CENTRO UNIVERSITÁRIO EURÍPIDES DE MARÍLIA
FUNDAÇÃO DE ENSINO "EURÍPIDES SOARES DA ROCHA"
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

MINERAÇÃO DE DADOS EM EMPRESAS DE TELECOMUNICAÇÕES

MAURO MACIEL MOREIRA CASTRO

Marília, 2013

CENTRO UNIVERSITÁRIO EURÍPIDES DE MARÍLIA
FUNDAÇÃO DE ENSINO "EURÍPIDES SOARES DA ROCHA"
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

MINERAÇÃO DE DADOS EM EMPRESAS DE TELECOMUNICAÇÕES

Monografia apresentada ao Centro Universitário Eurípides de Marília como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.
Orientador: Prof. Elvis Fusco

Marília, 2013

CENTRO UNIVERSITÁRIO EURÍPIDES DE MARÍLIA
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

TRABALHO DE CONCLUSÃO DE CURSO – AVALIAÇÃO FINAL

Mauro Maciel Moreira Castro

Implementação de Ambiente de Mineração de Dados.

Banca examinadora da monografia apresentada ao Curso de Bacharelado em Ciência da Computação do UNIVEM/F.E.E.S.R., para obtenção do Título de Bacharel em Ciência da Computação.

Nota: 8,5 (oito e meio)

Orientador: Elvis Fusco

1º. Examinador: Fabio Lucio Meira

2º. Examinador: Paulo Augusto Nardi

Marília, 03 de dezembro de 2013.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

Há muito conhecimento escondido na imensa quantidade de dados disponíveis nos bancos de dados das empresas, o que torna a extração de conhecimentos úteis para tomada de decisões algo lento e complicado. Observando as dificuldades encontradas em algumas empresas de telecomunicações de pequeno porte, este trabalho tem como objetivo a criação de um *Data Mart* para armazenar esta grande quantidade de informações e facilitar a geração de relatórios pelos usuários. Para a obtenção de conhecimentos foram utilizadas técnicas de Mineração de Dados, mais especificamente técnicas de classificação com a utilização de algoritmos de árvores de decisão, que transformam dados brutos em informações valiosas. O trabalho tem como cenário uma empresa real de telecomunicações, bem como sua base de dados.

Palavras-chave: Data Warehouse; Mineração de Dados; Bancos de Dados; Árvores de Decisão.

Abstract

There is much hidden knowledge in enormous amount of data available in the databases of companies, which makes the extraction of useful knowledge for decision making something slow and complicated. Watching the difficulties encountered in some small telecom companies, this work aims to create a Data Mart to store this large amount of information and to facilitate reports made by users. To obtain knowledge were used data mining techniques, more specifically classification techniques with the use of algorithms of decision trees that transform raw data into valuable information. The work has a real telecom business scenario, as well as its database.

Keywords: Data Warehouse, Data Mining, Databases, Decision Trees.

Introdução

Os sistemas de bancos de dados são componentes essenciais no cotidiano de uma empresa, pois durante um dia são inseridas e processadas uma grande quantidade de informações, quando executamos operações tais como: cadastro de um cliente, consulta dos fornecedores de uma empresa, geração de relatórios, reserva de materiais, notas, faturas de clientes, etc.

No passado, os dados e as tecnologias eram utilizados exclusivamente para decisões operacionais. Com o passar do tempo os desenvolvedores e clientes perceberam que apenas um banco de dados não era suficiente para processar essas transações e fazer processamento analítico ao mesmo tempo.

As informações obtidas dos ambientes operacionais (Bancos de Dados de Produção) são úteis apenas para manter o sistema em funcionamento. Ocorre que os funcionários responsáveis pela parte de tomada de decisões da empresa precisavam de informações específicas, tais como: quais regiões eles teriam que investir, qual linha de produtos estava obtendo melhor retorno financeiro ou quais mercados seriam vantajosos fortalecer (Ponniah, 2010).

Para o problema da necessidade de informações foram criados os *Data Warehouses*, desenvolvidos com base nos estudos do MIT (Instituto de Tecnologia de Massachusetts) realizados nos anos 70 que focavam o desenvolvimento de uma arquitetura técnica mais eficiente para sistemas de informações (Haisten, 1999).

Mesmo com os resultados satisfatórios do uso do *Data Warehouse*, muitos conhecimentos importantes ficavam escondidos devido à enorme quantidade de dados armazenados. Tornou-se indispensável a utilização de ferramentas ou métodos para análise de grandes quantidades de informações, ferramentas que fazem uso de *Data Mining* (Mineração de Dados) foram muito úteis nesses casos. Essas ferramentas selecionam o conteúdo das bases de dados e fazem uso de diversos métodos para obtenção de conhecimento.

Como proposta desta abordagem de estudo, será realizada a construção de um *Data Mart*, bem como meios de manipulá-lo e a implementação de um software responsável pela Mineração de Dados utilizando algoritmos de árvores de decisão, encontrados na ferramenta

Weka. O software desenvolvido tem como cenário os dados reais de uma empresa de telecomunicações. Servirá, portanto, para ajudar empresas deste ramo a obter informações que deem suporte às tomadas de decisão.

Quanto à motivação para o tema deste estudo, foi a de atender empresas de telecomunicações de pequeno porte que necessitam de informações que lhes deem suporte à decisão como: perfil de acesso do cliente, padrões de contração de planos, informações estas que são obtidas a partir da utilização de processos e ferramentas específicas, tais como: *Data Warehouse* e Mineração de Dados, para poderem se desenvolver.

No capítulo 1 é apresentado o embasamento teórico sobre *Data Warehouse*, os diferentes tipos de informação, as origens do DW, suas principais características e abordagens de desenvolvimento com suas vantagens e desvantagens.

Já no capítulo 2 será abordada a parte teórica da Mineração de Dados, conceitos e aplicações, suas principais técnicas com maior enfoque em árvores de decisão tendo em vista sua utilização no estudo de caso.

Com base nos conhecimentos descritos nos capítulos anteriores, no capítulo 3 será relatado o estudo de caso. Partindo da definição do cenário e das ferramentas utilizadas no processo de implementação, são descritos em seguida os passos necessários para o desenvolvimento e, por fim, os resultados obtidos.

As conclusões do trabalho realizado e sugestão para trabalhos futuros são apresentados após o capítulo 3.

Neste estudo são destacadas como principais restrições:

- O software implementado atende as necessidades dos clientes da empresa Alfa (empresas de telecomunicação) por ter sido utilizado apenas à base de dados e regras de negócio da mesma;
- Este trabalho visa estudar alguns dos principais algoritmos existentes atualmente na literatura mundial e testar alguns dos que estão presentes na ferramenta Weka em uma grande base de dados. Vale salientar que não foi criado ou implementado nenhum novo algoritmo.

Capítulo 1 – Data Warehouse

Neste capítulo será abordado o levantamento histórico dos sistemas de apoio a decisão, chegando até a criação dos *Data Warehouses*. Serão apresentadas suas características, arquitetura, abordagens de desenvolvimento e, por fim, será descrito o processo de extração, tratamento e carga (ETL).

1.1 Origens dos Sistemas de Apoio à Decisão

A tecnologia de dispositivos de armazenamento surgiu por volta de 1970 e, devido às dificuldades para armazenar informações e as limitações apresentadas nas antigas fitas de armazenamento, não era necessário um software para gerenciamento de dados. Com a criação dos discos de armazenamento foi possível gravar grandes quantidades de dados, surgindo também um novo tipo de software conhecido como Sistema Gerenciador de Banco de Dados (SGBD). Tal sistema permitia aos programadores maior facilidade, tanto para o armazenamento, quanto para o acesso a estes dados. Com esse novo sistema surgiram também os primeiros bancos de dados que foram definidos como “uma única fonte de dados para todo o processamento” (Wagner, 2003).

Em meados de 1975 foi desenvolvido o Processamento de Transações On-line (OLTP) de alto desempenho, permitindo aos usuários de computadores realizarem tarefas que, antes não eram viáveis, tais como: controlar sistemas de reservas, sistemas de caixas bancários, sistemas de controle de produção e estoque, entre outras.

Antigamente os dados e as tecnologias eram utilizados exclusivamente para decisões operacionais, ou seja, operações de inserção, atualização, remoção e consultas. Com o tempo perceberam que apenas um banco de dados não era suficiente para processar essas transações e fazer processamento analítico.

Durante a década de 80 surgiu um novo programa que era utilizado no processo de extração de dados. Tal programa consistia em percorrer as bases de dados, usar critérios de seleção para encontrar informações e transportá-las, então, para outros arquivos ou bancos de dados onde seriam analisados. Com isso não haveria uma sobrecarga do sistema principal, pois outro sistema ficaria responsável pela parte de geração de relatórios e gráficos para a empresa (Inmon, 2005).

Contudo, essa solução gerou diversos outros problemas como por exemplo,

divergência entre resultados, falta de integridade, diferentes algoritmos para as consultas e o grande espalhamento dos dados, pois eles eram extraídos para outras bases que poderiam ser consultadas e extraídas para outras bases (Inmon, 2005).

O crescimento das empresas, que ocorreu por volta de 1990, tornou-as complexas e por sua vez, as organizações se expandiram globalmente, acirrando ainda mais a competição pelo mercado, deixando os gerentes cada vez mais preocupados com a obtenção de informações que pudessem aprimorar sua linha de produção. As informações obtidas dos Bancos de Dados de Produção eram úteis apenas para manter o sistema em funcionamento. Ocorre que os gerentes responsáveis pela tomada de decisão precisavam de informações específicas tais como: quais regiões eles tinham que investir, qual linha de produtos estava obtendo melhor retorno ou quais mercados fortalecer (Ponniah, 2010).

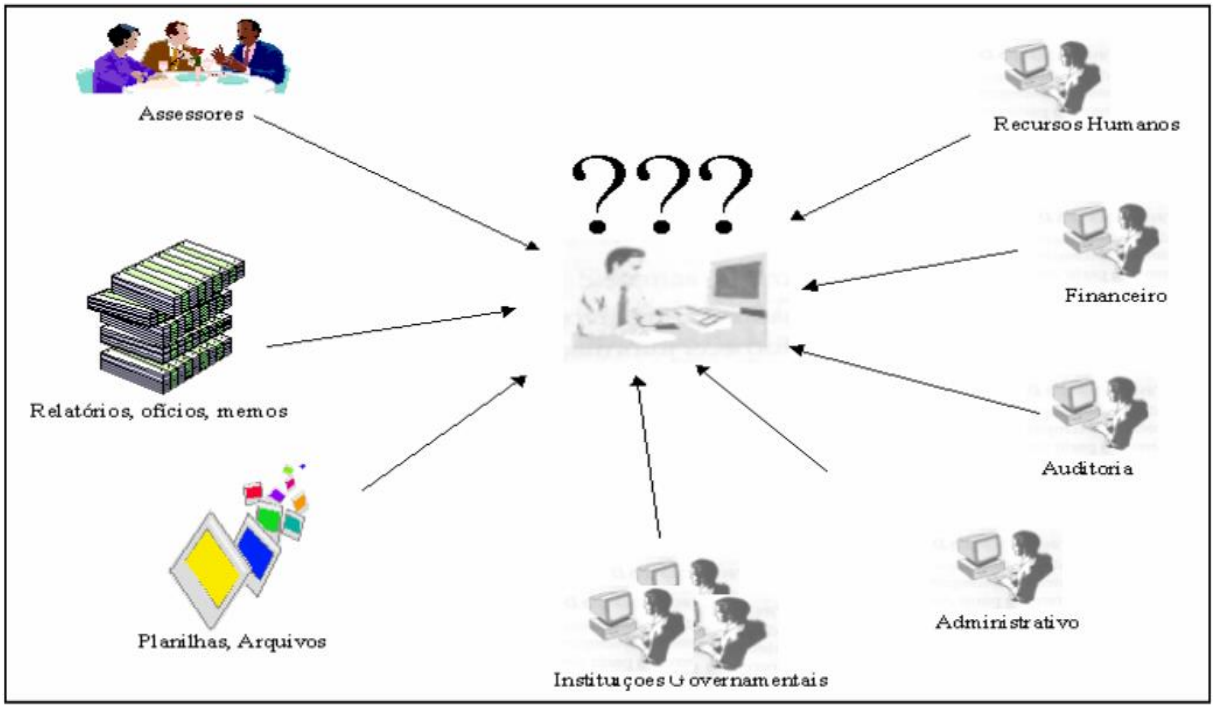


Figura 1 - Excesso de Informações (Clemes, 2001)

Também na década de 90 houve uma saturação de informações que complicaram ainda mais a análise e os processos de decisão, visto que os sistemas de informação armazenavam cada vez mais dados sem haver métodos para processá-los, conforme apresentado na Figura 1 (Clemes, 2001).

Com a globalização, as empresas que antes precisavam focar apenas em mercados de menor tamanho e que mudavam muito lentamente, depararam-se agora com um novo conceito de administração, onde várias companhias de grande porte têm de competir por novos clientes. Com isso, o processo de tomada de decisão necessitava de ferramentas

computacionais mais eficientes, que suportassem a grande quantidade de informações obtidas (Barbosa, 2003).

