juntamente a um conjunto de tabelas de dimensão conectadas que atende às necessidades de um grupo particular de negócio (Kimball e Merz, 2000).

Na literatura encontramos vários conceitos de Data Mart – DM. Segundo Berry, 1997 "Data Mart é um sistema especializado que fornece juntos os dados necessários para departamento ou uma aplicação relacionada". Os DM são subconjuntos de dados da empresa armazenados fisicamente em mais de um local, geralmente divididos por departamentos (Inmon, 1997).

Muitos modelos dimensionais de aparência semelhante podem ser construídos em torno da empresa, representando fontes diversas de dados e os interesses de grupos diferentes que desejam analisar tais dados. Cada um desses modelos dimensionais pode ser chamado de um DM (Kimball e Merz, 2000).

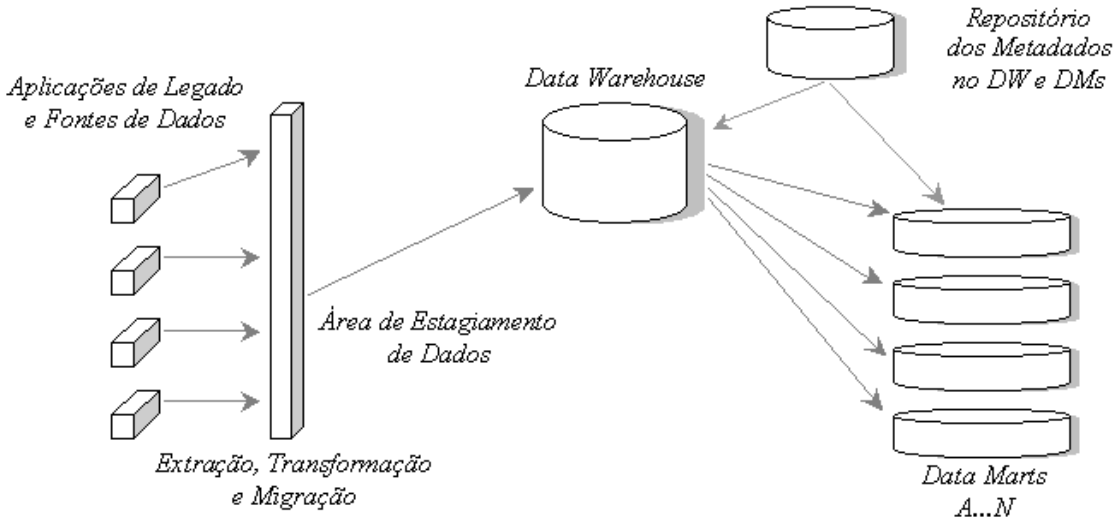
A escolha de uma arquitetura determinará ou será determinada por onde o DW ou DM estarão residindo. As arquiteturas conhecidas na literatura são três: Global, Independente e Integrada.

3.5.1 Arquitetura Global

A Arquitetura Global é considerada como a que suporta toda ou a maior parte dos requerimentos ou necessidades de um DW integrado com grande grau de acesso e utilização das informações para todos os departamentos de uma empresa. Para implementar essa arquitetura é utilizada a técnica *Top-Down*, um exemplo é apresentado na figura 8.

Proposta por Inmon (1997), ela requer um maior planejamento e trabalho de definições conceituais de tecnologia completos antes de se iniciar o projeto do DW propriamente dito.

Figura 8 - Implementação *Top-Down*



Fonte: Firestone, 1998.

O quadro 8 apresenta uma síntese das vantagens e desvantagens da utilização da implementação *Top-Down*.

Quadro 8 - Comparativo da Implementação *Top-Down*

VANTAGEM	DESVANTAGEM
Herança de arquitetura, permitindo uma fácil manutenção.	Implementação muito longa podendo levar facilmente mais de um ano.
Visão global do empreendimento.	Alta taxa de risco.
Controle e centralização de regras.	Herança de cruzamentos funcionais.
Repositório de metadados centralizado e simples.	Expectativas relacionadas ao ambiente, devido à demora do projeto.

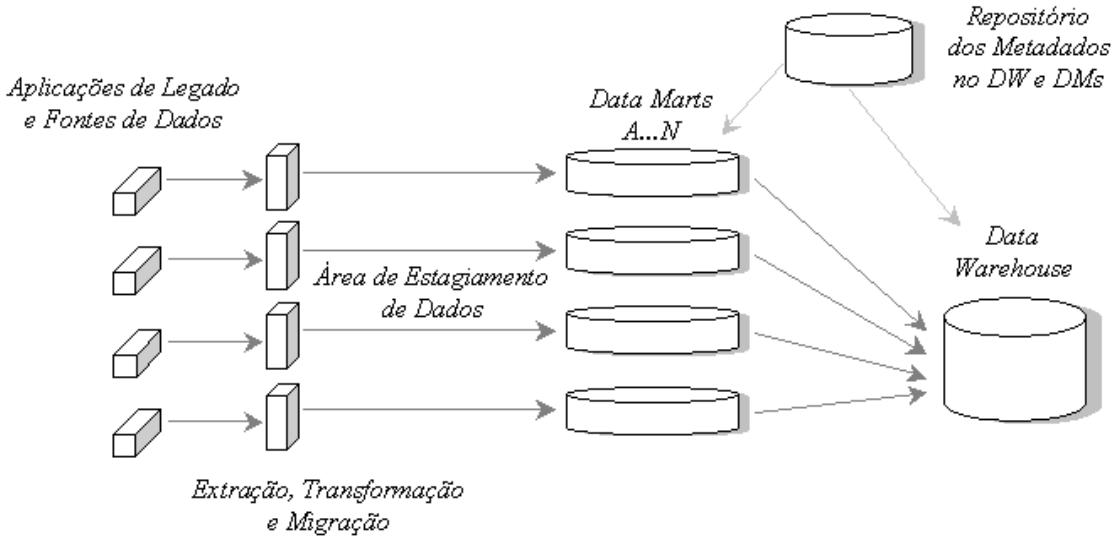
Fonte: Machado, 2000.

Os dados são extraídos de sistemas operacionais e fontes de dados externas em horário fora do pico das operações. São filtrados, eliminando-se dados não necessários e tratados para a qualidade dos requisitos levantados para o projeto. Por fim são carregados nas bases de dados apropriadas de data warehouse para acesso aos usuários finais (Machado, 2000).

3.5.2 Arquitetura de Data Marts Independentes

A arquitetura independente implica em Data Marts (DM) *stand-alone*, controlado por um grupo específico de usuários e que atende somente às suas necessidades específicas e departamentais, sem foco corporativo. Esse fato faz com que não exista conectividade entre Data Marts da organização. Está arquitetura é apresentada na figura 9.

Figura 9 - Implementação *Botton-Up*



Fonte: Firestone, 1998.

Quadro 9 - Comparativo da Implementação *Botton-Up*

VANTAGEM	DESVANTAGEM
Implementação rápida.	Perigo de *legamarts*: DM isolados.
Retorno rápido.	Desafio de possuir a visão do empreendimento.
Manutenção no enfoque da equipe.	Administrar e coordenar múltiplas equipes e iniciativas.
Herança incremental	Maldição do sucesso: enquanto um grupo fica feliz com os resultados, outros grupos pressionam por resultados.

Fonte: Machado, 2000.

Ao contrário da abordagem anterior, esta parte da implementação é chamada de *Botton-Up*. Dessa forma serão implementados data marts individualmente, sem partir da análise de requisitos para toda a empresa (Sell, 2000). O quadro 9 faz um comparativo das vantagens e desvantagens dessa implementação.

3.5.3 Arquitetura Integrada ou BUS

Essa forma de implementação tem o propósito de integrar a implementação *Top-Down* com a *Botton-up.* Essa abordagem é conhecida como desenvolvimento baseado em data marts incrementais e foi proposta por Kimball (1998b).

Na construção de um DW, há uma gama de fatores que afetam a complexidade como, por exemplo, a construção do projeto que é lenta e cara. Com o objetivo de equilibrar os gastos e oferecer resultados em prazos mais curtos, é possível construir data marts incrementais.

Um grande DW de clickstream distribuído por diversas empresas ou negócios tem de ter algum tipo de uniformidade previsível. Deve haver um conjunto de padrões que permita que as partes diferentes se reconheçam e se comuniquem. Além disso, novas partes do DW devem ser capazes de se unirem ao DW existente e dele participar de maneira eficiente.

A solução proposta por Kimball e Merz (2000), é utilizar-se de dimensões e fatos adaptados que estão distribuídos por todas as partes do DW. Uma dimensão adaptada é uma dimensão que significa a mesma coisa e que tem a mesma estrutura através de diversos data marts. Já um fato adaptado é uma definição de um fato medido que é consistente através de diversos data marts. Geralmente, uma instância específica de um fato existe em somente uma localização do data warehouse. É a definição do fato que é consistente através dos data marts.

Conforme citam Kimball e Merz (2000, p.126): “Dimensões adaptadas e fatos adaptados implementam uma arquitetura Data Warehouse Bus e a arquitetura Data Warehouse Bus é a base para um sistema completamente distribuído de data warehouse, tal como o DW para clickstream ”.

São somente os dados de dimensão que precisam e devem ser duplicados. Independentemente de qual seja a estrutura arquitetônica, uma estrutura comum como uma dimensão será sempre fisicamente duplicada, em um grande ambiente de data warehouse.

Os dados da tabela de fatos, que podem representar mais de 90 por cento do volume de um DW, são explicitamente não duplicados. Uma característica forte do DW Bus é o isolamento dos dados da tabela de fatos em localizações específicas. Um provedor de fato local é um grupo que controla uma fonte específica de dados. Os conjuntos destes provedores vivem todos em torno do DW. Todos os clientes que precisam dos dados abrem conexão diretamente com a cópia física mantida pelo grupo provedor de fatos locais.

3.6 Conclusão

Neste capítulo, apresentou-se uma revisão teórica sobre o data warehouse para clickstream, que pode ser utilizado para responder questões relacionadas ao ambiente de sites da Web na Internet. Foram mostrados os principais conceitos e componentes, com o propósito de fundamentar a elaboração e implementação do modelo proposto.

De acordo com o processo de construção de um data warehouse de clickstream apresentado anteriormente pode-se concluir:

- Planejamento do data warehouse: Mostrou-se que no planejamento existe uma infinidade de papéis necessários para a realização de um projeto completo de data warehouse. Se os passos de definição do projeto, identificação dos papéis, o planejamento das atividades, os requisitos do negócio e dados da auditoria forem bem gerenciados, e a equipe consegue entender a real necessidade da organização, o projeto tem uma boa chance de sucesso. Levando-se sempre em mente que os dados devem ser publicados corretamente.

- Modelagem: De acordo com a definição dos requisitos se esboça e se constrói o modelo de dados. É feita uma comparação entre MD e E/R, levando-se em consideração de que para esta aplicação a MD é a mais indicada. É apresentada a metodologia de implantação do warehouse a partir de data marts e os requisitos

desse passos são abordados.

- No processo de extrair/transformar/carregar que é chamado de pós-processamento de clickstream, são abordados os recursos específicos que um pós-processador deve conter.

- O warehouse entrega os dados na forma de conjuntos de observações prontos para utilização, que podem ser digeridos imediatamente pelas ferramentas de exploração de dados. Um conjunto dessas ferramentas é apresentado e se discuto onde e como utiliza-las.

- Observa-se com atenção especial o papel dos metadados na construção do data warehouse, demonstrando que eles devem passar por todas as etapas do projeto a fim de minimizar a possibilidade de insucesso.

- A utilização de informações confidenciais na ambiente da Internet exige um estudo a respeito da segurança dos dados. São apresentadas as principais recomendações com diretivas de segurança na Internet.

- Por fim, se discute as principais arquiteturas de implementação de data warehouse como a global, independente e a BUS, mostrando as vantagens e desvantagens de cada uma delas.

Os conceitos abordados neste capítulo serão referenciados nos próximos. No capítulo seguinte apresentam-se as metodologias para construção do data warehouse de clickstream e a arquitetura proposta.

CAPÍTULO 4 – O MODELO PROPOSTO

4.1 Introdução

O objetivo desse capítulo é apresentar um modelo para a implantação de um DM que possa contribuir na melhora do desempenho de um site da Web. O foco do presente trabalho, para efeito de aplicação, trata dos aspectos de navegação na Web e busca desvendar os problemas que podem ocorrer em um serviço oferecido 7 dias por semana, 24 horas por dia (7x24), na rede Internet.