Brackett (apud Clemes, 2001) previa que a cada dois anos o volume de dados estava dobrando, o que significa que se as organizações não utilizarem métodos mais eficientes para o tratamento das informações obtidas, tornar-se-ão cada vez menos competitivas.

1.2 Informações Estratégicas

Para ter maior conhecimento do andamento da empresa, os gerentes necessitavam não só de informações para rever e monitorar indicadores de desempenho, como também observar a relação entre os mesmos, além de manterem-se informados de como fatores chaves variavam com o tempo. Portanto, executivos e gerentes careciam de informações que os aproximassem de seus clientes, de novas tecnologias, de resultados de vendas e de marketing.

Como se pode observar, todos estes tipos de informações essenciais possuem um fator em comum, que, combinados, recebem a denominação de informações estratégicas. Vejamos abaixo, algumas características citadas por Ponniah (2010):

- **Integrada:** deve ser única e abranger uma visão geral da empresa;
- **Integridade de dados:** a informação deve ser precisa e seguir padrões de acordo com as regras de negócio;
- **Acessível:** deve possuir formas intuitivas de acesso e compreensíveis para qualquer nível de usuário;
- **Confiável:** as informações perante o sistema como um todo devem possuir apenas um valor;
- **Pontual:** os resultados devem estar disponíveis dentro de um período de tempo aceitável.

Assim sendo, estas informações tornam-se importantes para o processo de tomada de decisões estratégicas da empresa, visto que com elas é possível para os gestores obterem diferentes pontos de vista do mercado.

1.3 Ferramentas para Informações Estratégicas

Em decorrência da necessidade por informações estratégicas, novas ferramentas

foram desenvolvidas, tais como as descritas por Ponniah (2010):

- **Relatórios Ad Hoc:** programas que eram desenvolvidos para a geração de relatórios de um problema solicitado pelos clientes;
- **Programas Especiais para Extração:** programas escritos para a extração de dados das várias aplicações que seriam úteis para possíveis pedidos dos clientes;
- **Pequenas Aplicações:** programas que utilizavam os dados extraídos para a impressão de relatórios com a possibilidade dos usuários passarem parâmetros na pesquisa. Outros programas mais elaborados permitiam também a visualização dos relatórios na tela;
- **Centros de Informação:** centros onde os usuários podiam solicitar relatórios ad hoc ou visualizar informações especiais nas telas;
- **Sistemas de Apoio à Decisão:** pequenos sistemas parecidos com as soluções anteriores responsáveis por apresentar informações estratégicas, com a possibilidade de visualizar e imprimir relatórios, porém, com uma melhor aparência pela utilização de menus. Foram muito utilizados pelo marketing;
- **Sistemas de Informação Executiva:** sistemas que possuíam facilidade de uso para que executivos pudessem obter informações estratégicas, através da possibilidade de solicitar relatórios especiais. Não duraram muito devido às limitações de possuírem telas pré-programadas e quantidade limitada de relatórios disponíveis.

Como se pode observar, há uma mudança de enfoque por parte das empresas, ou seja, a princípio as empresas buscavam apenas armazenar as informações e, agora, devido aos avanços tecnológicos nos Sistemas de Apoio à Decisão (SADs), houve uma maior valorização das informações. Logo, não basta apenas armazenar as informações é preciso também interpretá-las (Souza, 2002).

Com o desenvolvimento de novas ferramentas e a ampliação de sua área de abrangência, surgiram para os SADs: o *Data Warehouse*, OLAP e o *Data Mining*.

Maior concorrência no mercado e a busca de mais qualidade por parte dos clientes obrigaram as corporações compreender o comportamento dos seus clientes e o que desejam. Para isso é preciso que haja ferramentas eficientes para a solução de problemas e criatividade no desenvolvimento de novos produtos (Barbosa, 2003).

1.4 Origens do Data Warehouse

Segundo Haisten (1999), a origem do *Data Warehouse* vem dos estudos do MIT

(Instituto de Tecnologia de Massachusetts) nos anos 70, que focavam o desenvolvimento de uma arquitetura técnica mais eficiente para os sistemas de informações.

Outrora os dados e as tecnologias eram utilizados exclusivamente para decisões operacionais, ou seja, operações de inserção, atualização, remoção e consultas. Com o tempo percebeu-se que apenas um banco de dados não era suficiente para processar essas transações e fazer processamento analítico, bem como ser responsável pela parte de geração de relatórios e gráficos para a empresa (Inmon, 2005).

Como resultado, os sistemas de informação foram divididos em dois grupos distintos:

- **Sistemas de Informação Operacionais:** são responsáveis por dar suporte a processos rotineiros que consistem em executar regras de negócio e registrar informações;

- **Sistemas Informacionais:** consistem em sistemas de informação criados para dar suporte aos processos de tomada de decisão.

1.4.1 – Sistemas de Informação Operacional

São sistemas que possibilitam fornecer serviços mais rápidos e eficientes, pois a todo momento estão executando operações que envolvem leitura e gravação de dados. Esses sistemas aceitam uma grande quantidade de dados de entrada e de saída, para tanto, necessitam de um processamento veloz. Utilizam operações matemáticas menos exigentes, tais como: adição, subtração, multiplicação e divisão. Estes sistemas podem ser acessados por vários usuários ao mesmo tempo (Souza, 2002).

De acordo com Reeves (2009) esses sistemas são chamados de Sistemas de Processamento de Transações On-line (OLTP). São especificamente voltados para funções como: baixa de estoque, ordens de pagamento, cadastros em geral, consultas, etc. Portanto, tais sistemas têm como função primordial assegurar que as transações sejam executadas no menor tempo possível e sem erros.

Os Sistemas de Informação Operacional são de grande importância para as empresas e, geralmente, são os primeiros a serem implementados em um sistema. Conforme a empresa se desenvolve, estes sistemas são reescritos, aumentados e mantidos ao ponto em que, as maiores organizações mundiais não operam sem a utilização dos mesmos (Sarkis, 2001).

Conclui-se, então, que os sistemas de informação operacional são responsáveis por manter o bom funcionamento da empresa executando os processos de negócio básico do dia-

a-dia.

1.4.2 – Sistemas Informacionais

Os planejamentos de marketing, de criação e análise financeira requerem um suporte diferente do que os sistemas de informação operacional podiam oferecer. Em razão dessa necessidade os sistemas informacionais são utilizados por terem suas funções essenciais baseadas em conhecimento. Enquanto os sistemas de informação operacional focam em uma única área, os sistemas informacionais fazem uso de uma grande quantidade de dados operacionais relacionados (Sarkis, 2001).

Em síntese, os sistemas informacionais têm como principal utilidade fornecer informações estratégicas para a utilização dos executivos e gerentes em processos de tomada de decisão. Podemos citar exemplos como o *Data Warehouse*, repositórios, mineração de dados, entre outros.

1.4.3 – Diferenças entre Sistemas de Informação Operacional e *Data Warehouse*

De acordo com Kimball (2004), Inmon (1997) e Reeves (2009) na Tabela 1 é agrupada a síntese das características dos Sistemas de Informação Operacional e do *Data Warehouse* apresentadas pelos mesmos:

Tabela 1 – Comparativo entre Sistemas Operacionais e *Data Warehouse*

SISTEMAS OPERACIONAIS (OLTP)	***DATA WAREHOUSE***
Utilizado por funcionários	Utilizado pela alta administração, comunidade gerencial
Valores atuais e voláteis	Valores históricos e imutáveis
Modelo Entidade-Relacionamento	Modelo Dimensional
Orientado a aplicações	Orientado a assuntos
Uso balanceado com alguns picos de processamento	Totalmente utilizado ou ocioso
Dados mudam constantemente	Dados estáveis
Alto detalhe dos dados	Dados sumarizados ou detalhados
Padrão de uso previsível	Difícil de prever
Suporte a processos da empresa	Suporte a análises estratégicas e relatórios
Os dados são consultados e mantidos inalterados	Os dados são explorados de acordo com as necessidades
Os dados são validados após a entrada	Os dados são validados depois da sua ocorrência
Os dados são organizados de acordo com o escopo de um sistema	Os dados devem ser integrados com vários sistemas
Acessados um registro por vez	Acessados um conjunto de registros por vez

Dados são atualizados quando regras de negócio são executadas	Os dados são atualizados periodicamente, em processos programados
Não contempla redundância	É permitida a redundância
Pequena quantidade de dados utilizados por processamento	Grande quantidade de dados utilizados em um processamento
Atende necessidades cotidianas	Atende necessidades gerenciais
Pode ser utilizado por milhares de usuários	Utilizado por centenas de usuários
De megabytes a gigabytes de tamanho	De gigabytes a terabytes de tamanho
Índices são utilizados para localizar um registro	Índices são utilizados para otimizar consultas

1.5 – Características do *Data Warehouse*

Um *Data Warehouse* (DW) ou Armazém de Dados possui uma característica que o diferencia de um banco de dados convencional, que consiste na sua capacidade de gerar e armazenar informações estratégicas. Para aumentar o desempenho do sistema, os *Data Warehouses* são armazenados fisicamente em ambientes diferentes do sistema principal da empresa, logo, as pesadas consultas realizadas nos DWs, não interferem no funcionamento geral da organização.

Vemos que Inmon (1997) conceitua um *Data Warehouse* como um banco de dados que armazena informações geradas pelas operações da empresa, que são extraídas de uma fonte única ou múltipla, permitindo um enfoque dos dados históricos da organização, o que será muito útil no suporte à tomada de decisão. Por ser dinâmico, ele oferece informações para uma grande variedade de usuários, com isso é possível a observação de informações relevantes que antes eram independentes.

Conforme Gardner (apud Wagner, 2003): "*Data Warehouse* é um processo, e não um produto, para a montagem e administração de dados provenientes de várias fontes com o propósito de obter uma visão simples e detalhada de parte de todo negócio".

Segundo Poe (apud Souza, 2002):

> "*Data Warehouse* é um banco de dados analítico que é usado como base para um SAD (Sistema de Apoio a Decisão). É planejado para armazenar um grande volume de dados somente de leitura, provendo acesso intuitivo às informações que serão usadas na tomada de decisões."

1.5.1 – Orientado a Assuntos

Para Inmon (1997) com o intuito de fornecer informações estratégicas, os dados armazenados no DW são organizados de acordo com os principais assuntos de interesse da organização, que podem ser setores da empresa, ou até mesmo ações realizadas pelo negócio.

Nos sistemas operacionais, os conjuntos de dados são separados para o funcionamento individual de uma determinada aplicação. Enquanto em DWs, os dados são organizados separados por assuntos ou eventos do mundo real, de tal forma que todo este conjunto de dados tenha relação apenas com um assunto (Ponniah, 2010).

Em resumo, Assuntos no DW são o conjunto de informações sobre determinada área de uma empresa.

Exemplos:

- Qualidade = componentes com defeitos, trocas e devoluções;
- RH = folha de pagamento, número de funcionários;
- Vendas = nota fiscal, quantidade de produtos vendidos.

1.5.2 – Dados Integrados

Como os dados armazenados nos sistemas operacionais podem estar em formatos e padrões diferentes, os mesmos devem ser modificados antes de serem alocados no DW, permitindo com isso um único padrão para todo o sistema (Inmon, 1997).

Por exemplo, as convenções de nomes, valores de variáveis, tais como sexo masculino e feminino, bem como atributos físicos de dados como os tipos dos dados, são formalmente unificados e integrados nessa base única.

Para uma correta tomada de decisão é necessário que todos os dados relevantes de todos os aplicativos do sistema sejam colocados juntos, independentemente de quais sistemas de informação operacional, SGBDs ou formato de dados eles vierem. Estes dados podem vir até de sistemas externos. Para que tudo funcione corretamente é necessário que haja um tratamento e uma padronização entre as informações coletadas (Ponniah, 2010).

1.5.3 – Dados Não-Voláteis

Nos *Data Warehouses* são executadas somente dois tipos de operações. São elas:

operações de inclusão de novos registros e consulta aos registros existentes. Assim sendo, as informações armazenadas no mesmo, não correm o risco de sofrer alterações, o que garante a segurança e integridade dos dados históricos (Inmon, 1997).

De acordo com Wagner (2003), depois dos dados serem inseridos nos sistemas analíticos (*Data Warehouses*, bancos de dados históricos, repositórios, etc.) os usuários poderão apenas consultá-los, visto que os bancos de dados analíticos são conhecidos como "somente leitura". Com isso, esses sistemas armazenam uma gigantesca quantidade de informações, o que lhes confere uma de suas principais características.

Em síntese, os dados armazenados nos *Data Warehouses* não devem nunca sofrer alterações visto que são dados históricos.

1.5.4 – Dados Variáveis com o Tempo

Em sistemas de informação operacionais do negócio, as informações precisam ser as mais atuais possíveis. No entanto, como a função do DW é a de análise, ele necessita tanto das informações atuais, quanto de informações históricas para poder perceber algum tipo de tendência.

As mudanças nos dados devem ser armazenadas para que nos relatórios gerados possam ser visualizadas. Portanto, armazenar dados que variem com o tempo é útil para: permitir a análise do passado, relacionar as informações com o presente e poder prever informações do futuro (Ponniah, 2010).