O capítulo inicia com a definição das tecnologias utilizadas no modelo, em seguida são apresentados os requisitos necessários para implantação do data mart de clickstream. As etapas de extração, armazenamento e apresentação são divididas e desenvolvidas objetivando colocar o DM em operação. Por último é proposto um modelo básico de segurança em redes para que o DM seja utilizado no ambiente da Internet.

4.2 Definição

É preciso entender por que a organização precisa dessa solução e se os clientes (os usuários do sistema) que serão satisfeitos com a solução estão realmente interessados e envolvidos com o desenvolvimento do projeto, pois eles darão a direção inicial do caminho que será percorrido no desenvolvimento do sistema.

O projeto de um data warehouse de clickstream deverá criar um esquema que otimize o processo de suporte a decisão. O esquema de dados deve ser simples de entender para um analista de negócios e os dados devem ser limpos, consistentes e corretos. O esquema de dados deve também suportar processamento de consultas rápidas. Conforme pesquisas apresentadas no capítulo anterior, o modelo dimensional satisfaz os requisitos acima e será o utilizado na modelagem.

Ao invés de se projetar um armazém de dados centralizado capaz de atender todas as

áreas da organização, será adotada uma abordagem departamental. Na implantação de um DW em um ambiente de produção contínua de dados e onde os sites da Web têm natureza distribuída, verifica-se que o modelo de data mart incremental é o mais indicado, pois o tempo de execução é menor podendo demonstrar resultados em poucos meses.

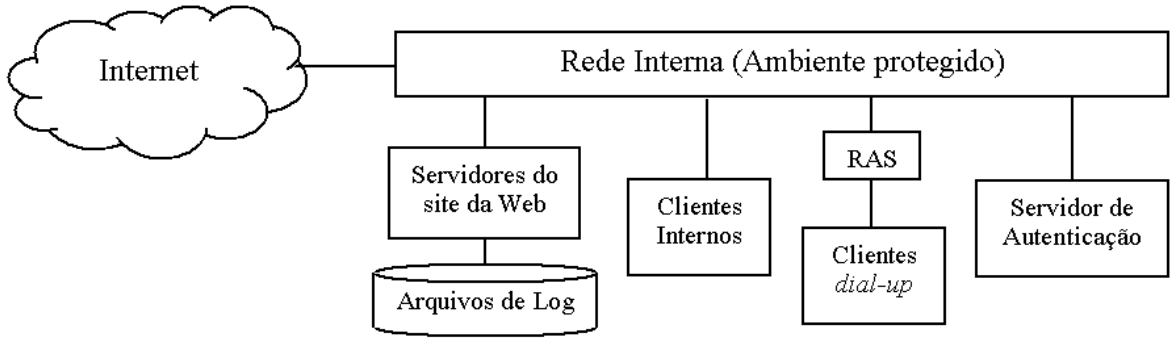
Para que a utilização do modelo da data mart incremental seja bem sucedida, segundo Kimball e Merz (2000), ele deverá utilizar a arquitetura DW Bus. Esta arquitetura demonstra um compromisso de consistência dos dados entre os diferentes DM que venham a ser implementados, adaptando dimensões e fatos a realidade da organização, podendo assim compor um futuro data warehouse de clickstream.

A utilização de questões OLAP, oriundas de *brainstorming* e de experiências em sites da Web, através de simulações de negócio são úteis para demonstrar algumas das possíveis necessidades que a organização precisa resolver e por serem importantes na formação dos atributos que habitarão as dimensões e fatos do data mart.

4.3 Requisitos

Antes de começar a apresentação dos componentes da arquitetura proposta para a construção do DM de clickstream é preciso analisar quais são as fontes de informações que abastecerão o warehouse e como esse sistema está disponível no ambiente de produção.

Figura 10 - Estrutura Característica de um Site da Web



A figura 10 apresenta uma típica estrutura de publicação de sites da Web em provedores de acesso à Internet de pequeno e médio porte no Brasil. Todos os servidores utilizados geram arquivos de log das transações ocorridas, sendo esses dados a principal fonte para o DM proposto. Os clientes do site da Web podem estar na Internet, ou na rede interna do provedor através de duas opções: ligado diretamente à LAN ou conectados através de um servidor de acesso remoto (RAS) em uma conexão *dial-up*.

Um conjunto de requisitos deve ser atendido para que os dados de logs coletados através dos sites da Web estejam em conformidade com a necessidade do projeto. O objetivo é formar subsídios para o estudo do comportamento dos usuários da Web, e por definição esses ambientes não possuem estados, ou seja, um modelo de aplicação básica da Web não mantém uma continuidade de sessão para o navegante.

Para implantação desse DM incremental, a fonte primária de informações considerada será o conjunto de logs armazenados nos servidores da Web no domínio da organização. É preciso definir quais são os fornecedores dos serviços utilizados (servidores) e o padrão para geração dos dados de log. Quanto mais informações coletadas e ricas em detalhes os servidores conseguirem, mais completo se tornará o DM.

Figura 11 - Requisitos no Site da Web

A figura 11 apresenta os pontos de requisito com os quais o modelo se preocupa em atuar para que as informações armazenadas tenham sentido no warehouse. Alguns caminhos para passar parâmetros entre transações HTTP e auxiliar na avaliação de estado de uma sessão foram apresentados no capítulo anterior. A proposta é utilizar duas das técnicas abordadas:

Strings de consulta e servidores de cookie. As duas técnicas se completam em seu objetivo e todo o site da Web deve ser reformulado para suportá-las formando assim os códigos de conteúdo.

A *string* de consulta deve estar presente nos links que levam para uma navegação fora do domínio da organização, como: busca, formulários, *banners*, entre outros. Fornecendo essas informações para o armazenamento nos servidores, quando acionadas.

O servidor de cookie é a técnica que contribuirá para identificação da máquina utilizada na navegação. A cada página do site da Web e link referenciado deve-se consultar a presença ou não de um cookie de sessão que representa o domínio do site da Web. Esse cookie deve ser persistente e único para o domínio, facilitando a identificação na fase de estagiamento dos dados.

No modelo proposto é utilizado um servidor de registro de logs centralizado para atender as plataformas provedoras da Web. Segundo ressalta Sweiger (2002), esse procedimento facilita a manipulação dos dados na fase de estagiamento.

Sugere-se que todas as páginas do site da Web sejam montadas manualmente e trabalhem de forma estática – não se alteram de usuário para usuário. Nessa implementação deve-se incorporar a codificação que é única para cada página identificada. A classificação da atividade e seu conteúdo são informações que ajudarão na compreensão do comportamento do usuário e formarão os rótulos de página.

Para que a seqüência de passos do navegante no site da Web, em relação ao tempo, seja armazenada com precisão é proposta a utilização de uma ferramenta de sincronização dos servidores que utiliza o protocolo NTP (*Network Time Protocol*). Através da implantação de um servidor NTP, que proverá informações de sincronia de tempo aos clientes cadastrados será possível manter um padrão aceitável de registro.

Depois que esses requisitos forem satisfeitos, deverá ser colocado em operação o sistema do site da Web fornecendo os serviços de forma transparente para os usuários navegantes, mas armazenando informações que serão importantes para o povoamento do warehouse.

4.4 Data Warehousing – Visão Geral

O próximo passo é detalhar os componentes da arquitetura DW agrupados por área. Ao enfocar o processo de construção e manutenção de um DW, são identificadas as seguintes áreas: captura, integração, limpeza, transformação, carga de dados e gerencia do DW. A cada uma dessas funções é necessário o suporte prestado pelos metadados.

4.4.1 O Processo ETL

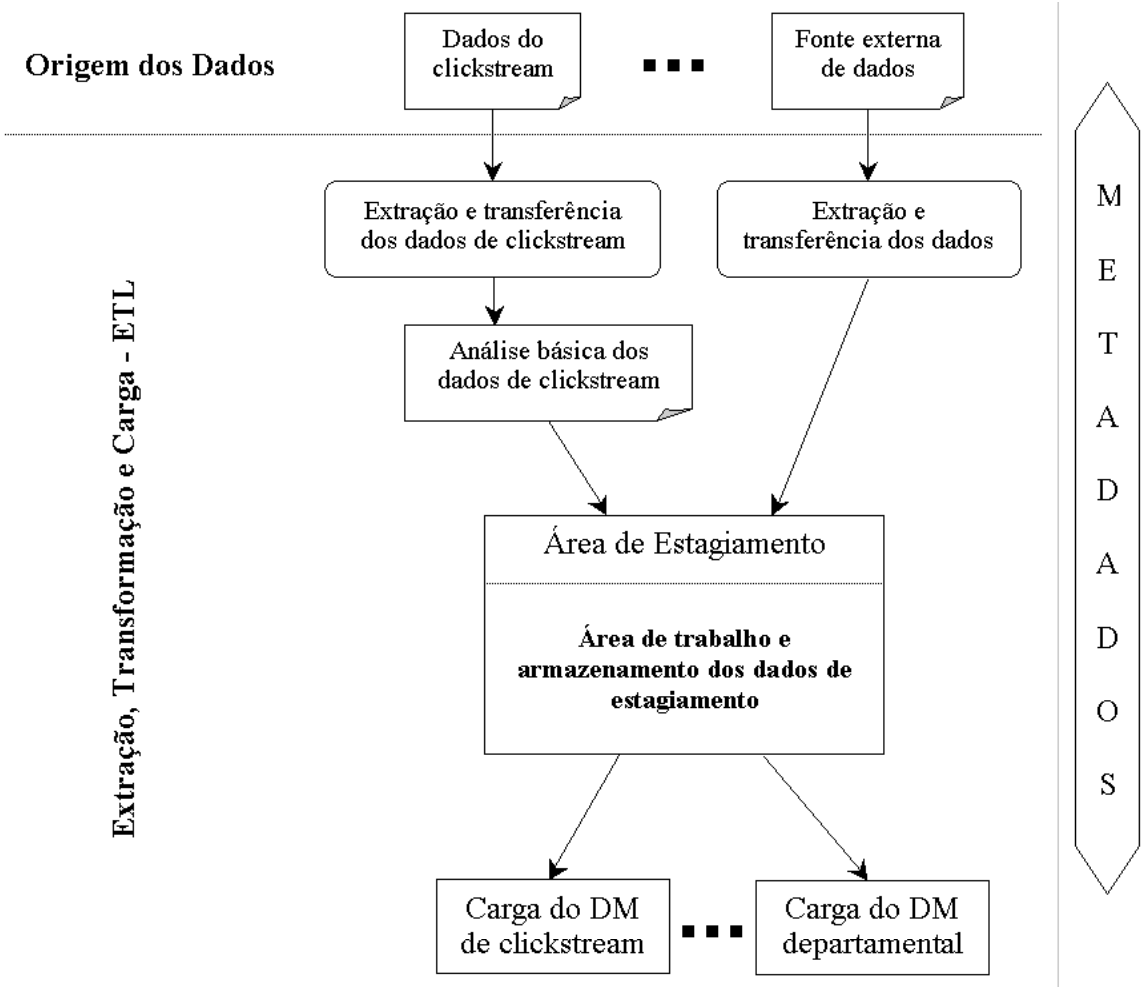
A natureza distribuída do conteúdo das páginas Web faz do processo ETL uma tarefa complexa, pois os dados podem estar armazenados em qualquer lugar onde a Internet alcança. Os componentes do ETL são construídos em uma área comum de estagiamento de dados, o esquema utilizado é apresentado na figura 12 e deve suportar os passos do processamento de ETL antes da carga final dos dados no warehouse.

À análise dos sistemas fontes será feita de forma a compreender os dados distribuídos pela organização para que possam ser integrados, neste caso além dos dados de clickstream podem ser incorporados dados adicionais dos sistemas de administração que podem ajudar na identificação de usuários e na formação de outros conjuntos de informações úteis para o data mart em desenvolvimento.

A carga dos dados para área de estagiamento deve utilizar o método incremental com alimentação contínua de novos dados, essa forma foi escolhida por não carregar o hardware que está servindo o site da Web e para que o volume de dados a ser transformado no próximo passo também não carregue o hardware de processamento ETL.

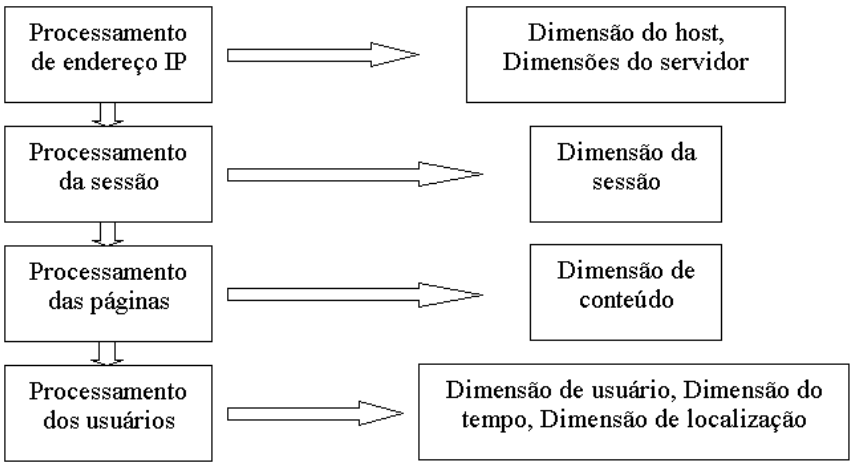
O objetivo da análise básica dos dados de clickstream é construir uma lista inicial da fonte de dados dos sistemas. Nesse ponto se analisa como os dados estão sendo gerados e quais as modificações são necessárias no site da Web para produção do formato dos dados esperados, é onde a equipe de desenvolvimento do DM tem maior contato com a equipe de desenvolvimento da página do site da Web.

Figura 12 - A Área de Estagiamento do Modelo



O processamento dos dados de clickstream não precisa ser feito serialmente e muitas tarefas podem ser executadas concorrentemente, mas as etapas de identificação, combinação de sessões, identificação da página e identificação do usuário devem ser executadas seqüencialmente para melhorar a eficiência da agregação dos dados (Sweiger, 2002). O modelo de estagiamento dos dados é mostrado na figura 13.

Figura 13 - Hierarquia de Processamento no Estagiamento



- O processamento do endereço IP objetiva esclarecer nomes de domínio e regiões geográficas onde o cliente está utilizando os serviços.

- O papel primário do processamento de sessão é coletar e marcar todos os eventos de página que ocorreram durante uma única sessão identificável de usuário. É preciso definir um tempo limite para o encerramento da sessão, como por exemplos cinco minutos sem atividade.

- O processamento da página identificará seu conteúdo e a atividade, além de identificar informações fornecidas pela URL destino e origem.

- O processamento de usuários deverá retirar todas informações e identificações de um usuário ou de um grupo de usuários. O modelo proposto utilizará um sistema de identificação para cada máquina, assim será possível identificar a máquina que está navegando pelo site da Web.

A carga dos dados será o passo final da fase de ETL e utilizará as regras existentes no metadados, as quais determina a documentação das tabelas e atributos na área de estágio e no DM destino.

4.5 Armazenamento

Antes que esta primeira versão do data mart seja implementada, segundo recomendações de Kimball e Merz (2000), os seguintes itens devem ser identificados, seguindo a ordem:

1. As questões de negócio a serem endereçadas por esse DM;
2. A fonte de dados de medição numérica (fato);
3. O grão de cada tabela de fatos proposta;
4. Todas as dimensões necessárias por todas as tabelas de fatos propostas;
5. Um plano e um compromisso para adaptar tais dimensões por toda empresa;
6. Todos os fatos numéricos a serem incluídos em cada tabela de fatos;
7. Um plano e um compromisso para adaptar quaisquer fatos que apareçam em mais de uma tabela de fatos por toda a empresa.

Para satisfazer as etapas um e dois, a primeira tabela de fatos do DM será implementada para analisar um simples *click* no site da Web, uma simples requisição para uma página e deve ser baseada nos dados de clickstream derivados dos registros do site da web em questão.

Neste modelo o objetivo é mapear todos os passos do visitante ao site da Web, para isso o grão escolhido é de um registro para cada *click* do cliente, satisfazendo o item três e tornando mensurável a quantidade de *clicks* no site.

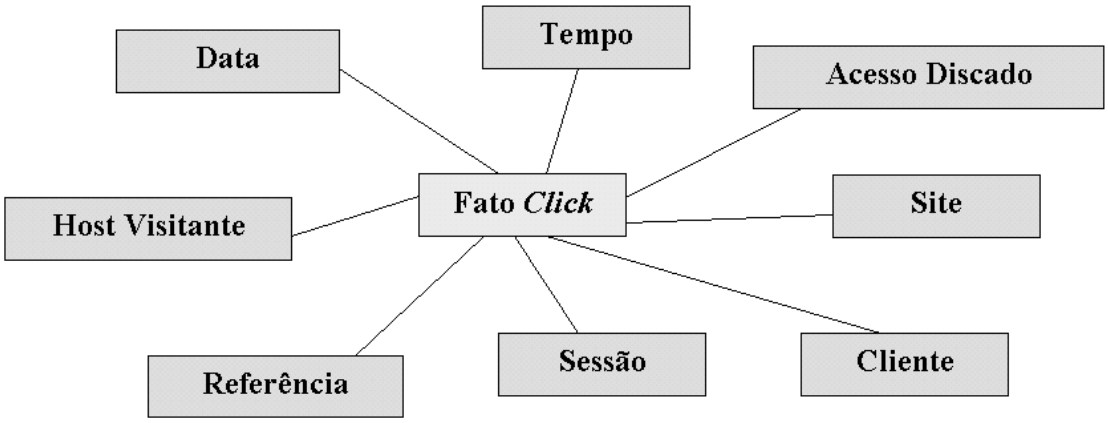
Com base na literatura pesquisada e também nas entrevistas com futuros usuários do sistema as dimensões julgadas apropriadas para essa primeira tabela de fatos são: data do calendário, horário do dia, host visitante, cliente, página, acesso discado, origem do link (referência) e sessão, satisfazendo o item quatro.

A dimensão origem do link ou URL captura qual página Web o usuário tinha visitado um passo antes, a dimensão data do calendário captura a data em que o usuário visitou o site, a dimensão sessão captura a sessão à qual esse *click* pertence e informações adicionais como, por exemplo, o começo e o fim da sessão de uma página. Como o foco do modelo é dar ênfase aos clientes do provedor, que atua regionalmente, a dimensão hora do dia captura o horário da solicitação no servidor da Web.

O acesso discado é a única dimensão deste modelo que não tem origem nos logs do site da Web, ela trás informações sobre os usuários que utilizam o serviço Web a partir de uma conexão *dial-up* fornecida diretamente pelo provedor em estudo.

Por fim, adiciona-se o conjunto de fatos medidos para esses *clicks*, incluindo quantidade de *clicks*, o status da solicitação do *click* e os bytes transferidos, satisfazendo o item seis. O esquema final é apresentado na figura 14.

Figura 14 - O esquema do Fato *Click*

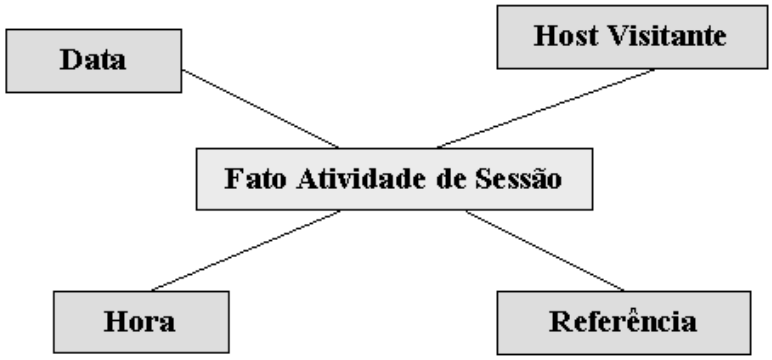


A vantagem de usar uma tabela de fatos de *click* é a alta qualidade dos dados armazenados, porque o grão tem uma fina granularidade, deste modo pode-se analisar um simples nível de *click*. Por outro lado, se o site da Web tem um volume de visitas considerado elevado o processo ETL pode ficar sobrecarregado, prejudicando o andamento do projeto.

4.5.1 Propondo um Nível Dual de Granularidade

A Tabela de fatos proposta anteriormente pode se tornar bastante grande devido ao alto nível de granularidade, dificultando, dessa forma, consultas específicas. Para solucionar esse problema é proposto um esquema de fatos de sessão, que é um modelo de informação sobre a sessão completa do usuário em um site da Web, a figura 15 ilustra esse modelo.

Figura 15 - O Esquema do Fato Sessão



A dimensão data captura a data do servidor em que a sessão ocorreu, a dimensão hora do dia captura a hora do servidor em que ocorreu a sessão, a dimensão host visitante captura informações do host e a identificação fornecida pelo servidor e a dimensão referência captura, quando houver, a origem do host visitante – de que site da Web parte o acesso ao site em estudo.

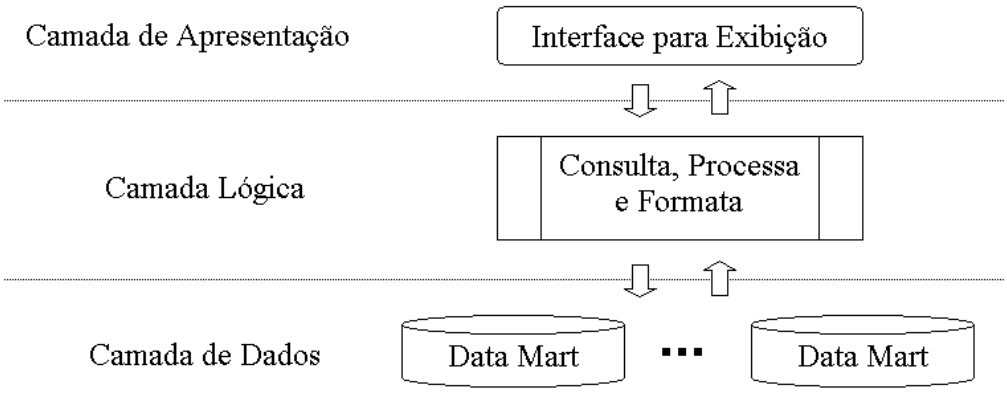
A vantagem de usar os fatos de sessão em relação aos fatos de *click* é a facilidade de seguir o usuário através do site da Web. Todavia, é impossível reconhecer um simples *click* depois da mudança de granularidade, porque está informação estará perdida no processo de transformação.

4.6 Apresentação

Para apresentação dos dados disponibilizados no DM é proposta a implantação utilizando uma arquitetura de particionamento em três camadas envolvendo a base de dados, os componentes lógicos e o aplicativo da camada de apresentação, conforme a figura 16.

Nessa topologia de três camadas, segundo Harrison (1998), é possível modificar, por exemplo, a camada de apresentação sem causar impacto sobre a camada lógica correspondente, facilitando eventuais manutenções necessárias a cada camada.

Figura 16 - Arquitetura em Três Camadas



O modelo de particionamento de aplicativos em três camadas permite que os componentes lógicos sejam executados em um servidor separado do Data Mart e da exibição, o que permite o uso de máquinas menos robustas na camada de apresentação. Com isso, às licenças de utilização de servidor são apenas para os servidores lógicos, representando uma boa economia financeira em software.

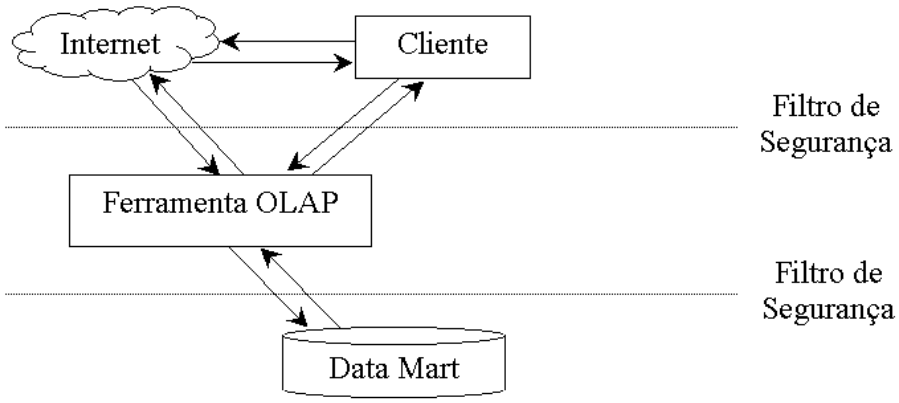
A função básica da camada lógica é transformar os dados brutos do DM em informações úteis que possam ser retornadas ao cliente. É proposta a utilização de uma ferramenta OLAP, que por definição (Harrison, 1998) executa cinco funções: interface, consulta, processo, formato e exibição.