Na opinião de Inmon (1997), geralmente os DWs armazenam informações pelo período que varia entre cinco e dez anos e nos sistemas operacionais os dados históricos são armazenados por, no máximo, noventa dias.

Com uma arquitetura que permite um melhor controle de informações referentes a tempo (dia, hora, mês, etc.), os DWs armazenam os dados históricos sem o risco de eventuais problemas de inconsistência. Todos os *Data Warehouses* possuem uma tabela-fato ou tabela de dimensão que registrarão exclusivamente informações referentes a tempo.

1.5.5 – Granularidade

Para Inmon (1997), como a maioria das consultas visa dados levemente resumidos que são compactos e de fácil acesso, quando os usuários realizarem consultas que exijam

maior nível de detalhes, esses dados também estarão disponíveis. Por isso um nível dual de granularidade seria a melhor solução para atender as exigências por parte dos usuários.

Os dados armazenados no DW possuem diferentes níveis de granularidade, pois, ao realizar uma consulta, o usuário pode necessitar apenas de informações resumidas e ir aprofundando aos poucos para obter mais detalhes. Diante disso, é necessário avaliar um bom nível de granularidade dos dados, porque quanto menor a granularidade, maior será o volume de informações armazenadas (Ponniah, 2010).

Segundo Wagner (2003) os níveis de granularidade podem ser divididos em dois níveis extremos:

a- Nível de granularidade muito alto: possibilita uma grande economia de espaço de armazenamento. Para isso são utilizados níveis de agregação e de sumarização. Como desvantagem há uma redução na capacidade de atender consultas mais detalhadas, pois não teremos acesso aos dados de baixo nível;

b- Nível de granularidade muito baixo: possibilita responder a praticamente qualquer consulta, uma vez que nenhuma informação foi sumarizada. Como desvantagem é necessário um maior espaço para armazenamento, o que em algumas empresas pode se tornar inviável.

Resumindo, a granularidade seria o nível de detalhes dentro do DW. Logo, quanto maior o nível de detalhes, maior será o esforço computacional para processar os dados, bem como armazená-los. Logo, um menor nível de granularidade permite uma melhor realização de Mineração de Dados (*Data Mining*). Maiores detalhes sobre Mineração de Dados serão descritos adiante.

1.5.6 – Modelo Dimensional

O modelo dimensional nada mais é do que uma técnica de projeto lógica que atende melhor as necessidades do *Data Warehouse*, pois é mais fácil de se consultar e analisar os dados. Esta facilidade ocorre por apresentar uma estrutura de dados padronizada e intuitiva, que permite o acesso às informações com alto desempenho (Kimball, 1998).

De acordo com Brito (2012), este modelo possui uma característica peculiar, ou seja, a de apresentar um dado assunto sobre diferentes aspectos que também são conhecidos por dimensões. Por exemplo:

- Tempo;

- Geografia;
- Produto;
- Quantidade;
- Cliente.

Tal modelo reduz consideravelmente a quantidade de tabelas, já que une diversas tabelas do sistema sobre um determinado assunto em uma única tabela, como por exemplo: rua, cidade e estado, poderão ser resumidas em uma tabela Geografia.

A modelagem dimensional, de acordo com Souza (2002), pode ser dividida em:

- Fatos: armazenam medidas numéricas do negócio denominados fatos. Exemplo: unidades vendidas, quantidade de atendimentos, valor de notas. Cada registro da tabela-fato representa um item, uma transação ou evento do negócio. Armazenam grande quantidade de dados, que podem variar entre gigabytes e terabytes. Dependem diretamente das combinações de atributos das tabelas de dimensão;
- Dimensões: armazenam textos referentes às dimensões de um negócio, são sempre valores constantes, por exemplo, a dimensão cliente apresentará: primeiro nome, último nome, sexo, profissão, entre outros. Armazenam pequena quantidade de dados e contêm dados descritivos do negócio. As tabelas-dimensão são simétricas em relação à tabela-fato. Cada chave primária da mesma corresponderá a uma chave estrangeira na tabela-fato, permitindo assim a ligação entre ambas (Kimball, 1998).

Sarkis (2001) completa afirmando que o modelo dimensional permite que as tabelas existentes possam ser alteradas localmente pela adição de novos registros na tabela, por exemplo: a regra de faturamento foi alterada, o que não implica na alteração do esquema do *Data Warehouse*. Com isso as ferramentas de relatório não precisarão ser reprogramadas.

Segundo Wagner (2003):

> "Cada eixo no espaço multidimensional corresponde a um campo ou coluna de uma tabela relacional, e cada ponto, um valor correspondente à interseção desses campos ou colunas. Assim, o valor para o campo "vendas", correspondente ao mês igual a "março" e "filial 8", é um ponto com coordenada [março, filial 8]. Neste caso, "mês" e "filial" são duas dimensões e "vendas", uma medida. Em teoria, quaisquer dados podem ser considerados multidimensionais. Entretanto, o termo normalmente se refere a dados que representam objetos ou eventos que podem ser descritos e, portanto, classificados, por dois ou mais de seus atributos."

Na Tabela 2 são listadas as características das tabelas Dimensão e Fato.

Tabela 2 - Características das tabelas de Dimensão e Fato (Kimball, 2004)

Tabela-Dimensão	Tabela-Fato
Chave sintética criada para identificar a dimensão	Chave primária concatenada com as chaves das dimensões
Grande quantidade de atributos	Pequena quantidade de atributos
A maioria dos atributos contém texto	A maioria dos atributos é do tipo numérico
Não normalizada	Não normalizada
Habilidade de aprofundar ou resumir as consultas (drill-down e roll-up)	Dependem do nível de detalhamento da tabela-dimensão
Múltiplas hierarquias	Não possui hierarquias
Poucos registros	Muitos registros
	Medidas totalmente ou parcialmente somáveis

1.6 – Data Mart

Segundo Sarkis (2001) os *Data Marts* seriam, em poucas palavras, um subconjunto do *Data Warehouse*, geralmente focados em um dado departamento de uma empresa, ou alguma regra de negócio. Em outras palavras, trata-se de uma coleção de dados e ferramentas com o objetivo de resolver um determinado problema empresarial. Devido aos altos custos de desenvolvimento de um *Data Warehouse*, os *Data Marts* são uma solução mais viável para as grandes companhias.

Para Souza (2002) os *Data Marts* são um *Data Warehouse* de pequeno porte usados para suprir as necessidades de informações estratégicas de um departamento empresarial. Os dados armazenados podem conter diferentes níveis de granularidade, assim como nos *Data Warehouses*. Sem o devido planejamento global antes do desenvolvimento, a criação de *Data Marts* podem gerar fragmentação nos dados dos diferentes departamentos, o que dificulta a utilização das informações de forma integrada.

De acordo com Wagner (2003) “muitas empresas iniciam o processo a partir de uma área especifica, normalmente carente de informação e cujo trabalho é relevante para os negócios da empresa. Criam os chamados *Data Marts* (um *Data Warehouse* Departamental)”.

Na Tabela 3, é apresentado um comparativo entre os conceitos de *Data Mart* e *Data Warehouse*.

Tabela 3 - Comparativo entre *Data Mart* e *Data Warehouse*. Fonte (Souza, 2002)

DATA MART	***DATA WAREHOUSE***
Menor custo e esforço para implementação inicial	Inclusão de requisitos de todas as funções de negócio
Aumento de performance a partir da experiência dos usuários	Definições de dados e regras de negócios consistentes
Controle do *data mart* pela própria área de negócio a qual atende	Redundância de dados minimizada

1.7 – Abordagens de Desenvolvimento

Nos itens a seguir serão descritas as principais características das abordagens *Top-Down* e *Bottom-Up*.

1.7.1 – *Top-Down*

Este tipo de abordagem parte de um *Data Warehouse* completo e gera a partir dele pequenos *Data Marts*. Em geral, os analistas preferem a abordagem top-down que propicia maior flexibilidade e a possibilidade de carregar os dados quando mudanças ocorrerem (Sarkis, 2001).

Na Figura 2 é apresentada a arquitetura *Top-Down*.

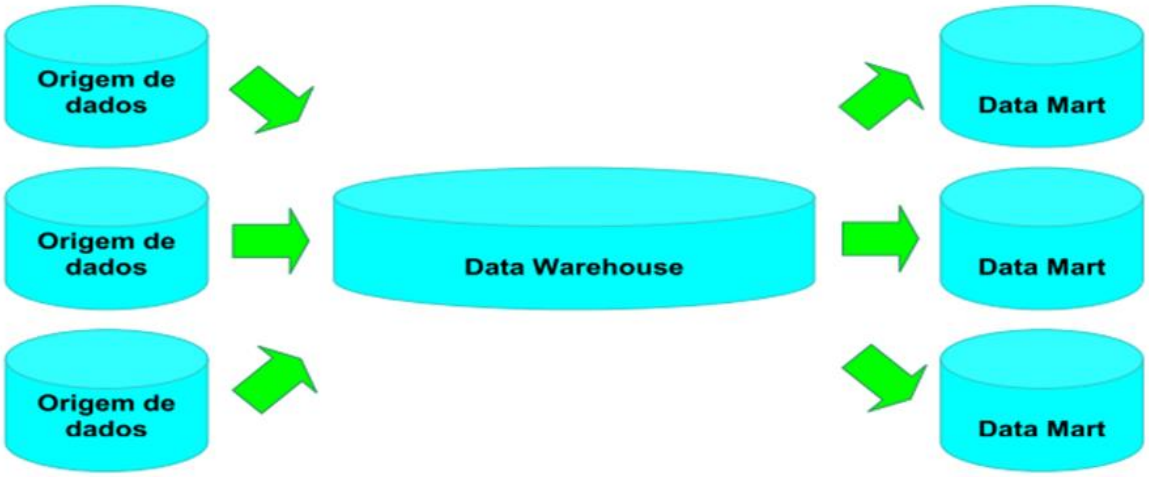


Figura 2 - Arquitetura Top-Down (Almeida, 2006)

De acordo com Souza (2002), o desenvolvimento top-down consiste em um *Data Warehouse* centralizado e com grande facilidade de acesso, e vários *Data Marts* dividindo um mesmo mecanismo de extração. Assim os dados contidos no DW também estarão presentes nos *Data Marts* permitindo o compartilhamento por todos os departamentos da empresa. Para que esta abordagem possa lograr êxito é necessário que toda a empresa participe definindo regras de negócio de forma corporativa.

Segundo Almeida (2006), nesta abordagem os dados encontram-se resumidos, dimensionados e disseminados para um ou mais *Data Marts* que, por sua vez, derivam todos os dados de um *Data Warehouse* centralizado.

As afirmações dos autores acima citados sobre vantagens e desvantagens desta abordagem, podem ser visualizadas na Tabela 4:

Tabela 4 - Vantagens e Desvantagens da abordagem Top-Down

Vantagens	**Desvantagens**
Fácil manutenção: os *data marts* utilizam a mesma arquitetura do *data warehouse* central.	Implementação muito longa e cara.
Arquitetura integrada e flexível.	Não existem garantias para retorno do investimento.
Única visão sobre a informação: o *data darehouse* consistente e padronizado é o ponto de partida de todos os *Data Marts*.	Necessita de uma equipe altamente treinada para o desenvolvimento.

1.7.2 – *Bottom-Up*

Nesta abordagem o desenvolvimento começa a partir de alguns *Data Marts* até chegar a um *Data Warehouse*, permitindo com isso flexibilidade, resultados rápidos e baixo custo inicial. Para empresas de pequeno porte é preferível este tipo de abordagem (Sarkis, 2001).

Este enfoque caracteriza-se por ser uma arquitetura de *Data Marts* independentes. Os dados são extraídos por mecanismos de extração desenvolvidos por cada departamento, o que permite a utilização de ferramentas e bases de dados distintas (Souza, 2002).

A arquitetura da abordagem *Bottom-up* é apresentada na Figura 3.

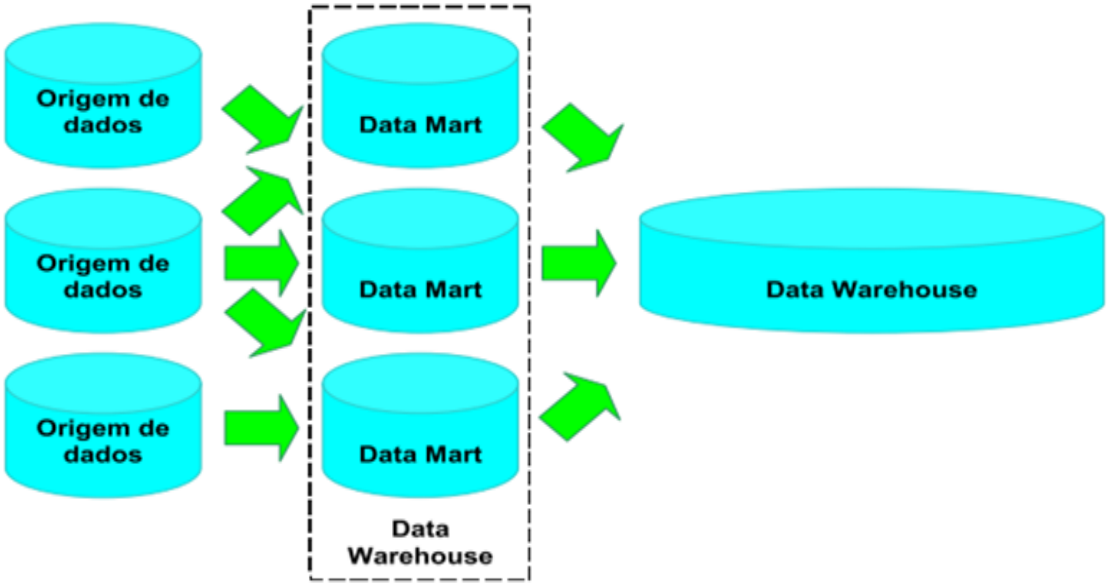


Figura 3 - Arquitetura *Bottom-Up* (Almeida, 2006).