Uma arquitetura de três camadas que separa a lógica de acesso aos dados da lógica do aplicativo e da lógica de apresentação oferece três vantagens distintas: desempenho, flexibilidade e escalabilidade.

4.7 Segurança no Data Warehouse de Clickstream

Como mencionado no capítulo anterior, a instalação do acesso de data warehouse no ambiente Internet aumenta a preocupação com a segurança sendo preciso traçar diretivas que proporcionem um ambiente seguro. No modelo proposto deverá haver uma hierarquia de acesso aos dados: O data mart só poderá ser acessado pela camada lógica, ficando o usuário impossibilitado de acesso direto, a figura 17 ilustra esse modelo.

Figura 17 - O Ambiente de Produção



Os aspectos de segurança que precisam ser abordados pela estratégia do aplicativo incluem o uso de nomes e senhas de usuários, incluindo os meios pelos quais os nomes e senhas viajam entre as camadas de aplicativos multicamadas distribuídas. No mínimo, a estratégia deverá definir exigências para a criptografia de nomes e senhas, mas poderá se estender a diretrizes para criptografia de todas as transações entre o cliente e o servidor.

Como foi abordada no capítulo anterior, a estrutura de segurança do DM pode ser baseada em quatro elementos:

- ✓ Dois fatores para autenticação;
- ✓ Uma conexão segura;
- ✓ Definição dos papéis dos usuários;

- ✓ Acesso a todos os objetos do DM controlados por papéis.

Também é necessário definir os métodos de detecção de lapsos de segurança em bancos de dados e implementação de ações corretivas para minimizar os danos no caso de uma incursão, fatores esses que podem ser implementados quando o data mart estiver em produção.

4.8 Conclusão

Neste capítulo foi apresentado o modelo para implantação de um data warehouse de clickstream com todas as etapas, a partir de um data mart incremental. Discutiu-se o porque da utilização da arquitetura DW Bus mostrando suas vantagens para a aplicação em questão. Os principais pontos da arquitetura, para que o sistema possa suportar a análise de dados que vem da Web, segundo seus componentes são:

- Os requisitos para que os dados que comporão o warehouse tenham sentido devem ser definidos antes da coleta, através das etapas de inclusão de código de conteúdo, rótulos de página e um sistema integrado de sincronismo de servidores;

- O processo de extração proposto define, além dos dados de clickstream, fonte de dados externa. O processamento incremental foi adotado por não sobrecarregar a área de estagiamento;

- Foi apresentada uma seqüência de execução na área de estagiamento de dados que se seguida, deverá melhorar a agregação dos dados;

- Para o armazenamento dos dados no data mart foi sugerida uma lista de requisitos, originalmente proposta por Kimball e Merz (2000), que impõe limitações para que a implantação de um primeiro modelo produza resultados o mais rápido possível. Como esse modelo pode-se tornar volumoso, dificultando a navegação dos dados, uma tabela de fato sessão é proposta baseada na tabela de fato *click* original;

- Para a apresentação dos dados foi proposta uma arquitetura de três camadas: base de dados, componentes lógicos e apresentação. Os motivos foram modularidade e

possível redução de custo financeiro em software;

- O tópico de segurança foi abordado por que os dados do data mart serão publicados no ambiente da Internet. Algumas regras de segurança foram propostas para entrar em operação juntamente com o data mart quando este estiver em operação para os clientes.

No próximo capítulo será apresentado um estudo de caso (protótipo) onde foi empregada a arquitetura apresentada neste capítulo.

CAPÍTULO 5 – APLICAÇÃO

5.1 Introdução

Neste capítulo é apresentado o desenvolvimento do protótipo proposto no modelo apresentado no capítulo anterior de um data mart de clickstream para um provedor de acesso à Internet. O objetivo é implementar o processo completo da construção do ambiente DM, apreciando na prática os pontos de maior dificuldade. O módulo do sistema de informação foi implantado a partir de dados colhidos nos servidores da organização.

Primeiramente, foram relacionados os aspectos abordados na análise da evolução do portal e o momento atual, além dos fatores considerados no processo de tomada de decisão para evolução do site da Web. Em seguida, foram apresentados detalhes da implementação das fases da arquitetura. Dentro das possibilidades e restrições do modelo foram apresentados alguns resultados obtidos.

5.2 Requisitos do Sistema de Informação

A empresa em que o protótipo foi desenvolvido atua no mercado de provimento de acesso à Internet desde novembro de 1995. Sua área de atuação é a região oeste do Paraná envolvendo as cidades de Cascavel, Toledo, Medianeira, Marechal Cândido Rondon e Capitão Leônidas Marques. Com essa área geográfica a empresa atende diretamente cerca de 5.000 usuários que estão divididos entre acesso discado, conexão ADSL e via rádio freqüência.

Devido a natureza exploratória do trabalho de pesquisa, em um primeiro momento não se obtém uma especificação completa de todos os requisitos. Através de reuniões e entrevistas com os potenciais usuários do DM, seguindo recomendações expostas por Kimball e Merz (2000), foram levantados alguns objetivos e, também, identificadas as necessidades de informação a partir de algumas situações descritas a seguir:

- ✓ Valores quantitativos a respeito de visitas ao site da Web.

- ✓ Estudo da utilização do site da Web versus a forma de acesso seja por rede local do provedor, de forma discada ou de visitas externas.

- ✓ Formatação de sessões completas no site para estudos temporais que contribuam para melhoria da seqüência de informações disponibilizadas no site.

- ✓ Quais os hosts externos e internos que mais fornecem visitas ao site e no caso de busca externa, quais as palavras-chave utilizadas na pesquisa.

- ✓ Quais os principais navegadores utilizados pelos visitantes.

A partir da identificação inicial dos requisitos, foi dado início a uma investigação das possíveis fontes de dados para concepção de um sistema de informações. Como os servidores que fornecem os dados são ricos em detalhes e podem prover os requisitos iniciais, a única fonte externa de dados foi a base de autenticação de usuários discados.

5.3 Requisitos do Site da Web em Estudo

Para que os dados coletados através da navegação no domínio em estudo tivessem sentido, todo o site teve que ser reestruturado. Cada página Web recebeu uma identificação própria com informações sobre o seu conteúdo, como: Identificação da página, versão e data de última atualização. Esse conjunto de informações forma os metadados das páginas e contribuem para a formação dos rótulos de conteúdo.

Todos os *links* envolvidos no site, que de alguma maneira chamam sites externos ou fazem referências a *banners* de propaganda, são providos de *strings* de consulta. Cada *string* foi classificada para atender determinada ação, por exemplo, quando no site se faz referência a um *banner* de propaganda. A string de consulta que aciona o link é “frame_banner.asp”. No caso deste exemplo, é interessante a informação de como o navegante chega ao site referenciado, contribuindo para a clareza da seqüência de passos desenvolvida no site.

5.3.1 Sincronismo

Através do serviço NTP – *Network Time Protocol*, aceito mundialmente como uma referência padrão de tempo e conhecida como UTC (*Universal Time Coordinated*), todos os servidores da rede local da organização que, de alguma forma participam dos serviços do site da Web, tiveram seus relógios sincronizados.

O NTP implementa um modelo de sincronização hierárquico distribuído e para o modelo foi utilizado o modo cliente/servidor. No topo encontram-se os servidores de tempo, conhecidos como relógios de referência de altíssima precisão que são atualizados por receptores GPS *(Global Positioning Systems)*. Estes servidores são disponibilizados pela RNP – Rede Nacional de Pesquisa através dos endereços "ntp1.rnp.br" e "ntp.pop-zz.rnp.br" (RNP, 2002).

Na rede local foi implementado o servidor NTP denominado cliente que busca informações de sincronismo com os servidores principais e está disponível no endereço "ntp.certto.com.br". Este cliente é o responsável pela sincronização dos relógios das máquinas que servem a aplicação em estudo na LAN. Com isso a seqüência de dados armazenados nos servidores de log segue uma lógica temporal e dá sentido a várias transações que podem ocorrer quase que simultaneamente ao longo do tempo.

5.3.2 Servidores Web

A organização trabalha com servidores de site da Web de dois fornecedores, a Fundação Apache – com o servidor Apache® e a Microsoft® – com o servidor IIS *(Internet Information Server)*. Nesse ponto do trabalho, o objetivo foi normalizar os dados de log gerados pelos servidores em um único repositório, para facilitar o processo de extração e transformação dos dados efetuados posteriormente.

O servidor IIS da Microsoft® foi o escolhido como repositório central de logs por ter o maior volume de páginas Web do domínio e também por oferecer facilidades de configuração e operação. O armazenamento desses dados foi separado por arquivos diários automaticamente, visando também facilitar o processo de carga para a área de estagiamento de

dados.

5.3.3 Servidores de Log

Para identificar o host que navega pelo site da Web foi concebido um servidor de cookie persistente que atendende a todas as páginas que fazem parte do domínio em estudo e fornece uma identificação padrão cookie armazenada na máquina do cliente até um prazo estipulado por programação. O *script-cookie* responsável pela geração do cookie no host visitante cria uma identificação numérica única para cada host, acrescentando esta informação em uma tabela da base de dados do domínio. Caso o host já tenha sua identificação anterior o campo da tabela não é incrementada.

Com o servidor centralizado, qualquer consulta a uma página Web no domínio acarreta na execução do *script-cookie*. Se o servidor de site da Web é o IIS o *script*, que é escrito em ASP, executa em modo nativo, se o servidor acessado é o Apache®, cada página publicada tem um campo especial que invoca o *script-cookie* no servidor para consulta na tabela e eventual criação da identificação. O código do *script-cookie* é apresentado no anexo D.

Com a utilização do servidor de logs da Microsoft®, percebeu-se que consegue-se gerar uma identificação de sessão automaticamente, cada *click* disparado pelo navegante armazena um conjunto de informações no servidor. O formato dos dados é separado em dois campos codificados: um que representa a data e o outro representando *clicks*. O conjunto dessas informações forma sessões distintas para cada janela do navegador facilitando a classificação das sessões dos hosts.

5.4 Modelagem Dimensional

O objetivo do protótipo é apresentar aos tomadores de decisão da organização, em um curto espaço de tempo, o que vem a ser um sistema de DM e como esse sistema poderá contribuir para melhorar a qualidade das decisões e uma conseqüente melhoria dos serviços disponibilizados aos seus clientes.