Para Kimball (2004), com a utilização dessa abordagem o fluxo de desenvolvimento tem início com os *Data Marts* a partir de dados extraídos dos departamentos. Apesar da independência entre os dados, os *Data Marts* podem ser unidos para formar um *Data Warehouse* corporativo.

Na Tabela 5 são apresentadas as vantagens e desvantagens desta abordagem:

Tabela 5 - Vantagens e Desvantagens da arquitetura Bottom-Up (Souza, 2002)

Vantagens	**Desvantagens**
Implementação rápida (prazo de alguns meses) e enfoque em áreas essenciais ao problema.	Dificuldade de integração entre os *Data Marts*.
Retorno rápido, o que garante confiança e futuros investimentos.	Problemas de redundância e inconsistência entre os dados.
Facilidade na integração de novos *Data Marts* que forem sendo desenvolvidos.	Necessidade de coordenar múltiplas equipes de desenvolvimento de *Data Marts* em paralelo.
Menor infraestrutura inicial necessária para abrigar todo DW permitindo menores custos.	Esforços concentrados em um único *Data Mart* podem atrasar o desenvolvimento geral.

1.8 – Tipos de Arquitetura

Em Ponniah (2010), são citadas as diferentes arquiteturas de *Data Warehouse*:

- ***Data Warehouse* Centralizado:** Supre as necessidades de informação em nível

empresarial. Antes de seu desenvolvimento, são previamente calculados todos os requisitos necessários para sua infraestrutura. Os dados nele armazenados encontram-se na terceira forma normal. As consultas são realizadas diretamente no *Data Warehouse* central, sem intermediários. Nesse tipo de arquitetura, não são utilizados os *Data Marts*;

- ***Data Marts* Independentes:** Em empresas onde os departamentos possuem necessidades específicas, são criados os *Data Marts*. Cada um desses *Data Marts* separados, não fornecem uma visão geral da empresa. Como resultado é possível a existência de dados inconsistentes, visto que o uso de diversos padrões e diferentes definições de dados, prejudicam o relacionamento entre *Data Marts* de setores distintos;
- ***Hub-and-Spoke*:** Esta é uma abordagem sugerida por Inmon. Faz uso de um *Data Warehouse* centralizado e de *Data Marts* que dependem dos dados consultados do DW centralizado. Os dados do DW encontram-se inseridos na terceira forma normal, enquanto que nos *Data Marts* eles podem ser desnormalizados, resumidos, multidimensionais, entre outras formas, dependendo da necessidade da utilização. Geralmente as consultas são realizadas nos *Data Marts*, o que nada impede que sejam realizadas diretamente no *Data Warehouse* central. Para o desenvolvimento dessa arquitetura, é utilizada a abordagem *Top-Down*;
- ***Data Mart Bus*:** Esta abordagem foi criada por Kimball. Consiste na criação de um grande *Data Mart* (*supermart*) usando dimensões de negócios e métricas. Depois dele os próximos *Data Marts* criados compartilharão algumas das dimensões do supermart, obtendo com isso supermarts logicamente integrados, o que permitirá uma visão geral dos dados. Os dados armazenados são atômicos e utilizam o modelo de dados dimensional. Para o desenvolvimento dessa arquitetura, é utilizada a abordagem *Bottom-Up*.

1.9 – ETL

Considerado como parte fundamental para o desenvolvimento de um *Data Warehouse*: a correta extração dos dados de suas múltiplas fontes de origem e o devido tratamento dos dados obtidos, bem como a carga (inserção) dos mesmos, absorvem em alguns casos, até oitenta por cento de todo o processo de criação de um *Data Warehouse* (Inmon, 2005).

Para a realização desse processo faz-se necessária uma infraestrutura robusta o bastante para receber, processar e armazenar grandes quantidades de dados obtidos nas fontes. Vale lembrar que os dados presentes no *Data Warehouse* precisam ser de ótima qualidade, logo, a adequada realização desse processo garante o sucesso do desenvolvimento de um *Data*

Warehouse (Almeida, 2006).

Os processos que compõem o ETL (extração, transformação e carga) serão descritos em mais detalhes nos itens subsequentes.

1.9.1 – Extração

Conforme citado por Pereira (apud Souza, 2002):

> "A extração é o processo para obtenção dos dados existentes nos sistemas fontes para o ambiente do DW. Na maioria das vezes, esses dados provêm de várias fontes distintas e independentes. O processo pode ser conduzido através da construção de programas extratores executados sobre estes sistemas de modo a gerar os arquivos com os dados desejados. Outra opção é utilizar ferramentas de extração específicas, de forma a obter os dados necessários".

Segundo Kimball (2004), este processo costuma utilizar mais ou menos sessenta por cento das horas de desenvolvimento de um *Data Warehouse*. Este processo consiste em buscar informações relevantes em bases de dados da empresa, ou externos, que sejam úteis de acordo com modelagem do *Data Warehouse*.

Para a extração de dados de diversas fontes, é necessária uma técnica adequada para cada uma das fontes, tornando assim mais complicada a realização do processo. Os dados podem estar contidos em máquinas com diferentes arquiteturas, bancos de dados podem estar em formatos diferentes, dados podem estar presentes em outras redes ou até mesmo armazenados em planilhas, arquivos de texto, entre outros (Ponniah, 2010).

Os principais problemas, citados por Kimball (2004), encontrados na busca de dados são:

- Distribuição das origens dos dados em diversas plataformas, o que podem gerar a necessidade de utilizar vários produtos de terceiros, tais como: *plug-ins*, *odbcs*, etc;
- Dados que apresentem diferentes versões em um mesmo sistema;
- Um mesmo dado é representado de diferentes formas;
- Mudança de esquema dos bancos de dados de origem não documentada;
- Dados inválidos ou que sua representação não tem significado para os usuários.

1.9.2 – Transformação

Neste processo serão aplicadas regras de limpeza e transformações sobre os dados. Tal processo garante a qualidade dos dados que serão inseridos no *Data Warehouse*. Após a realização deste processo, os dados apresentarão as seguintes características apresentadas por (Kimball, 2004):

- **Unicidade dos dados:** evita a duplicação de informações;
- **Precisão dos dados:** os dados apesar de estarem em um formato diferente, ainda contêm suas características originais;
- **Integração dos dados:** visto que os dados parciais comprometem a realização de análises;
- **Consistência:** os dados são consistentes com as dimensões que os compõem.

Conforme descrito em Ponniah (2010), a transformação de dados é sempre muito importante para a implementação de um *Data Warehouse*. Ela consiste em realizar conversões a partir de dados de sistemas anteriores. As conversões podem ser divididas em pequenas tarefas individuais tais como:

- Limpar os dados extraídos de cada fonte;
- Correção de erros de ortografia;
- Resolução de conflitos, CEPs ou endereços com valores incorretos;
- Fornecimento de valores padrões quando houver falta de informações;
- Eliminação de duplicatas, quando o mesmo dado vier várias vezes de diversas fontes;
- Padronização semântica, utilizada para resolver problemas com sinônimos e homônimos;
- Aplicação de regras de cálculos para gerar novos valores, como a sumarização de dados ou realizar agregações.

Almeida (2006) ressalta que alguns problemas podem surgir durante o processo de transformação dos dados, tais como: ambiguidade de dados, ou seja, uma mesma palavra possui vários significados, ou em alguns outros casos, várias palavras apresentam um mesmo significado; integridade referencial dos dados, quando alguns fatos importantes não são levados em conta no processo de extração ou, são inexistentes na origem; bases com conjuntos de caracteres distintos podem apresentar problemas ao serem representadas na base de destino.

1.9.3 – Carga

A carga é o último passo do processo de ETL e tem como função carregar os dados já transformados, que podem estar armazenados em uma área de estagiamento, para o *Data Warehouse* de destino. Para Kimball (2004) é imperativo que alguns cuidados sejam levados em consideração antes da realização dessa etapa, como por exemplo:

- Realizar esse processo em paralelo e, se possível, quando houver pouca ou nenhuma utilização do *Data Warehouse*;
- Tabelas de Dimensão devem ser carregadas antes das tabelas de Fato;
- Deve ser analisada a integridade das chaves artificiais das tabelas de Dimensão;
- Verificar se a presença de valores nulos não está invalidando registros históricos dos fatos;
- Realizar uma carga incremental dos valores para assim não sobrecarregar o sistema, por exemplo, carregar apenas dados novos ou alterados.

Depois de concluída toda a modelagem do *Data Warehouse* é necessária uma carga inicial dos dados que, por sua vez, movimentará um grande volume de informações, o que acarretará grande quantidade de tempo.

Quando o *Data Warehouse* já estiver em funcionamento, o carregamento dos dados deverá ser realizado no menor tempo possível logo, se faz necessário o carregamento dos dados que sofreram algum tipo de alteração ou foram inseridos recentemente (Ponniah, 2010).

No capítulo a seguir serão abordadas algumas das principais técnicas de Mineração de Dados, bem como a sua origem.

Capítulo 2 – Mineração de Dados

Este capítulo trata da origem da Mineração de Dados, seus conceitos e sua aplicação. São descritas as técnicas de Mineração de Dados existentes e alguns dos algoritmos utilizados.

2.1 – Origem da Mineração de Dados

Claro está que estamos vivendo a "era da informação", cujo desafio é a implementação de novas técnicas que possuam a capacidade de medir e descobrir, de forma eficiente e clara, padrões relevantes presentes na grande quantidade de informações processadas. A criação dessas técnicas e a melhoria nos processos operacionais determinará a sobrevivência ou não de uma organização (Barbosa, 2003).

Segundo Beckmann (2010), tem-se tornado cada vez mais difícil o entendimento do grande volume de dados gerados pelos atuais processos automatizados. Na tentativa de sanar as dificuldades encontradas, as empresas têm procurado utilizar ferramentas para as tarefas de descoberta de conhecimento, identificação de padrões, atividades de classificação, entre outras.

Com a maior utilização do computador e da Internet a capacidade de produzir informações tem crescido de forma gradativa. De maneira geral, as empresas mesmo estando presentes neste cenário, não fazem uso de recursos computacionais para a transformação dos seus dados brutos em informações úteis, que permitam aos executivos a tomada de decisões (Bartolomeu, 2002).

2.2 – Conceitos

Mineração de dados também recebe os nomes de: garimpagem de dados, extração de conhecimento, arqueologia de dados, análise de dados, *Data Mining*, etc.

Estatísticos, pesquisadores de inteligência artificial, DBAs (administradores de banco de dados), pessoal de marketing, utilizam o termo Mineração de Dados para se referir a seleção de dados para o apoio de hipóteses em particular. A Mineração de Dados também pode ser conceituada como uma técnica para a detecção automática de tendências e associações presentes nos dados (Moraes, 2003).

Segundo Harrison (apud Moraes, 2003), a Mineração de Dados consiste em descobrir modelos e regras presentes nas grandes quantidades de dados por meios automáticos ou semiautomáticos, utilizando técnicas presentes na estatística, na ciência da computação e na inteligência artificial.

Conforme Martins (apud Bartolomeu, 2002), a Mineração de Dados é um processo não trivial de extração de informações previamente desconhecidas, porém, relevantes.

Utilizando métodos estatísticos e de inteligência artificial é possível a descoberta de relações e padrões entre os dados. Em outras palavras, Mineração de Dados significa extrair de grandes bases de dados informações importantes que podem ser utilizadas na tomada de decisões.

Para Pilot (apud Bartolomeu, 2002), ferramentas de mineração de dados servem também para pesquisas dos utilizadores, que geralmente tomam muito tempo para serem respondidas ou, que até mesmo as respostas encontradas surpreendam as expectativas dos especialistas.

2.3 – Aplicações

A Mineração de Dados por possuir uma vasta área de aplicação e necessitar de diversas técnicas para chegar ao conhecimento, entre elas podemos destacar: associação, classificação, agrupamento, previsão de séries temporais, entre outros.

A seguir são descritos alguns exemplos de aplicação apresentados por Halmenschlager (2002) e Thearling (2010):

- **Análise de riscos:** com a utilização da técnica de agrupamento é possível, por exemplo, uma empresa detectar os bons e os maus pagadores, com isso poderão ser evitados riscos de inadimplência;
- **Marketing direto:** a área de marketing é uma das que mais utiliza Mineração de Dados para tentar obter o perfil dos seus clientes e, através disso, conseguir enviar propostas direcionadas a clientes com maior probabilidade de adquirir seus produtos;
- **Segmentação de mercado:** através da análise de compras de clientes, pode-se gerar associações entre os produtos, fazendo com que os clientes que compram um tipo de produto, passem a se interessar pela compra de outros tipos, também;
- **Análise de mercado:** identificando a relação entre produtos ou serviços que são adquiridos juntos, em uma mesma transação, pode-se desenvolver melhores estratégias de

disposição dos produtos e promoção dos mesmos;

- **Manufatura:** realizando previsão de vendas, torna-se possível determinar níveis confiáveis de estoque e programas de distribuição de mercadorias;
- Uma empresa farmacêutica pode analisar a sua recente força de vendas, bem como seus resultados, para melhor direcionar seu investimento em pesquisas de medicamentos e determinar quais atividades de marketing terão maior impacto nos próximos meses;
- Uma empresa de cartão de crédito pode aproveitar seu grande armazém de dados, para identificar quais os clientes com maior probabilidade de adquirir novos produtos;
- Uma grande empresa de bens de consumo, pode aplicar mineração de dados visando melhorar o seu processo de vendas para os varejistas, como por exemplo: selecionar estratégias promocionais que poderão melhor alcançar seus clientes.

A Tabela 6 apresenta a evolução na utilização da Mineração de Dados:

Tabela 6 - Estágios de Evolução da Mineração de Dados (Pilot apud Bartolomeu, 2002)

Estágio Evolutivo	**Perguntas de Negócio**	**Tecnologias Utilizadas**	**Provedores de Produtos**	**Características**
Coleção de Dados (1960)	"Qual foi minha renda total nos últimos cinco anos?"	Computadores, gravadores, discos.	IBM, CDC	Retrospectiva, distribuição estatística dos dados.
Acesso aos Dados (1980)	"Qual foi a unidade de venda na Inglaterra em março passado?"	Banco de Dados Relacional, Linguagem de Consulta Estruturada (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Restrospectiva, apresentação de dados dinâmicos no nível de registros.
***Data Warehousing* e Suporte a Decisão (1990)**	"Qual foi a unidade de venda na Inglaterra em março passado? Drill down para Londres."	Processamento analítico on-line (OLAP), banco de dados multi-dimensionais, *data warehouses*	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospectiva, apresentação de dados dinâmicos em múltiplos níveis.
***Data Mining* (Dias Atuais)**	"Qual será a mais provável unidade de venda em Londres no próximo mês? Por quê?	Algoritmos avançados multipro-cessadores, banco de dados massivos	Pilot, Lockheed, IBM, SGI, novos grupos de desenvolvimento (indústria recente)	Apresentação de informações pró-ativas.

2.4 – Técnicas de Mineração de Dados

As tarefas de mineração de dados (técnicas), segundo Jesus (2004), são divididas em dois principais grupos:

- **Descoberta direta de conhecimento:** este tipo de mineração parte de um

objetivo pré-determinado, ou seja, o usuário seleciona um campo em questão e solicita ao sistema uma estimativa, classificação ou previsão do mesmo;

- **Descoberta indireta de conhecimento:** como não há alvo, podemos perguntar ao sistema quais padrões são relevantes para o entendimento dos dados. Deste modo, a técnica de mineração será responsável pela descoberta de padrões que possam ser úteis.

A seguir, serão descritas, resumidamente, as principais tarefas de Mineração de Dados:

- **Classificação:** consiste em estudar características de um objeto ou situação e atribuir-lhe uma classe pré-definida;
- **Estimação:** a estimação trabalha com valores numéricos contínuos. A partir de algumas variáveis conhecidas, podemos utilizar a estimação para a obtenção de um valor para algumas variáveis desconhecidas;
- **Previsão:** o que difere esta técnica das técnicas de classificação ou estimação é o fato de classificar registros previstos ou estimados de alguma atividade futura. Em outras palavras, consiste em determinar o futuro de alguma grandeza. Também podem ser ainda aproveitadas as técnicas utilizadas na estimação ou classificação;
- **Agrupamento por Afinidade:** consiste em gerar regras de associação a partir de dados. Por exemplo, quem compra leite adquire também pão com a probabilidade de 40%;
- **Segmentação:** é a técnica que realiza agrupamentos de populações heterogêneas, em vários subgrupos mais homogêneos. A técnica de segmentação difere-se da classificação por não utilizar classes pré-definidas. É muito útil quando não se tem conhecimento sobre os dados a serem analisados;
- **Descrição:** tem como propósito descrever o que ocorre numa base de dados facilitando, assim, o entendimento sobre variáveis relevantes do sistema, como: produtos, clientes, etc.

A Tabela 7 apresenta as técnicas mais adequadas para cada tarefa de Mineração de Dados.

Tabela 7 - Técnicas mais adequadas para cada tarefa (Harrison apud Bartolomeu, 2002)

	Classificação	*Estimativa*	*Previsão*	*Agrupamento por afinidade*	*Segmentação*	*Descrição*
Estatística padrão	✓	✓	✓	✓	✓	✓
Análise de seleção estatística			✓	✓	✓	✓
Raciocínio baseado em casos	✓		✓	✓	✓	
Algoritmos genéticos	✓		✓			
Detecção de agrupamentos					✓	
Análise de vínculos	✓		✓	✓		
Árvores de decisão	✓		✓		✓	✓
Redes neurais artificiais	✓	✓	✓		✓	

A seguir serão descritas com mais detalhes as técnicas de mineração, com maior enfoque na técnica que foi utilizada no desenvolvimento deste projeto.

2.4.1 – Técnicas de Associação

As técnicas de Associação possuem como objetivo, conforme apresentado em Garcia (2003) “o método de associação busca estabelecer relacionamentos entre um conjunto de dados, a fim de encontrar afinidades entre eles. A partir de uma transação, as regras de associação tentam encontrar itens que envolvem a presença de outros itens”.

Como, por exemplo, uma base de dados de vendas de um mercado, onde cada transação possui um ou mais itens adquiridos pelos fregueses. Podem-se extrair algumas regras de associação que mostram que 30% das transações que contêm cerveja, também contêm fraldas descartáveis, sendo que 2% do total de transações contêm ambos os itens. As regras geradas pelo método de associação apresentam o formato “Se X, então Y”, em que X e Y são conjuntos de itens. A avaliação destas regras é obtida, conforme Garcia (2003), através dos parâmetros de: fator de suporte e fator de confiança.

a) **Fator de suporte:** indica a ocorrência desta regra obtida dentro do total de transações. No caso 2% para o exemplo supracitado;

b) **Fator de confiança:** é o grau no qual a regra é verdadeira ao considerar os registros de forma individual. No caso 2% de todas as vendas continham cerveja, e desse valor 30% foram de vendas de cerveja e fraldas.

Segundo Ninho (2011) “o problema na mineração por regras de associação está em

gerar todas as regras que sejam realmente úteis. Regras úteis são regras que ocorrem com frequência, são confiáveis e fazem predições interessantes.” Outro problema encontrado está na grande carga de processamento exigida, para isso faz-se necessária a pesquisa de novas técnicas e algoritmos eficientes.

2.4.2 – Técnicas de Classificação

A tarefa mais comum de Mineração de Dados é a Classificação, que consiste na localização de propriedades comuns entre um conjunto de dados e, posteriormente, os classifica em diferentes classes pré-definidas (Bartolomeu, 2002).

As regras geradas possuem a estrutura de “Se <condição> então <conclusão>”. Onde a condição, premissa, corpo da regra ou complexo é uma condição de testes de atributos na forma de Xi {=, ≠, >, <, ≤, ≥} A, onde Xi é um atributo e A é um valor constante. Já a conclusão ou cabeça da regra, sendo constituída de uma classe Ci. (Halmenschlager, 2002)

Comumente as regras de classificação são usadas para representar conhecimentos em sistemas especialistas, podendo ser interpretadas por especialistas humanos sem maiores dificuldades. Estas regras também são modulares, ou seja, não dependem umas das outras.

Conforme citado em Beckmann (2010), os principais algoritmos de Classificação são:

- **Classificador Bayesiano:** utiliza a regra de Bayes para efetuar uma classificação supervisionada. Este algoritmo é indicado para conjuntos de dados que seguem uma distribuição normal, e classes que possam ser separadas completamente. Contudo, possui como desvantagem não considerar os casos com menores frequências, classificando estes de forma errônea. Esse problema ocorre geralmente com dados desbalanceados;
- **Vizinhos mais próximos:** este método não necessita de um modelo pré-definido, sua classificação consiste em calcular a distância entre uma nova instância e todas as instâncias do conjunto de treinamento;
- **Algoritmos genéticos:** em situações em que os problemas são muito complexos ou levam muito tempo para se calcular a solução exata, é necessária a utilização de heurística. A heurística presente nos algoritmos genéticos é utilizada para encontrar soluções mais próximas do ótimo, porque ela parte de uma abordagem probabilística e não determinística. Ou seja, ela tenta chegar a uma solução mais próxima de 100%. Os algoritmos genéticos apresentam como vantagem, a capacidade de realizar uma exploração simultânea,

não partindo apenas de um ponto, e sim de uma população, logo, trabalha muito bem de forma paralela. Como desvantagem, os algoritmos podem exigir um longo tempo de execução e não garantem a convergência para o ótimo global;

- **Árvores de decisão:** uma árvore de decisão permite a criação de regras de decisão, isto é, conjuntos de "se <condição>, então <conclusão>" os quais permitem aprender e então distinguir as classes dos registros de dados. Mais detalhes serão descritos à seguir.

2.4.2.1 - Árvores de Decisão

Segundo Ninho (2011), a árvore de decisão é uma maneira gráfica de visualizar um conjunto de condições necessárias para um determinado fim. Para a construção dessa árvore é imprescindível a realização de uma grande sequência de testes. Cada nó da árvore corresponde a um teste do valor de uma das propriedades, os ramos deste nó são identificados com os possíveis valores do teste. Os nós folhas correspondem às classes pré-definidas.

Árvores de Decisão apresentam o modelo de árvore, que é um tipo de estrutura de dados não linear e que possui um número finito de elementos ou nodos. Os nodos dessa árvore que não possuírem filhos serão chamados de folhas (classe pré-definida), enquanto que as raízes são os nomes dos atributos e, por fim, os galhos são os possíveis valores para estes atributos.

A profundidade de uma árvore é definida pela maior distância entre uma folha e a raiz, existindo árvores com profundidade uniforme em todas as folhas e outras não. Uma árvore é considerada binária quando possuir sempre dois filhos em cada nodo, ternária quando for três e mista quando o número de filhos for variável (Halmenschlager, 2002).

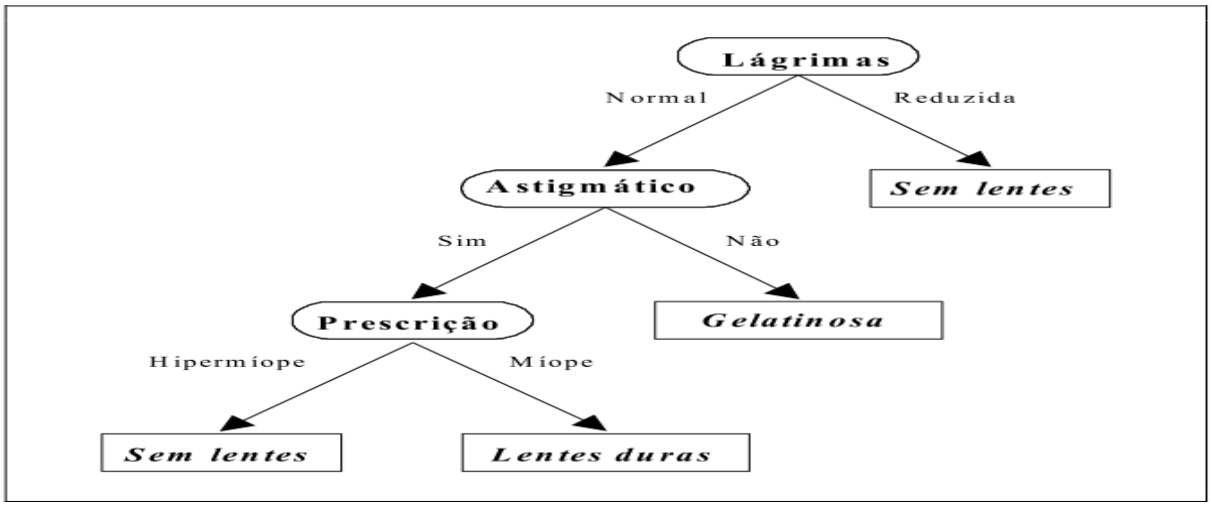


Figura 4 - Exemplo de Árvore de Decisão (Witten apud Halmenschlager, 2002)

Na Figura 4 é ilustrado um exemplo de uma árvore de decisão para a prescrição do uso, ou não, de lentes de contato. Nessa figura, cada elipse é um teste do valor de atributo e cada retângulo é uma classe pré-definida.