Os dados armazenados nos servidores de log são irrelevantes, se eles não estiverem

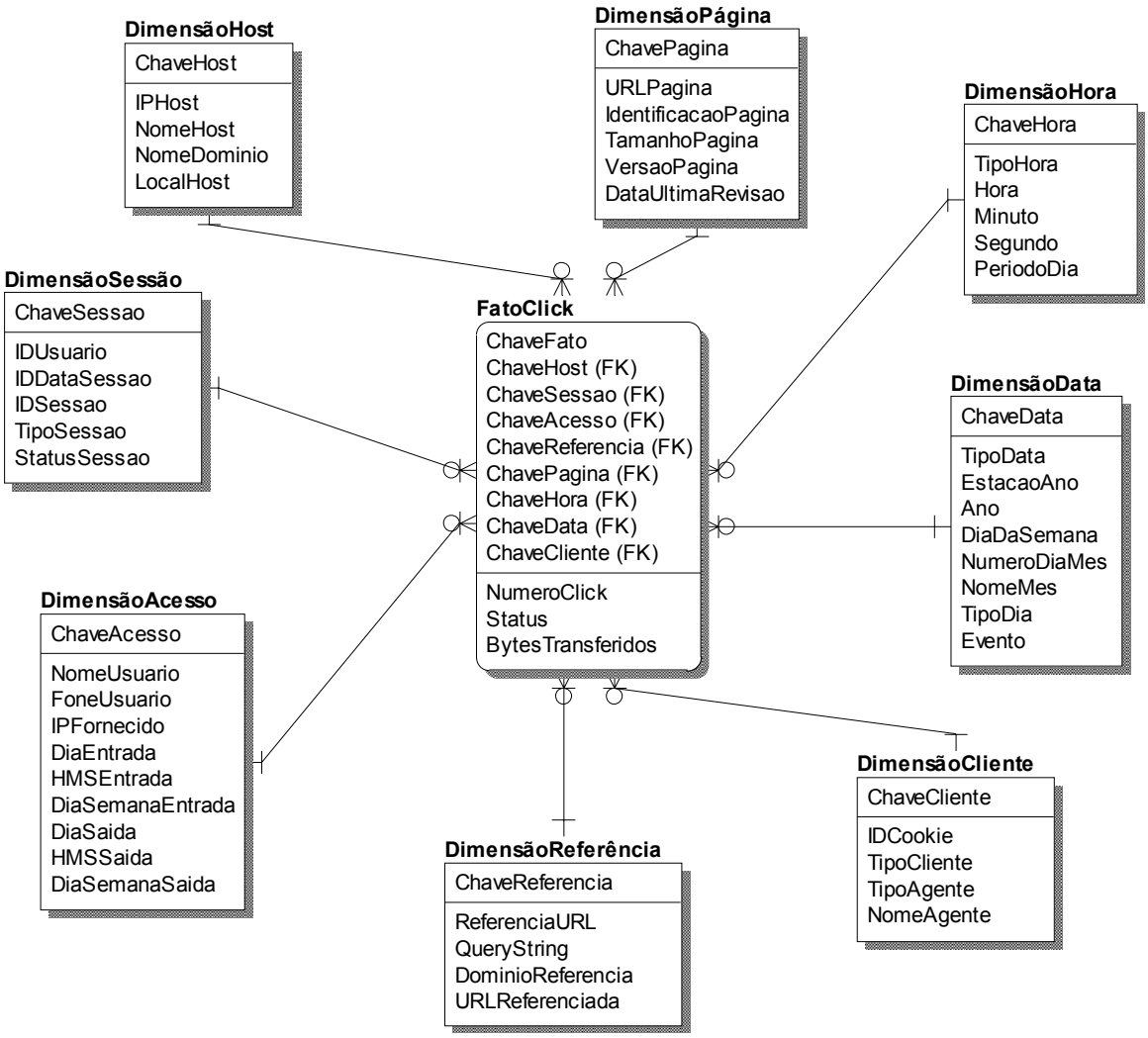
organizados de forma a agregar valor. O esquema estrela ou modelagem dimensional é uma abordagem que pode dar sentido a esses dados provendo de informações os usuários envolvidos no processo de tomada de decisão. O protótipo é construído segundo o modelo estrela e o quadro 10 apresenta os elementos, seus conceitos e as medidas relativas ao seu desenvolvimento.

Quadro 10 - Elementos do Modelo Desenvolvido

ELEMENTO	CONCEITO	MEDIDAS
Tabela de Fatos	Tabela onde cada fato representa um *click* medido por visitante do site da Web.	Dados quantitativos abordados no fato *click* • Numero de *Click* • Status • Bytes Transferidos
Tabelas de Dimensão	Tabelas que guardam um conjunto de relações relativas ao fato medido	Dados relativos • Host Visitante • Sessão • Página • Hora • Data • Cliente • Referência • Acesso

Após análise junto aos usuários do protótipo para discutir quais informações poderiam ser relevantes nesse primeiro modelo, sem afetar o tempo de construção do protótipo, foram apresentados os atributos de cada dimensão. O esquema estrela do fato *click* é apresentado na figura 18 e foi implementado em um servidor SQL Server da Microsoft®.

Figura 18 - Esquema Estrela do Fato *Click* Proposto no Modelo



5.5 Ferramentas de Implementação

O conjunto de ferramentas utilizado para desenvolvimento do protótipo foi o *Microsoft SQL Server – Enterprise Edition*® e para implementação do protótipo os seguintes componentes foram utilizados:

Microsoft® SQL Server™ 2000 Data Transformation Services (DTS) que é composto de um conjunto de ferramentas gráficas e objetos programáveis com o objetivo de extrair, transformar e consolidar dados de fontes diversas em simples ou múltiplo destino. Essa ferramenta foi escolhida para fase de transformação dos dados pelas seguintes razões:

- Apresenta funcionalidades de um *workflow*[1];

- Permite especificar fontes e destinos de dados OLE DB ou ODBC, incluindo SGBDR, arquivos texto ou planilhas Excell.

Microsoft® SQL Server™ 2000 com um conjunto de componentes que trabalham adequadamente para armazenamento e análise de dados necessários para organização com sistemas de processamento de dados. Os tópicos da arquitetura SQL Server descreve como os vários componentes trabalham juntos para gerenciar os dados efetivamente. Os seguintes motivos foram considerados para o uso dessa ferramenta:

- Integração completa com a Internet;

- Fácil Instalação e uso;

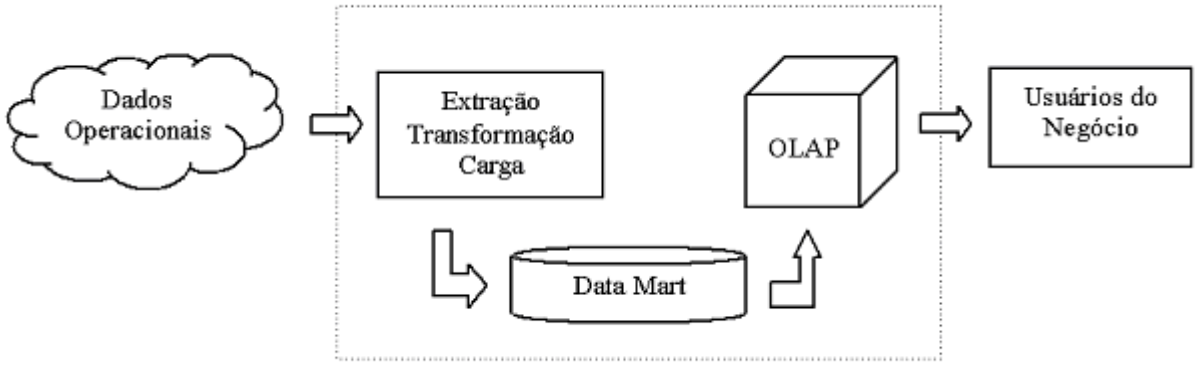
- Incorpora ferramentas de extração e análise de dados para processamentos OLAP.

Como ferramenta para serviços de análise foi utilizado o *Microsoft Analysis Management®*. Esta ferramenta oferece capacidade de navegação *drill-through*, exporta relatórios para *Excell*, HTML e *Power Point* além de possuir algumas funcionalidades para construir modelos de Data Mining.

Através da configuração do *Analysis Services* para conexão, a fonte de dados que contém as informações necessárias para construção do objeto em estudo, que se encontra no *Microsoft SQL 2000®*, é construído um cubo multidimensional de dados que seta fatos, dimensões e valores mensuráveis. Depois da construção dos cubos, se optou pelo modelo de análise OLAP multidimensional (MOLAP). Por último, o cubo é processado e disponibilizado para acesso. A figura 19 ilustra uma visão global do processo.

[1] Um *Workflow* é definido como uma coleção de tarefas organizadas para realizar um processo, quase sempre de negócio. Essas tarefas podem ser executadas por um ou mais sistemas de computador, por um ou mais agentes humanos, ou então por uma combinação destes. A ordem de execução e as condições pelas quais cada tarefa é iniciada também estão definidas no *Workflow*, sendo que o mesmo é capaz ainda de representar a sincronização das tarefas e o fluxo de informações (Moro, 1998).

Figura 19 - Visão Global do Processo de Construção do DM



Utilizando o Analysis Server o particionamento em três camadas, proposto pelo modelo pode ser testado, funcionando adequadamente. Para um futuro desenvolvimento do projeto completo do DM em produção podem ser utilizados os recursos técnicos de segurança que a ferramenta oferece.

Faz-se oportuna a citação de que foram efetuados testes com a ferramenta de extração de dados *Data Junction*® (2002). Essa ferramenta deveria ser utilizada no processo ETL por que se mostrou eficaz e rápida na transformação e filtragem dos dados e também por que tem a importante característica de automatização do processo de carga em uma base de dados pré-estabelecida. Infelizmente não pode ser utilizada porque a cópia cedida pelo fornecedor analisa somente os primeiros 1000 registros truncando processamentos posteriores.

5.6 Geração de Chaves

Com relação às chaves do DM de clickstream desenvolvido, foram seguidas duas estratégias sugeridas por Kimball:

- ✓ Criação de chaves seqüenciais inteiras;
- ✓ Criação de chaves genéricas.

Ambas as alternativas sobre as chaves foram fundamentadas na economia de espaço de armazenamento em grandes tabelas de fatos; melhoria de performance em operações de junção

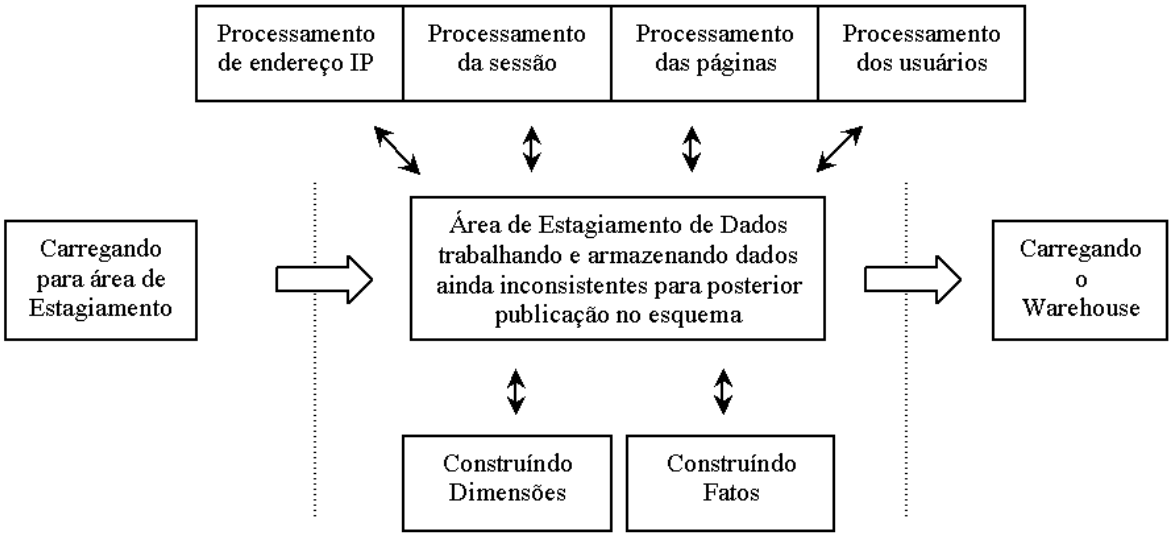
entre tabelas de fatos e de dimensões; e flexibilidade apresentada pelas chaves seqüenciais no acompanhamento de mudanças nas dimensões, visto que assumem valores independentes dos sistemas transacionais.

5.7 Carga dos Dados

O processo de carga dos dados tem um conjunto de procedimentos necessários para alimentar o banco de dados. Neste protótipo, a carga não é automatizada. Uma vez que o esquema seja validado a implementação do processo de carga deve ser automatizada, bem como prover a eliminação dos dados que expiram no DM.

A carga ocorre através da área de estagiamento de dados, uma estrutura intermediária que contempla o isolamento do DM em relação aos sistemas operacionais. Na área de estagiamento de dados várias transformações foram verificadas. Foi a etapa que exigiu mais processamento de hardware e procedimentos sendo considerada a mais cara no processo. A figura 20 ilustra a seqüência em que os dados foram tratados.