O modelo gerado pelo classificador de árvore de decisão pode ser compreendido e interpretado muito facilmente por humanos, por apresentar regras claras. O algoritmo permite a classificação de conjuntos de dados complexos, produzindo bons resultados, não somente com valores numéricos ou categóricos, mas também podendo ser combinado com outros métodos de Mineração de Dados (Beckmann, 2010).

O processo é recursivo e repetido em cada subconjunto de outros atributos, até quando a separação se tornar impraticável, ou quando uma classificação única pode ser aplicada a cada elemento do subconjunto derivado.

Na Figura 5 é ilustrado um algoritmo genérico de Árvores de Decisão.

Se todos os exemplos no atual conjunto de exemplos S satisfazem um critério de parada
então
cria um nodo folha com algum nome da classe e pára;
senão
seleciona um atributo A para ser utilizado como um atributo de particionamento e cria um nodo com o nome do atributo de particionamento;
escolhe um teste sobre os valores de A, com resultados mutualmente exclusivos e coletivamente exaustivos $R_1,...,R_k$, e cria um ramo, a partir do nodo recentemente criado, para cada teste;
particiona S nos subconjuntos $S_1,...S_k$, tal que cada S_i, i=1..,k, contenha todos os exemplos em S com resultado R_i do teste escolhido;
aplica este algoritmo recursivamente para cada subconjunto S_i, i=1,...k;
fim_senão
fim_se

Figura 5 - Algoritmo genérico para a construção de árvores de decisão (Garcia, 2003)

Em Garcia (2003) são apresentados três dos principais algoritmos de árvore de decisão:

- ID3: um dos primeiros algoritmos de árvores, criado por Ross Quinlan, no final dos anos 70, tendo como base sistemas de inferência e conceitos de aprendizagem já existentes na época. Este algoritmo foi desenvolvido visando a resolução de problemas que contenham atributos categóricos. Para sua utilização os valores devem estar tratados;
- CART: apresentado pelos estatísticos: Leo Breiman, Jerome Friedman, Richard

Oslen e Charles Stone, o algoritmo CART (Classificação e Árvores de Regressão) no trabalho intitulado “Classification and Regression Trees” publicado em 1984. Possui como principal característica a capacidade de gerar árvores de níveis reduzidos e poder trabalhar, tanto com valores categóricos como quantitativos;

- C4.5: este algoritmo é um aperfeiçoamento do algoritmo ID3, possibilitando o uso de atributos categóricos e quantitativos, bem como o uso de valores desconhecidos e, por último, adotando o sistema de poda.

2.4.3 – Técnicas de Agrupamento

Técnicas de Agrupamento (análise de Agrupamento), também conhecido por clusterização, está relacionado a diferentes algoritmos de classificação, todos focados em organizar os dados de acordo com as características comuns de cada um. Tyron (apud Ninho, 2011)

Ainda em Ninho (2011) são apresentados métodos para a formação dos clusters: “Para a formação dos clusters são utilizadas técnicas estatísticas multivariadas, com conotação exploratória, que verificam a similaridade dos objetos, através de coeficientes específicos para cada tipo de variável”.

Em Fayyad (1996), encontramos algumas funções para a técnica de Agrupamento:

- Utilizado para auxiliar na identificação de áreas similares de solo;
- Perfil de consumo;
- Níveis de renda;
- Catalogação de livros;
- Incidência de doenças, entre outras.

É muito utilizado também como uma ferramenta de pré-processamento para outros algoritmos, tais como: classificação e regras de associação.

As técnicas de agrupamento têm sido cada vez mais utilizadas em bases de dados geográficos, devido à grande quantidade de dados coletados. Por agrupamento é possível identificar regiões mais densas e esparsas, logo, é possível descobrir padrões de distribuição global e as correlações interessantes entre os atributos dos dados (Seixas, 2011).

Ao contrário do que ocorre na técnica de Classificação, onde os registros são atribuídos a classes pré-definidas, no Agrupamento os registros são segmentados de acordo com a semelhança que, a partir do seu resultado, poderemos determinar um significado. Por

exemplo: agrupando atributos de folhas, poderemos identificar diferentes espécies de vegetais (Bartolomeu, 2002).

Na Figura 6 visualiza-se um exemplo de agrupamento.

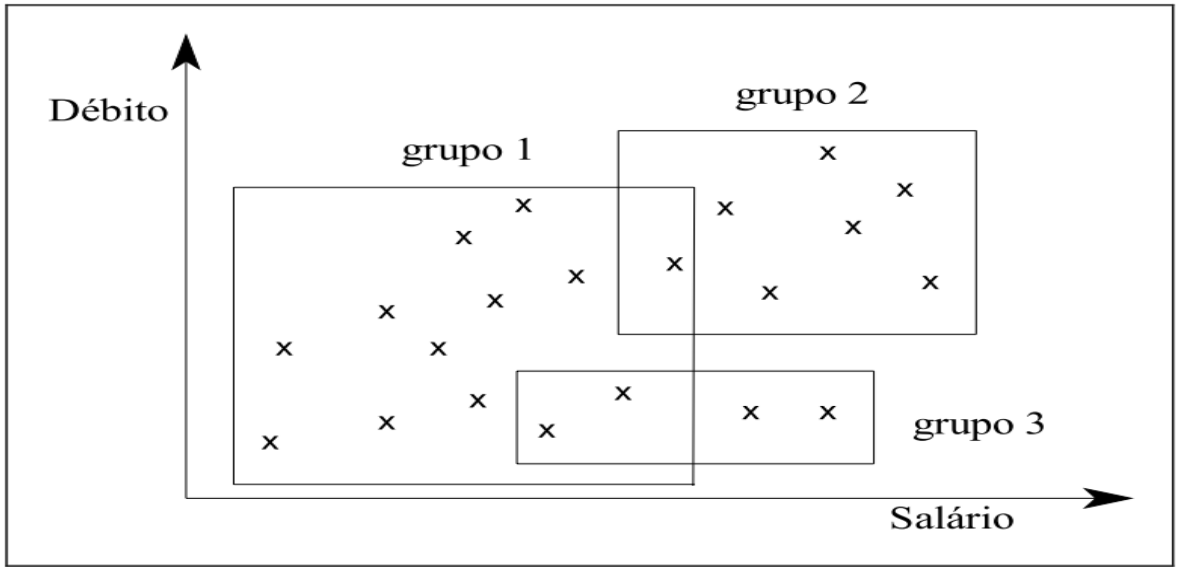


Figura 6 - Exemplo de Agrupamento (Halmenschlager, 2002)

São vários os métodos de agrupamento, sendo os três principais conforme descritos em Seixas (2011):

- **Métodos de particionamento:** Dado um conjunto de n objetos, estes métodos irão criar k ≤ n partições (grupos). Tendo em cada um destes grupos ao menos um objeto e cada objeto deverá pertencer a apenas um grupo. O algoritmo k-means é um dos algoritmos que implementa o método de particionamento;
- **Métodos hierárquicos:** Decompõem de forma hierárquica um conjunto de dados. Baseados em como é realizada esta decomposição, os métodos podem ser divididos entre: divisivos (a partir das folhas até a raiz) ou aglomerativos (a partir da raiz até as folhas). A decomposição hierárquica é representada por um dendrograma. Uma falha encontrada nestes métodos é que, depois de realizada uma fusão ou divisão dos dados, não é possível realizar ajustes;
- **Métodos baseados em densidade:** enquanto a maioria dos métodos de particionamento utilizam como base a distância entre os objetos, os métodos baseados em densidade, utilizam o limite definido de objetos presentes na vizinhança de um determinado grupo em crescimento. Dois dos métodos mais conhecidos são o DBSCAN e o OPTICS.

Aqui conclui-se a abordagem sobre a Mineração de Dados, bem como a revisão bibliográfica, cujas metodologias e técnicas serão de grande utilidade para o desenvolvimento deste projeto, apresentado no capítulo a seguir.

Capítulo 3 – Estudo de Caso

Neste capítulo será apresentada a aplicação da Mineração de Dados, em uma base de dados empresarial. As variáveis consideradas indispensáveis para a realização do estudo, bem como as ações adotadas para a obtenção dos dados utilizados, também serão apresentadas. Serão descritos também a forma de consulta dos dados (OLAP) e, finalmente, a descrição das etapas necessárias para a mineração dos dados.

O estudo de caso proposto tem como objetivo desenvolver um sistema que possibilite a uma empresa de telecomunicação, analisar os dados armazenados e identificar o perfil de seus assinantes. Por exemplo, os horários que um assinante mais consome banda, a relação entre sua profissão e o plano contratado, sua idade e gênero, entre outros.

Para que isso seja possível, foi necessária a criação de um *Data Mart* e a implementação de um programa responsável pela mineração dos dados.

3.1 – Descrição do Cenário

Em uma determinada empresa de desenvolvimento de software de controle de rede e monitoramento de Internet, que chamaremos de Alfa, são vendidos programas para empresas de telecomunicações. A Alfa possui uma regra de negócio que se denomina Consumo de Banda, esta regra consiste em, a cada período de tempo, enviar comandos SNMP[1] com OIDs[2] para consultar a quantidade de Bytes baixados (download) e subidos (upload) pelos equipamentos de seus clientes. Esta regra também é responsável por armazenar os valores obtidos em uma tabela do banco de dados, através de uma procedure que vai encontrar o registro desse assinante, bem como calcular as velocidades média de download e upload.

A base de dados real utilizada refere-se às informações de cadastros dos clientes, contratação de planos, equipamentos da empresa e consumo de banda, onde são relacionados

[1] *Simple Network Management Protocol*, ou Protocolo Simples de Gerenciamento de Rede, é um protocolo que permite a gerência de dispositivos através da rede. O "simples" do nome se refere ao fato do protocolo ser leve e fácil de ser implementando nos mais diversos dispositivos (RNP, 2004).

[2] Uma tabela MIB define "índices", chamados de OIDs, e conteúdos. Exemplo: o OID .1.3.6.1.2.1.1.4.0 contém como valor uma string com o contato técnico responsável pelo agente SNMP. Os OIDs são organizados em forma de árvore, e cada ramo da árvore pode receber um nome. (RNP, 2004)

os clientes com os equipamentos e horário de consulta.

Na Figura 7 é apresentado o script da tabela de consumo de banda.

```
CREATE TABLE consumobanda
(
  id bigint NOT NULL,
  equipamentoid bigint NOT NULL,
  download bigint NOT NULL,
  upload bigint NOT NULL,
  datacoleta date NOT NULL,
  horacoleta character varying(8) NOT NULL,
  macequipamento character varying(17) NOT NULL,
  intervalo integer NOT NULL,
  assinanteid bigint NOT NULL,
  downloadcoletado bigint NOT NULL,
  uploadcoletado bigint NOT NULL,
  planoid bigint NOT NULL,
  tipodownload integer NOT NULL,
  tipoupload integer NOT NULL,
  version integer NOT NULL,
  CONSTRAINT consumobanda_pkey PRIMARY KEY (id),
  CONSTRAINT consumo_fk_assinante FOREIGN KEY (assinanteid)
      REFERENCES assinante (id) MATCH SIMPLE
      ON UPDATE NO ACTION ON DELETE NO ACTION,
  CONSTRAINT consumo_fk_equipamento FOREIGN KEY (equipamentoid)
      REFERENCES equipamento (id) MATCH SIMPLE
      ON UPDATE NO ACTION ON DELETE NO ACTION,
  CONSTRAINT consumo_fk_plano FOREIGN KEY (planoid)
      REFERENCES plano (id) MATCH SIMPLE
      ON UPDATE NO ACTION ON DELETE NO ACTION
)
```

Figura 7 – Script SQL da tabela de Consumo de Banda

Com estes valores armazenados é possível:

- Verificar a quantidade de download ou upload excedentes de um determinado cliente e, a partir daí, limitar sua velocidade de tráfego;
- Verificar quais equipamentos apresentaram problemas na execução da regra;
- Criar gráficos para representar a velocidade, tanto de download como upload, dos clientes;

- Aplicar mineração de dados para obter conhecimentos sobre os hábitos de acesso dos clientes.

A Figura 8 mostra a representação do diagrama de entidade-relacionamento do banco de dados da empresa Alfa.