Figura 20 - Porção da Arquitetura Referente ao Processo ETL

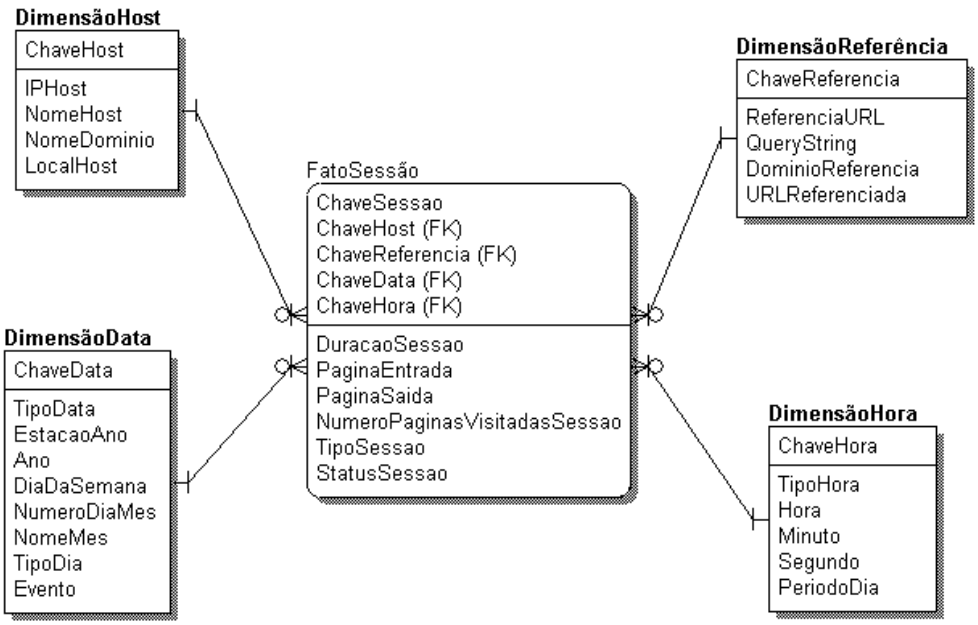


Nas tabelas dimensionais: página, host visitante, referência, cliente, acesso e sessão, onde a cada estagiamento de dados os campos são incrementados e um processo de pré-análise e incremento de novos dados foi implementada. Assim quando se dá à carga do fato *click* a integridade dos dados fica garantida, e a carga final para o warehouse só ocorre se as etapas intermediárias são bem sucedidas.

5.8 Construção do Nível Dual de Granularidade

Foi sugerido no modelo a construção do fato sessão, que sairá dos níveis de detalhe impostos pelo fato *click*, de acordo com a hierarquia contida na dimensão sessão do modelo original. O resultado foi uma tabela adicional e resumida do fato sessão e, neste caso, o grão utilizado no fato foi uma sessão completa. A figura 21 ilustra o esquema da tabela fato sessão, com o anexo E trazendo o script de carga.

Figura 21 - Esquema Fato Sessão Proposto do Modelo



As dimensões utilizadas são herança do fato *click* e os atributos escolhidos para povoar o fato sessão: duração de sessão, página de entrada e saída, número de páginas visitadas na sessão, tipo e status da sessão formam um importante conjunto de informações que podem ser

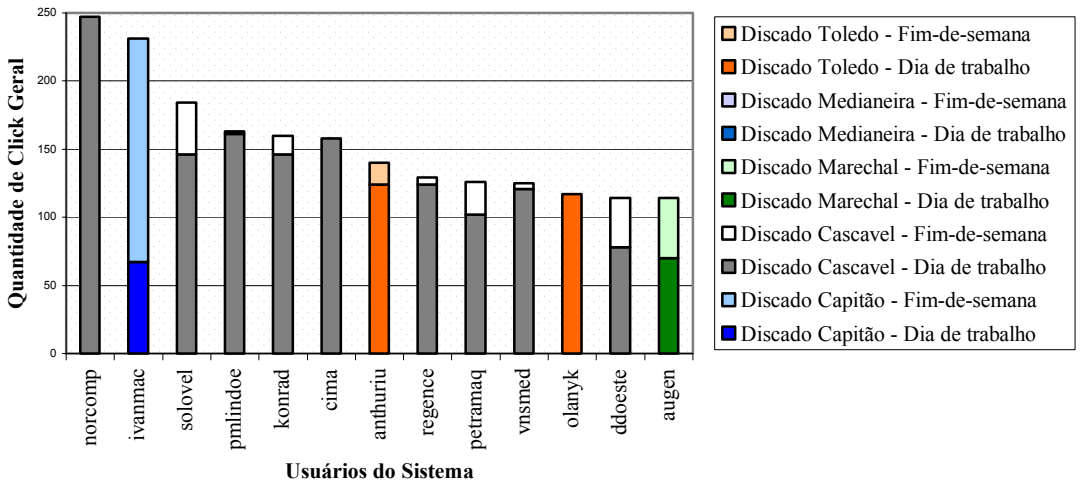
utilizadas para análise do comportamento de hosts visitantes que navegam pelo site da Web em estudo.

5.9 Análise de Resultados

Com base no levantamento de requisitos, a seguir são apresentadas algumas consultas executadas no DM implantado, com o objetivo de buscar informações relevantes para o aumento da carteira de clientes e também onde se pode melhorar o site da organização.

A ferramenta utilizada para moldar os dados do DM para este propósito foi o Analises Server Microsoft®, através da construção de Cubos específicos e enxutos, que obedece alguns procedimentos, como a seleção das tabelas a serem utilizadas como os atributos, forma de acesso aos atributos e cálculos. Como *front end* para consulta, foi utilizado o Microsoft Excell 2000®, que oferece a facilidade de apresentar tabelas e gráficos dinâmicos diretamente do cubo OLAP pré-processado.

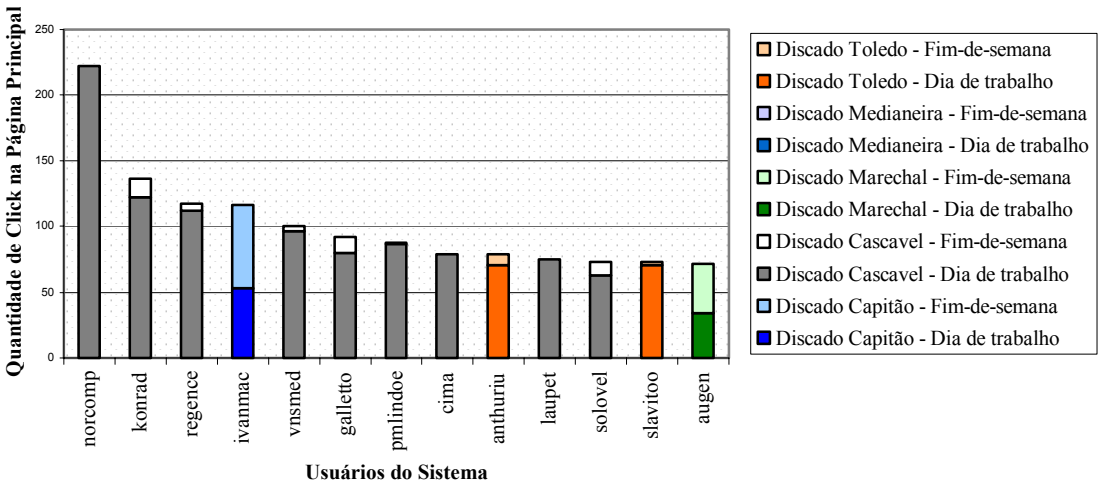
Figura 22 - Clientes Discados com o Maior Número de *Clicks* no Site



O estudo apresentado na figura 22 agrupa os clientes da organização atendidos pelo serviço de acesso discado (*dial-up*) que mais executaram *clicks* no site da Web do domínio em estudo. Neste caso o tempo considerado foi de uma semana completa separada em dia de

trabalho (segunda a sexta-feira) e fim de semana (sábado e domingo).

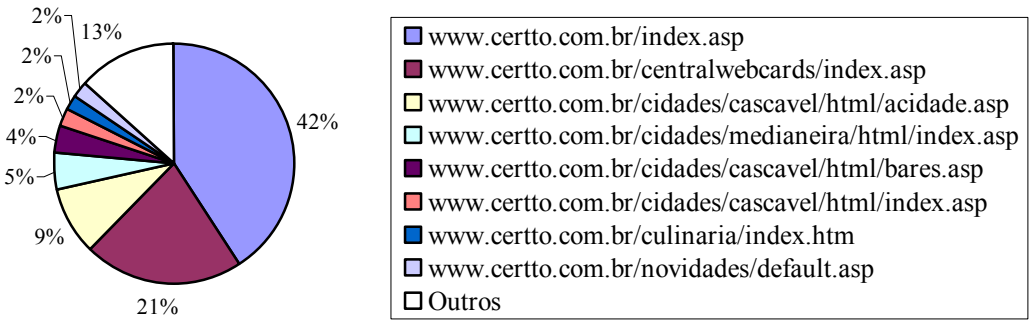
Figura 23 - Clientes Discados Versus Página Principal do Site do Domínio



Na figura 23 as mesmas características são mantidas, mas somente são computados os *clicks* na página principal. A grande maioria dos clientes que aparecem no gráfico é formada por empresas, apesar de não se ter feito distinção anterior. Procedendo-se uma análise, observa-se que nas empresas o acesso vinculado aos clicks do site é, basicamente, nos dias de trabalho. Com base nessas informações o departamento comercial da organização pode oferecer para estas empresas formas de acesso à Internet mais eficazes que forneçam um maior valor agregado ao faturamento da organização juntamente com uma diminuição dos custos do cliente.

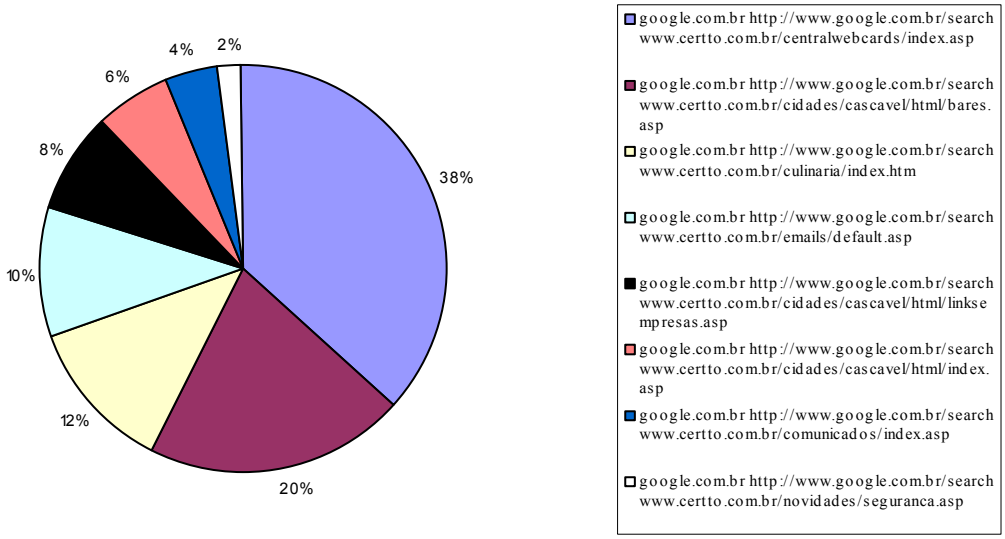
Numa segunda análise, o foco do estudo apresentado são as páginas de destino que todos os hosts considerados visitantes (todos os clientes conhecidos da organização foram descartados) ao site da organização apresentam em uma sessão completa. Como era de se esperar a página principal é a mais acessada com 42% da preferência. A figura 24 ilustra o gráfico.

Figura 24 - Páginas Destino dos Hosts Visitantes



A partir da generalização do gráfico da figura 24, foi elaborada uma pesquisa considerando os sites de busca como referência para chegada ao domínio da organização. O site de busca mais utilizado por hosts visitantes no intervalo de tempo pesquisado foi o Google® Brasil e as principais páginas da organizações que estabeleceram uma sessão são apresentadas no gráfico da figura 25.

Figura 25 - Páginas mais Pesquisadas Através do Site da Google® por Hosts Visitantes



A página mais referenciada da pesquisa, com 38% trata do serviço de cartões virtuais.

O segundo assunto mais pesquisado diz respeito à cidade virtual de Cascavel e aborda dois temas distintos: bares e empresas na cidade, com 34%. Em terceiro lugar aparece o tema do site que trata de culinária com 12% das preferências.

Com este estudo, a organização pode dar ênfase na melhoria dos serviços oferecidos à comunidade para sua revitalização, pois até então esses serviços eram tratados sem a devida atenção. Num segundo momento pode-se analisar as páginas que não obtiveram pesquisas. Esta informação pode ser utilizada para sustentar um plano de revitalização desses serviços.

5.10 Conclusão

Num futuro desenvolvimento do data mart de clickstream para a organização é preciso definir um repositório do conjunto de metadados de todo o processo, desde informações das páginas do site da Web, definições e restrições impostas até o formato como os dados são armazenados para análise. Com isso tem-se uma organização de todo conjunto dando sentido aos desenvolvedores e usuários do sistema de apoio à decisão.

Neste capítulo, apresentou-se o processo de implementação do protótipo para o modelo de DM de clickstream proposto. A aplicação da arquitetura se mostrou complexa e os principais pontos desenvolvidos, para que o sistema possa suportar a análise de dados que vem da Web, segundo seus componentes são:

- Para que o modelo implementado tivesse sentido foram abordados alguns requisitos iniciais propostos pelos potenciais usuários do DM, com isso uma linha inicial de desenvolvimento foi traçada.

- Antes de desenvolver os trabalhos específicos da construção de um DM o site em estudo teve que ser todo reestruturado, desde suas páginas, sincronismo no tempo dos hosts envolvidos, normalização dos servidores de aplicação utilizados até os servidores que armazenam os dados utilizados no DM.

- Foram apresentadas as vantagens e desvantagens do conjunto de ferramentas utilizadas na execução do protótipo, com ênfase nas ferramentas da Microsoft®

que juntas oferecem uma boa relação custo/beneficio.

- A fase de estagiamento dos dados foi a que apresentou maior dificuldade, pois além da utilização da ferramenta DTS da Microsoft®, um conjunto de scripts ASP teve que ser desenvolvido para adequar e enquadrar os dados no formato proposto pelo modelo lógico. Ficou constatado que a utilização de scripts ASP para esse serviço necessita de hardware poderoso, pois o servidor de BD e o servidor IIS devem estar executando concomitantemente.

- De acordo com o modelo proposto, foi criado o fato sessão, para responder questões específicas de sessões no site da Web.

- Por fim, alguns resultados foram apresentados utilizando-se a ferramenta Analysis Service da Microsoft® e o aplicativo Excell®, que corresponderam às expectativas de uma conexão modular e independente do data mart desenvolvido.

CAPÍTULO 6 – CONCLUSÃO E TRABALHOS FUTUROS

O trabalho desenvolvido teve como objetivo o desenvolvimento de um ambiente de data warehouse de clickstream voltado para um provedor de serviços Internet. Para suprir esta finalidade, um conjunto de tarefas teve que ser executado.

No capítulo 2 foi apresentada a trajetória da Internet no Brasil e como a organização em estudo se insere nesse contexto, apresentando algumas razões para o desenvolvimento de um DW, para que esta ferramenta possa colaborar em seu sucesso no mercado.

No capítulo 3 são apresentadas as técnicas de construção de um data warehouse voltado para clickstream, desde o planejamento do DW, análise, modelagem, processo, utilização até a segurança que deve ser empregada para ambientes que trabalham na Internet, a partir da visão de diversos autores com o objetivo de fundamentar o desenvolvimento.

No capítulo 4 é proposto o modelo de implantação de um data mart departamental com o objetivo de atender algumas das necessidades dos responsáveis a respeito da utilização do site da Web da organização e para isso são consideradas as particularidades do funcionamento operacional do provedor.

Os resultados da implantação do protótipo, que foi gerado a partir do modelo proposto, são apresentados no capítulo 5. As principais fases de criação de um data mart foram percorridas com a utilização de diversas ferramentas de apoio.

A Web é uma rica e incrível fonte de inteligência para negócios. A ferramenta de data warehouse construída apresenta o conjunto de conhecimento contido nos dados de clickstream através dos sites da Web e foi concebida para que analistas de negócios expandam seu marketing em direção ao estreitamento do relacionamento com o cliente. Também foi possível identificar os principais problemas que circundam a edição de sites da Web, ajudando a equipe de desenvolvimento a reduzir sua manutenção, promovendo assim uma diminuição no tempo de trabalho gasto para esse fim.

O estudo contribuiu para apresentar na prática um conjunto de ferramentas de suporte

à decisão para administração antes só vislumbrada por organizações de grande porte. Alguns dos resultados apresentados foram recebidos com grande interesse e despertaram a curiosidade por parte da organização em utilizá-la como diferencial competitivo no mercado.

Devido à extensa abrangência do tema foi necessário delimitar o escopo de atuação. Sendo assim, gostaria de deixar, a título de sugestões futuras, estudos a respeito de data warehouse de clickstream para sites da Web dinâmicos e um estudo da construção de componentes baseados em técnicas de data mining, para integração na arquitetura.

CAPÍTULO 7 – REFERÊNCIAS BIBLIOGRÁFICAS

APACHE SOFTWARE FOUNDATION. **Prove suporte ao projeto do software de código aberto Apache**. Disponível em: < http://www.apache.org>. Acesso em: 20 abril 2002.

BRACKETT, Michael H., **The Data Warehouse Challenge – Taming Data Chaos**. New York: J. Wiley & Sons Inc, 1996. 579p.

CRAIG, Robert S., et. al. Microsoft Data Warehousing – Building Distributed Decision Support Systems. New York: J. Wiley & Sons Inc, 1999. 384p.

FIRESTONE, Joseph M., **Architectural Evolution in Data Warehousing** (White Paper No. Eleven, July 1, 1998). Disponível em: <http://www.dkms.com/ARCHEV.html>. Acesso em 30 abril 2002.

HARRISON, Thomas H., **Intranet data warehouse.** São Paulo: Berkeley, 1998. 358 p.

HOMER, Alex et al. **Professional Active Server Pages 3.0.** Rio de Janeiro: Ciência Moderna, 2000. 1440 p.

HTTP/1.1 - Hypertext Transfer Protocol. Pedido para Comentários (RFC) 2616, referente as características de funcionamento do protocolo HTTP. Disponível em: <http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc2616.html>. Acesso em 20 abril 2002.

IBM - INTERNATIONAL BUSINESS MACHINES CORPORATION (USA). **Data modeling Techniques for Data Warehousing.** 1. ed. California, 1998.

IETF - The Internet Engineering Task Force. **Mecanismo de gerenciamento de estado do HTTP (Cookie)**. Disponível em: <http://www.ietf.org/rfc/rfc2109.txt>. Acesso em 20 abril 2002.

INMON, William H.,**Como Construir o data warehouse.** Rio de Janeiro: Campus, 1997a. 387p.

INMON, William H., HACKARTHORN, Richard D. **Como usar o data warehouse.** Rio de Janeiro: IBPI Press, 1997b. 277p.

JUNCTION, Data. **Software Provedor de Soluções Integradas**. Disponível em: < http://www.datajunction.com/>. Acesso em 2 abril 2002.

KIMBALL, Ralph. **Data warehouse toolkit.** São Paulo: Makron Books, 1998a. 387 p.

KIMBALL, Ralph, et. al. The data warehouse lifecycle toolkit: Expert methods for designing, developing data warehouses. New York: J. Wiley & Sons Inc, 1998b. 771 p.

KIMBALL, Ralph; MERZ, Richard. **Data Webhouse: Construindo o Data Warehouse para a Web**. Tradução: Edson Furmankiewicz, Joana Figueiredo. 1. ed. Rio de Janeiro: Campus, 2000. 367p.

LURIE, Martin P. **Web Click Stream Analysis using Linux Clusters.** O artigo explora o assunto de análise de clickstream usando clusters em Linux. Biblioteca on-line IBM. Disponível em: <http://www7b.boulder.ibm.com/dmdd/library/techarticle/lurie/0111lurie.html>. Acesso em: 23 fevereiro 2002.

MACHADO, Felipe N. R. **Projeto de Data Warehouse: Uma Visão Multidimensional**. São Paulo: Érica, 2000. 251p.

MARK, Madsen. **Implementing Aggregates for a Dimensional Data Warehouse**. O artigo explora como criar agregados apropriados para um data warehouse baseado em esquema estrela. Disponível em: <http://www.clickstreamdatawarehousing.com/aggregate_article.html>. Acesso em 27 março 2002.

MONCLA, Brenda. **Business Meta Data Integration – Making a Fundamental Paradigm Shift.** O artigo fala sobre a integração dos metadados de acordo com sua classificação. Setembro 1999. Disponível em: <http://www.thinkfast.com/_vti_bin/shtml.dll/whitepapers.html>. Acesso em: 28 abril 2002.

MORO, Mirela Moura. **Workflow na WEB**. Trabalho que apresenta definições de *workflow* para Sistemas de Bancos de Dados Distribuídos. Setembro 1998. Disponível em: <http://www.inf.ufrgs.br/~mirella/workflow/home.html>. Acesso em: 07 novembro 2002.

SEGURANÇA Máxima: **O guia de um hacker para proteger seu site na Internet e sua rede**. 2. ed. Rio de Janeiro: Campus, 2000. 826p. Autor desconhecido.

PEREIRA, Denise Maciel. **Uso do padrão OIM de metadados no suporte às transformações de dados em ambientes de data warehouse.** 2000. 217 f. Dissertação (Mestrado em Informática) – Instituto de Matemática e Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

RNP, Rede Nacional de Pesquisas. **Manual de a implementação de servidores NTP em redes locais.** Disponível em: <http://www.rnp.br/cais/ntp/ntp_manual.html>. Acesso em 20 março 2002.

ROGERS, M. O modelo CRM. **Revista HSM Management**, São Paulo, número 23, p.56-62, nov./dez. 2000.

SELL, Denílson. **Uma arquitetura para distribuição de componentes tecnológicos de sistemas de informações baseados em data warehouse.** 2001. 89 f. Dissertação (Mestrado em Engenharia de Produção) – Curso de Pós-graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis.

SEYBOLD, P. B., MARSHAK, R. T. Customers.com: How to create a profitable business strategy for the Internet and Beyond, New York: Random House, 1998. 564p.

SONG, Yeol; LEVAN-SHLTZ, Kelly. **Data Warehouse Design for E-commerce Environment.** New Jersey. NEC Research Institute, 1999. Disponível em: <http://citeseer.nj.nec.com/song99data.html> Acesso em: 20 novembro 2001.

SWEIGER, Mark et al. **Clickstream Data Warehousing.** New York: J. Wiley & Sons Inc, 2002. 452 p.

ANEXO A – ELEMENTOS DE DADOS DE LOG EM SERVIDORES DA WEB

Quadro 11 - Elementos de dados de log em servidor da Web

ELEMENTO DE DADOS	CLF	ECLF	DESCRIÇÃO
Host	✓	✓	Nome de domínio qualificado do cliente, ou seu endereço IP se o nome não estiver disponível.
Ident	✓	✓	Informações de identidade fornecidas pelo cliente, se o identd estiver ativado.
Authuser	✓	✓	Se a solicitação foi para um documento protegido por senha, então esse é o ID do usuário utilizado na solicitação.
Time	✓	✓	A data/hora em que a solicitação alcançou o servidor em formato de tempo CLF {zone dd/Mmm/aaaa:hh:mm:ss}.
Request	✓	✓	A primeira linha de solicitação do cliente.
Status	✓	✓	Código de status de três dígitos retornado para o cliente.
Bytes	✓	✓	Número de bytes retornado para o cliente excluindo cabeçalhos de http.
Referrer		✓	URL do servidor de referência.
User-agent		✓	Nome e versão do cliente (navegador).
Filename			Nome do arquivo.
Time-to-server			Tempo para atender a solicitação.
IP-Address			Endereço IP do host remoto.
Server-port			Porta canônica do servidor que satisfaz a solicitação.
Process-ID			ID do processo filho que fez a manutenção da solicitação.
Formatted-time			A data/hora.
URL-requested			O caminho do URL solicitado.
Server-name			O nome canônico do servidor que atende a solicitação.
Cookie			O valor do cookie recuperado do arquivo de cookie do cliente.

Fonte: Adaptado de Kimball e Merz, 2000.

ANEXO B – IDENTIFICAÇÃO DE PAPÉIS

Quadro 12 - Identificação de Papéis

Linha de Frente (Em itálico estão os papéis comuns, necessários em quase todo o projeto).	
Patrocinador do negócio	É o cliente final. Fornece a demanda e os recursos para o projeto e deve estar comprometido em realizar negócios na Web.
Patrocinador de IT	Disponibiliza os recursos para finalizar o projeto de Warehouse. Ele precisa compreender o ritmo do webmaster e o que torna o projeto mais fácil ou mais difícil.
Condutor do negócio	Quando o principal executivo patrocinador do negócio não estiver permanentemente disponível, um gerente de nível médio poderá servir como um *proxy*, ou um elo, com o patrocinador no negócio.