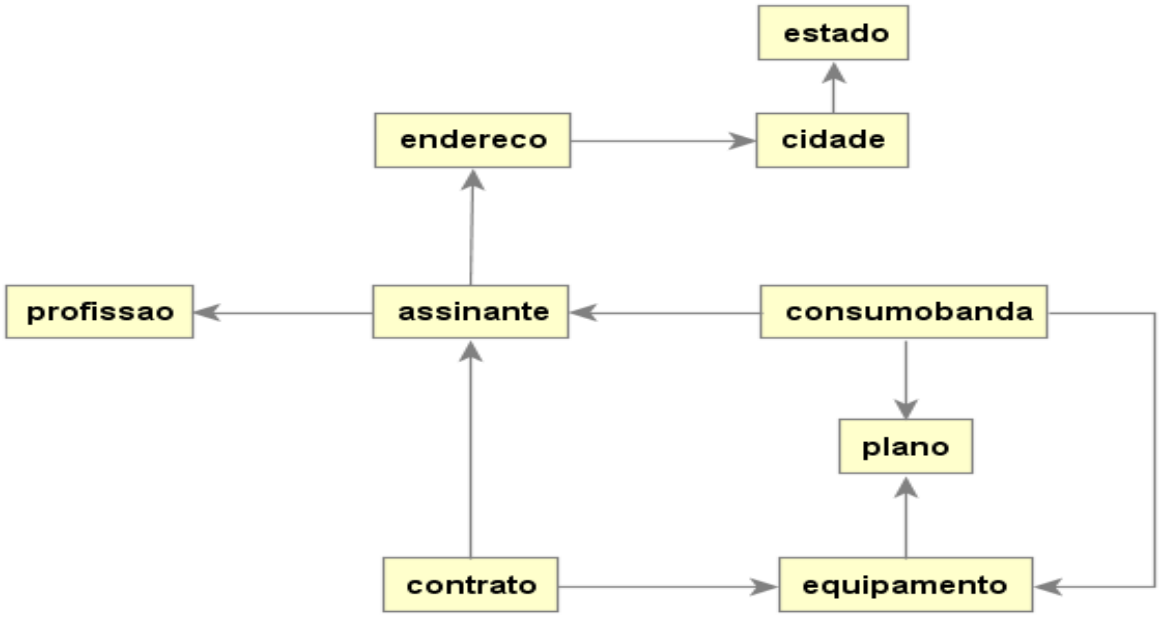


Figura 8 - Diagrama de entidade-relacionamento do banco de dados da empresa Alfa

É preciso lembrar que esta regra gera uma quantidade excessiva de dados, pois ela pode ser configurada para ser executada em diferentes intervalos de tempo, ou seja, quanto menor o intervalo maior a quantidade de registros armazenados.

Com a utilização de métodos convencionais, a obtenção de informações relevantes não é tão eficiente. Portanto, a criação de um *Data Warehouse* bem como a utilização de técnicas de Mineração de Dados se torna necessária.

No estudo de caso em questão, uma empresa de telecomunicação, que utiliza o sistema da empresa Alfa, possui cadastrados 11.765 assinantes, sendo 4.618 assinantes com algum plano de Internet. A quantidade de 20.243 equipamentos cadastrados, sendo destes 11.878 equipamentos ativos. A tabela de Consumo de Banda possui ao todo 2.097.393 registros.

Na Figura 9 é apresentado um exemplo de consulta dos registros da tabela de consumo de banda de um determinado assinante, nela podemos ver a variação da velocidade

de download e upload de acordo com a quantidade coletada no equipamento em um determinado horário.

equipamentoid bigint	download bigint	upload bigint	datacoleta date	horacoleta character varying(8)	macequipamento character varying(17)	intervalo integer	assinanteid bigint	downloadcoletado bigint	uploadcoletado bigint	planoid bigint
13114	0	3	2013-02-23	04:00:00	00:1E:6B:22:B4:2E	1200	17083	191355870	2423726673	139
13114	0	3	2013-02-23	04:20:00	00:1E:6B:22:B4:2E	1200	17083	191358171	2424297831	139
13114	0	3	2013-02-23	04:40:00	00:1E:6B:22:B4:2E	1200	17083	191360450	2424897150	139
13114	0	4	2013-02-23	05:00:00	00:1E:6B:22:B4:2E	1200	17083	191362785	2425506403	139
13114	0	3	2013-02-23	05:20:00	00:1E:6B:22:B4:2E	1200	17083	191365669	2426068195	139
13114	0	3	2013-02-23	05:40:00	00:1E:6B:22:B4:2E	1200	17083	191367970	2426566249	139
13114	0	3	2013-02-23	06:00:00	00:1E:6B:22:B4:2E	1200	17083	191370215	2427071767	139
13114	0	3	2013-02-23	06:20:00	00:1E:6B:22:B4:2E	1200	17083	191372550	2427625452	139
13114	0	3	2013-02-23	06:40:00	00:1E:6B:22:B4:2E	1200	17083	191374817	2428150651	139
13114	0	3	2013-02-23	07:00:00	00:1E:6B:22:B4:2E	1200	17083	191377577	2428701523	139
13114	0	4	2013-02-23	07:20:00	00:1E:6B:22:B4:2E	1200	17083	191379912	2429400806	139
13114	0	4	2013-02-23	07:40:00	00:1E:6B:22:B4:2E	1200	17083	191382672	2430018934	139
13114	0	4	2013-02-23	08:00:00	00:1E:6B:22:B4:2E	1200	17083	191385007	2430671793	139
13114	0	4	2013-02-23	08:20:00	00:1E:6B:22:B4:2E	1200	17083	191387308	2431329265	139
13114	3	28	2013-02-23	08:40:00	00:1E:6B:22:B4:2E	1200	17083	191862978	2435641772	139
13114	9	187	2013-02-24	00:00:00	00:1E:6B:22:B4:2E	1200	17083	204457778	2567484731	139
13114	11	190	2013-02-24	00:20:00	00:1E:6B:22:B4:2E	1200	17083	206144678	2596112363	139
13114	3	41	2013-02-24	00:40:00	00:1E:6B:22:B4:2E	1200	17083	206637031	2602339730	139
13114	0	3	2013-02-24	01:00:00	00:1E:6B:22:B4:2E	1200	17083	206639332	2602916472	139
13114	0	4	2013-02-24	01:20:00	00:1E:6B:22:B4:2E	1200	17083	206641667	2603557761	139
13114	0	4	2013-02-24	01:40:00	00:1E:6B:22:B4:2E	1200	17083	206644483	2604175481	139
13114	0	4	2013-02-24	02:00:00	00:1E:6B:22:B4:2E	1200	17083	206646762	2604782768	139
13114	0	4	2013-02-24	02:20:00	00:1E:6B:22:B4:2E	1200	17083	206649063	2605393218	139
13114	0	4	2013-02-24	02:40:00	00:1E:6B:22:B4:2E	1200	17083	206651364	2606079980	139
13114	0	4	2013-02-24	03:00:00	00:1E:6B:22:B4:2E	1200	17083	206654158	2606711141	139
13114	0	4	2013-02-24	03:20:00	00:1E:6B:22:B4:2E	1200	17083	206656459	2607339567	139

Figura 9 - Consulta dos dados da tabela de Consumo de Banda

3.2 – Metodologia

A metodologia utilizada para o desenvolvimento deste projeto é composta pelos seguintes passos:

- Efetuar revisão bibliográfica sobre Data Warehouse e Mineração de Dados;
- Criar um Data Mart para a geração de relatórios através do OLAP e limpeza dos dados;
- Desenvolver um software para mineração de dados com o uso de algoritmos de árvores de decisão;
- Realizar testes quanto às funcionalidades do software;
- Analisar os resultados obtidos.

Com a utilização dessa metodologia espera-se obter um software que possa ser utilizado por empresas de telecomunicações, clientes da empresa Alfa, para a obtenção de informações que deem suporte ao processo de tomada de decisão.

3.3 – Processo de Criação do *Data Mart*

Nos itens a seguir serão descritos os passos necessários para a criação de um *Data Mart* e também a criação do OLAP para manipulá-lo de forma mais amigável. Inicialmente será descrito o cenário, para só então ser explicado o processo de extração dos dados. Em seguida será descrito como foi realizada a criação do OLAP e, por fim, seus resultados.

Mesmo com os conhecimentos adquiridos sobre *Data Warehouse*, durante sua implementação foi necessária a busca por outros materiais técnicos para complementar a parte de desenvolvimento. Em outras palavras, os livros teóricos estudados, não abordavam de maneira clara a criação de um *Data Warehouse*. O material necessário foi encontrado no livro (Sarka et all, 2012), onde é explicada de maneira técnica a criação de *Data Warehouses* em bancos SQL Server 2012.

3.3.1 – Análise do Cenário

Enquanto a arquitetura utilizada no banco de dados da empresa Alfa era a de Entidade-Relacionamento, a arquitetura utilizada no *Data Mart* foi a multidimensional (esquema estrela), com várias tabelas de dimensão relacionadas com uma tabela fato. O banco de dados relacional possui também muitas tabelas com campos irrelevantes para o presente estudo.

Portanto, foi necessária a seleção dos campos, bem como a incorporação de outras tabelas, como por exemplo, as tabelas Estado, Cidade e Endereço formaram a tabela dimensão Geografia. A Figura 10 apresenta o diagrama do *Data Mart* criado.

Figura 10 - Diagrama do *Data Mart*

3.3.2 – Importação e Tratamento dos Dados

Para o processo de ETL (extração, transformação e carga) foi criado um projeto do Integration Services presente na ferramenta Visual Studio.

Na Figura 11 é apresentado o fluxo de importação de dados criado.

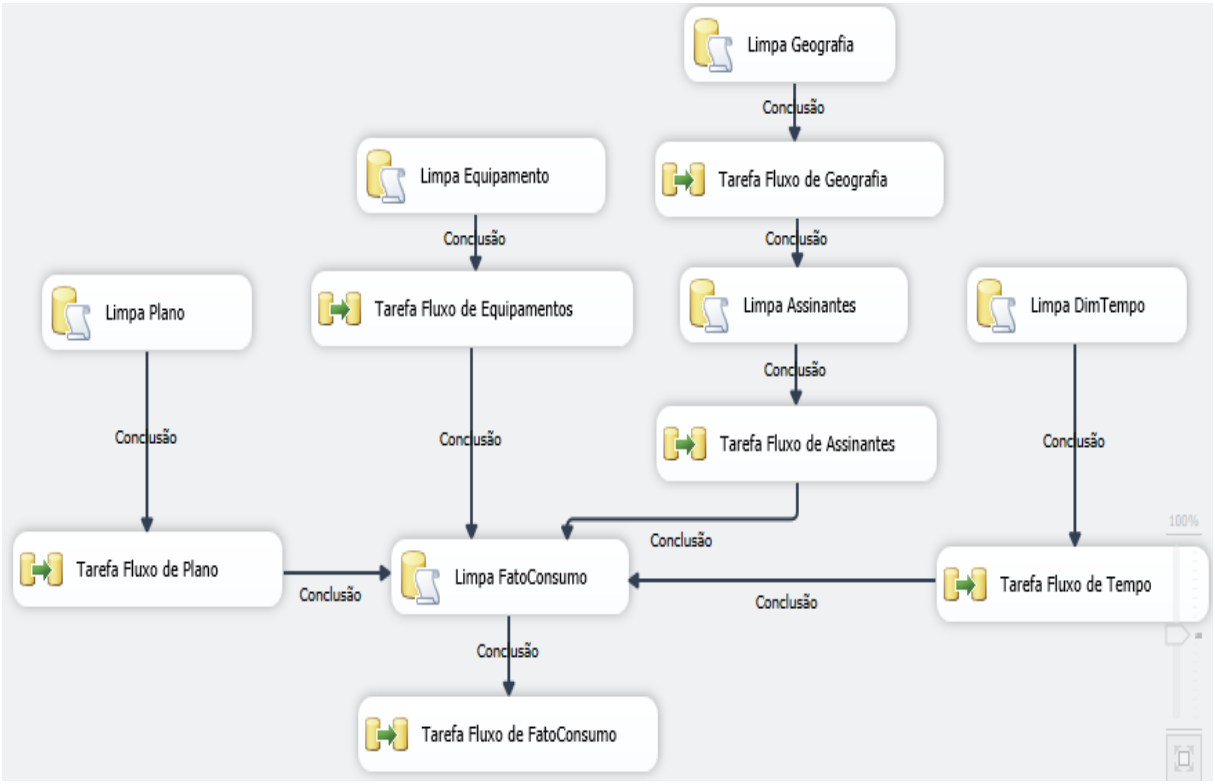


Figura 11 - Fluxo de importação criado no Visual Studio

Antes de os dados serem inseridos no *Data Warehouse*, é necessária a passagem pela etapa de transformação dos dados, onde os mesmos foram tratados. Esse tratamento foi responsável por substituir valores incorretos por valores padrão, bem como, a padronização de medidas e criação de tipos derivados.

A seguir são listadas algumas das transformações realizadas durante o processo:

- Criação do campo de idade derivado do campo de data de nascimento do Assinante;
- Criação dos campos primeiro nome e último nome, derivados do campo de nome do assinante;
- Utilização de valores para representar o sexo por ‘Masculino’, ‘Feminino’ e ‘Outros’, onde antes era apenas 1,2 ou nulo;
- Criação de diversos campos para representação de tempo como, por exemplo: ano, semestre, semana, etc.

Após a realização da transformação, foi efetuada a carga dos dados no banco SQL Server 2012.