Treinadores: Gerentes de projeto e lideres	
Gerente-geral do projeto	Essa pessoa provavelmente vê a si mesmo como uma pessoa de data warehousing e não como um webmaster. Essa é uma posição de tempo integral, e o gerente de projeto precisa ser bom em múltiplas habilidades. Nunca esquecendo o fato de que ele precisa manter a perspectiva de todo o projeto.
Líder de projeto do negócio	Essa pessoa deve estar disponível diariamente para coordenar os trabalhos com o gerente do projeto, tendo contato direto com todas as etapas de desenvolvimento do projeto.

Linha de base: A equipe básica do projeto (Em negrito aparecem os novos papéis, necessários para o Warehouse de Clickstream)

Analista de negócio	Conduz no início do projeto as definições do requisito do negócio e, mais tarde, é o responsável pela arquitetura de consultas e relatórios padrão .
Modelador de dados	Precisa entender os dois tipos de modelagem: E/R e MD. Precisam ter um bom conhecimento da diferença entre os relacionamentos da modelagem de dados reais e da modelagem de dados abstratos.
Administrador de banco de dados	Responsável pela implementação do banco de dados, desempenho, utilização do espaço, backup e recuperação.
Administrador de sistemas do site da Web	Administra todos os aspectos operacionais da hospedagem da Web. Responsável por escolher as plataformas de hardware e software. Pode ser responsável pela segurança do site.
Projetista de sistemas de pré-consolidação (staging) de dados	Responsável pelo sistema de ETL que alimenta o Warehouse. Deve ter experiência em servidores Web, bancos de dados, arquitetura cliente/servidor e especificação de vários componentes de ETL.
Desenvolvedor de aplicativos de usuário final	Precisa ocupar-se com os usuários finais, entendendo suas necessidades e aspirações.
Desenvolvedor de CRM	Deve estar preocupado em entregar em tempo real a mais eficiente interface com o usuário ao cliente que bate a porta do site da Web. Deve ter interesse no resultado da análise de seqüência de cliques.
Analista de comportamento de cliente	Compreender a seqüência de cliques e o relacionamento das transações capturadas por outros sistemas.
Professor de Warehouse de Clickstream	Entender e ensinar o conteúdo de dados, bem como, a utilização de ferramentas.
Webmaster	Responsável pela aparência e comportamento globais do site da Web.
Gerente de conteúdo	È a chave para o projeto de Warehouse e pode, de fato, ser um consumidor primário da produção do Warehouse.
Gerente de segurança do Warehouse	Papel tático. Responsabilidade principal de definir e administrar as definições de papéis do usuário.
Especialista em descrição de página de site da Web	Define os atributos para página da Web, o que as torna compreensíveis em um contexto de banco de dados.
Arquiteto de rede e segurança	Papel estratégico. Responsável por toda arquitetura de segurança do Warehouse.
Especialista em suporte técnico	Responsável pelas várias partes de infra-estrutura que o Warehouse depende.
Programador da pré-consolidação de dados	Responsável pela implementação de *back room* dos sistemas ETL.
Especialista em extração de logs da Web	Trabalha com o webmaster para fornecer logs da Web que são os mais prolixos e significativos possíveis.

Administrador de dados	Responsável pela definição no Warehouse das dimensões adaptadas e dos fatos adaptados. O administrador de dados possui um papel central na definição dos metadados orientados para o negócio, tornando esses metadados úteis para os desenvolvedores de aplicativos e para usuários finais.
Especialista em suporte do sistema de produção	São atribuídos os sistemas de legado da empresa.
Gerente de garantia de qualidade	Papel estratégico na definição de uma qualidade aceitável de dados para o Warehouse.
Coordenador de garantia de qualidade	Faz o julgamento humano final de quando os dados estarão aptos para a publicação.

Fonte: Adaptado de Kimball e Merz, 2000.

ANEXO C – ENTREVISTA: REQUISITOS DO NEGÓCIO

Quadro 13 - Entrevista: Requisitos do Negócio

1. Preparando para entrevista O conceito mais importante para entrevista com o usuário final é ouvir o que eles fazem e não discutir o conteúdo técnico do DW. Em entrevistas de auditoria de dados, o objetivo é focalizar as principais fontes de produção dados, compreendendo de maneira realista a qualidade dos dados. 2. Conduzindo a entrevista A principal função do entrevistador é fazer com que o usuário final fale. As perguntas devem ser abertas e do tipo por quê, o que aconteceria se (cenários hipotéticos), e depois disso. Não é recomendado que as entrevistas sejam gravadas, evitando assim a intimidação. Um dos objetivos finais talvez seja estabelecer algum nível de confiança pessoal. 3. Assimilando os resultados da entrevista Imediatamente após a entrevista, as anotações devem ser examinadas e concluídas. Quando todas as entrevistas estiverem completas, todos os achados importantes devem ser destacados e organizados em categorias. 4. Publicando os resultados da entrevista Os resultados das entrevistas devem ser apresentados aos entrevistados, à gerência de usuários finais e à gerência de IT. Eles servirão como uma base muito valiosa para todas as decisões a serem tomadas em seguida.

Fonte: Adaptado de Kimball e Merz, 2000.

ANEXO D – SCRIPT DE CRIAÇÃO DO ID COOKIE

```
<%'==Classe que estabelece uma conexão com o servidor de Banco de Dados ===
  Class ClassConexao
    Public Conn
      Public Registro
      Private RegAberto
      Private StrConn

    Private Sub Class_Initialize()
        StrConn = "Provider = SQLOLEDB.1; Persist Security Info=False; "&_
                  "User ID=CERTTO; Password=CERTTO; "&_
                        "Initial Catalog=Certto2002; Data Source=localhost"
      Set Conn = Server.CreateObject("ADODB.Connection")
      Conn.Open StrConn
      RegAberto = False
    End Sub

    Private Sub Class_Terminate()
      If RegAberto then
        Registro.Close
        Set Registro = Nothing
      End If
      Conn.Close
      Set Conn = Nothing
    End Sub

    Public Sub CriaRegistro(SQL)
      If RegAberto then
        FechaRegistro()
      End If
      Set Registro = Server.CreateObject("ADODB.RecordSet")
      Registro.Open SQL, Conn, 1, 2
      RegAberto = True
    End Sub

    Public Sub FechaRegistro()
        Set Registro = Nothing
      RegAberto = False
    End Sub
  End Class
'================================================================================
  Set Conex = New ClassConexao

  SQL = "Select * from tb_contador"
  Conex.CriaRegistro(SQL)

  if Conex.Registro.eof then
     response.write ("fim de arquivo")
  elseif request.cookies("CerttoID")("Status") <> "ok" then
     Conex.Registro("Cont") = Conex.Registro("Cont")+1
     Conex.Registro.update
     response.cookies("CerttoID")("ID") = Conex.Registro("Cont")
     response.cookies("CerttoID")("Status") = "ok"
     response.cookies("CerttoID").expires = CDate("1/1/2010")
     response.cookies("CerttoID").domain = "certto.com.br"
  else
    ' response.write (request.cookies("CerttoID")("ID"))
  end if
  %>
```

ANEXO E – SCRIPT DE CARGA DO AGREGADO SESSÃO

```
<%@ language="VBScript" %>
<!--#include file="adovbs.inc"-->
<!--#include file="conexao.asp" -->
<!--#include file="funcoes.asp" -->

<%
dim rs_fato, sql, cont1, cont2

Server.ScriptTimeout = 90000000

set rs_sessao = Server.CreateObject ("ADODB.Recordset")
sql = "select * from DimensaoSessao"
rs_sessao.open sql, ConexaoBase, 1, 2

cont1 = 0
cont2 = 0

do while not rs_sessao.eof
    VarChaveSessao = rs_sessao("ChaveSessao")

      set rs_fato = Server.CreateObject ("ADODB.Recordset")
    sql = "select * from FatoClick where ChaveSessao = " & VarChaveSessao &
" order by ChaveData, ChaveHora"
    rs_fato.open sql, ConexaoBase, 1, 2

      PrimeiroRegistro = true
    VarNumeroPaginas = 0

      if not rs_fato.eof then
          do while not rs_fato.eof

               if PrimeiroRegistro then
                  PrimeiroRegistro = false

                    ' Selecionando Chaves de Data e Hora
                    VarChaveDataInicio = rs_fato("ChaveData")
                    VarChaveHoraInicio = rs_fato("ChaveHora")

                    ' Selecionando Chaves de Host e Referencia
                    VarChaveHost = rs_fato("ChaveHost")
                    VarChaveReferencia = rs_fato("ChaveReferencia")

                    ' Selecionando Data e Hora de Início da Sessão
                VarInicioSessao =
DataHora(VarChaveDataInicio,VarChaveHoraInicio)
                    VarFimSessao = VarInicioSessao

                    ' Selecionando Pagina de Entrada
                    if (rs_fato("ChavePagina") <> 0) then
                          set rs_pagina = Server.CreateObject
("ADODB.Recordset")
                          sql = "select * from DimensaoPagina where
ChavePagina = " & rs_fato("ChavePagina")
                          rs_pagina.open sql, ConexaoBase, 1, 2

                          VarPaginaEntrada = rs_pagina("URLPagina")

                          if rs_pagina.eof then
                           VarPaginaEntrada = "indefinida"
                          else
```

```
                            VarPaginaEntrada = rs_pagina("URLPagina")
                           end if

                           rs_pagina.close
                      else
                         VarPaginaEntrada = "indefinida"
                      end if

                      VarPaginaSaida = VarPaginaEntrada
                 else
                      ' Selecionando Chaves de Data e Hora
                      VarChaveDataFim = rs_fato("ChaveData")
                      VarChaveHoraFim = rs_fato("ChaveHora")

                      ' Selecionando Data e Hora de Fim da Sessão
                  VarFimSessao = DataHora(VarChaveDataFim,VarChaveHoraFim)

                      ' Selecionando Pagina de Saida
                      if (rs_fato("ChavePagina") <> 0) then
                          set rs_pagina = Server.CreateObject
("ADODB.Recordset")
                          sql = "select * from DimensaoPagina where
ChavePagina = " & rs_fato("ChavePagina")
                          rs_pagina.open sql, ConexaoBase, 1, 2

                            if rs_pagina.eof then
                             VarPaginaSaida = "indefinida"
                            else
                             VarPaginaSaida = rs_pagina("URLPagina")
                            end if

                           rs_pagina.close
                      else
                          VarPaginaSaida = "indefinida"
                      end if
                 end if

                 VarNumeroPaginas = VarNumeroPaginas + 1

               rs_fato.movenext
              loop

              if (VarInicioSessao = null) or (VarFimSessao = null) then
                 VarDuracaoSessao = null
              else
                 VarDuracaoSessao =
datediff("s",VarInicioSessao,VarFimSessao)
              end if

              Set ListaData = Server.CreateObject("ADODB.Recordset")
              Set ListaData.ActiveConnection = ConexaoBase

              ' Selecionando o Registro
              ListaData.Source = "SELECT * FROM FatoSessao2"
              ListaData.CursorType = adOpenStatic
              ListaData.LockType   = adLockOptimistic
              ListaData.Open

              ' Atualizando Registro
              ListaData.Addnew
              ListaData("ChaveDataInicio")= VarChaveDataInicio
            ListaData("ChaveHoraInicio")= VarChaveHoraInicio
            ListaData("ChaveHost")= VarChaveHost
            ListaData("ChaveReferencia")= VarChaveReferencia
```

```
            ListaData("IdSessao")= rs_sessao("IdSessao")
            ListaData("InicioSessao")= VarInicioSessao
            ListaData("FimSessao")= VarFimSessao
            ListaData("DuracaoSessao")= VarDuracaoSessao
            ListaData("PaginaEntrada")= VarPaginaEntrada
            ListaData("PaginaSaida")= VarPaginaSaida
            ListaData("NumeroPaginas")= VarNumeroPaginas
            ListaData("TipoSessao")= rs_sessao("TipoSessao")
            ListaData("StatusSessao")= rs_sessao("StatusSessao")

              ' Grava o registro no Banco de Dados
              ListaData.Update
              ListaData.close

            cont1 = cont1 + 1
        else
            cont2 = cont2 + 1
        end if

        rs_fato.close
        rs_sessao.movenext
loop

response.write ("<br>" & cont1 & " registros gravados <br>")
response.write ("<br>" & cont2 & " registros desprezados <br>")

rs_sessao.close
ConexaoBase.close

response.write("TERMINOU PROCESSAMENTO")
%>
```