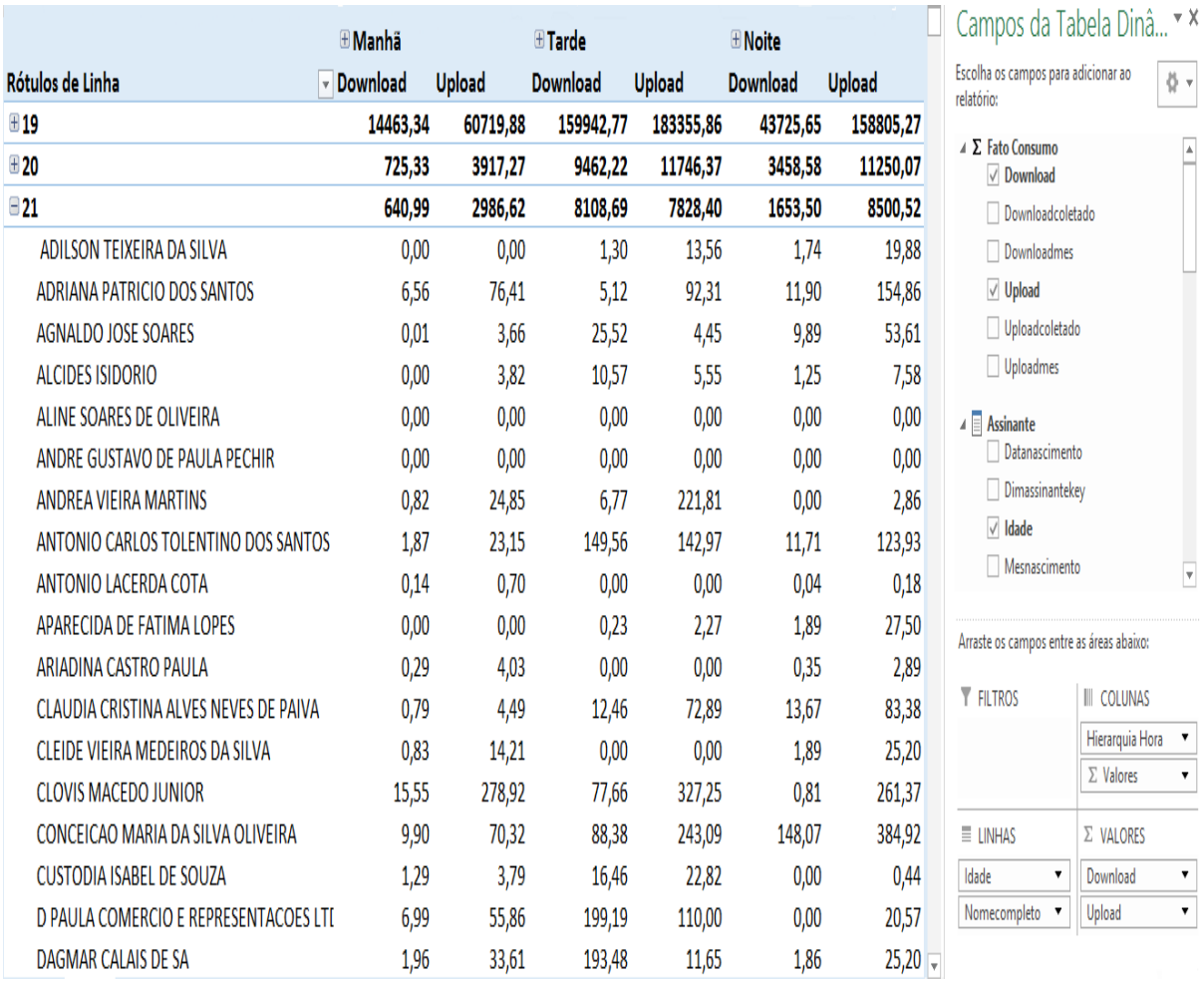
3.3.3 – Criação do OLAP para as consultas

Para o desenvolvimento do OLAP foi criado um projeto do Integration Services, onde foram descritos os tipos de dados, assim como sua utilização nas consultas. Por exemplo, campos relacionados com data, receberam o tipo Time para separá-los em: dia, hora, minuto, etc.

Foram criadas igualmente hierarquias entre os campos de uma tabela-dimensão para facilitar sua manipulação pelo usuário. Na tabela fato foram determinados o tipo de agregação que cada campo iria receber, como: soma, média, mínimo, máximo, etc.

Depois de finalizado este processo foi necessária a criação de uma conexão entre o Excel e o *Data Mart* implementado. As consultas realizadas pelos usuários ficaram assim mais fáceis e intuitivas, pois os mesmos apresentam mais habilidade com o uso de planilhas, do que com a as consultas SQL.

Na Figura 12 visualiza-se um exemplo da utilização do OLAP pelo Excel.



	Manhã		Tarde		Noite	
Rótulos de Linha	Download	Upload	Download	Upload	Download	Upload
19	14463,34	60719,88	159942,77	183355,86	43725,65	158805,27
20	725,33	3917,27	9462,22	11746,37	3458,58	11250,07
21	640,99	2986,62	8108,69	7828,40	1653,50	8500,52
ADILSON TEIXEIRA DA SILVA	0,00	0,00	1,30	13,56	1,74	19,88
ADRIANA PATRICIO DOS SANTOS	6,56	76,41	5,12	92,31	11,90	154,86
AGNALDO JOSE SOARES	0,01	3,66	25,52	4,45	9,89	53,61
ALCIDES ISIDORIO	0,00	3,82	10,57	5,55	1,25	7,58
ALINE SOARES DE OLIVEIRA	0,00	0,00	0,00	0,00	0,00	0,00
ANDRE GUSTAVO DE PAULA PECHIR	0,00	0,00	0,00	0,00	0,00	0,00
ANDREA VIEIRA MARTINS	0,82	24,85	6,77	221,81	0,00	2,86
ANTONIO CARLOS TOLENTINO DOS SANTOS	1,87	23,15	149,56	142,97	11,71	123,93
ANTONIO LACERDA COTA	0,14	0,70	0,00	0,00	0,04	0,18
APARECIDA DE FATIMA LOPES	0,00	0,00	0,23	2,27	1,89	27,50
ARIADINA CASTRO PAULA	0,29	4,03	0,00	0,00	0,35	2,89
CLAUDIA CRISTINA ALVES NEVES DE PAIVA	0,79	4,49	12,46	72,89	13,67	83,38
CLEIDE VIEIRA MEDEIROS DA SILVA	0,83	14,21	0,00	0,00	1,89	25,20
CLOVIS MACEDO JUNIOR	15,55	278,92	77,66	327,25	0,81	261,37
CONCEICAO MARIA DA SILVA OLIVEIRA	9,90	70,32	88,38	243,09	148,07	384,92
CUSTODIA ISABEL DE SOUZA	1,29	3,79	16,46	22,82	0,00	0,44
D PAULA COMERCIO E REPRESENTACOES LTI	6,99	55,86	199,19	110,00	0,00	20,57
DAGMAR CALAIS DE SA	1,96	33,61	193,48	11,65	1,86	25,20

Figura 12 - Utilização do OLAP pelo Excel

3.3.4 – Análise dos Resultados

Mesmo com uma grande quantidade de dados armazenados, a utilização do *Data Mart* diminuiu o tempo de espera na realização de consultas por parte dos usuários. Já com o OLAP, as consultas ficaram mais intuitivas, pois tornou-se possível a seleção os campos desejados e a sua forma de apresentação na obtenção de informações para os relatórios.

Os relatórios são criados pelos programadores da empresa Alfa através de solicitações dos usuários. Esse processo demanda tempo e nem sempre os resultados obtidos atendem as necessidades do usuário. Com a possibilidade de criação das consultas dinâmicas, filtragem dos resultados e a criação de gráficos com os dados através do Excel, tornaram as informações mais esclarecedoras e sem intermediários.

3.4 – Processo de Mineração de Dados

Nos próximos subitens serão descritos os passos da implementação do programa responsável pela mineração dos dados, obtidos do *Data Mart*.

3.4.1 – Ferramenta Weka

A ferramenta Weka foi julgada a mais adequada para ser utilizada na parte de implementação deste projeto, por ser uma ferramenta desenvolvida sob licença GNU, ou seja, possuir o código fonte aberto e ser bastante recomendada pelos peritos da área, além de ser free.

Weka é uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados. Os algoritmos podem ser aplicados diretamente a um conjunto de dados ou chamados a partir do seu próprio código Java. O Weka contém ferramentas para pré-processamento de dados, classificação, regressão, clustering, regras de associação e visualização (WEKA, 2013).

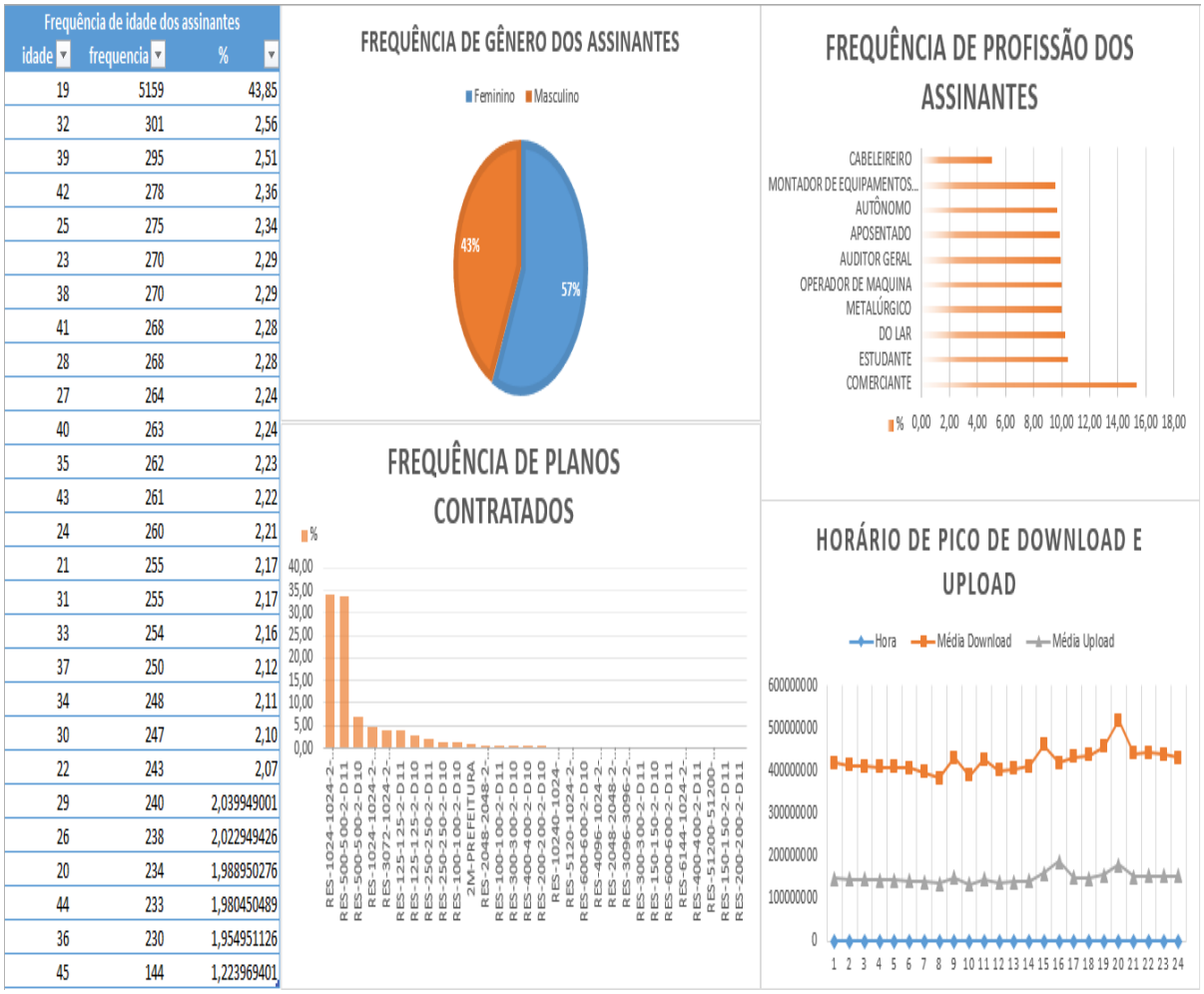
A Weka consolidou-se como a ferramenta de mineração de dados mais utilizada em ambiente acadêmico. Ela é aplicada não apenas em pesquisas científicas, mas principalmente para fins didáticos. Através de sua interface gráfica (conhecida como Weka Explorer) é possível conduzir processos de mineração de pequenas bases de dados, realizando a avaliação dos resultados obtidos e a comparação de algoritmos. Embora a Weka tenha um vasto número

de usuários, a maior parte desconhece que ela disponibiliza uma API que torna possível a utilização de suas classes dentro de programas Java. Trata-se do que os autores da ferramenta chamam de “forma programática” de utilizar a Weka. Com a utilização dessa API, torna-se possível a utilização da Weka em projetos reais de mineração de dados (Gonçalves, 2013).

3.4.2 – Análise Estatística dos Dados

Antes da realização da mineração de dados, é importante um conhecimento prévio das informações da base de dados. Assim sendo, recomenda-se a verificação dos resultados de algumas análises estatísticas relativas ao cenário em questão: frequência de idade dos assinantes, profissão, planos contratados e horários de pico de download e upload.

Foram escolhidos diferentes tipos de gráficos devido a diferença entre os tipos de dados. A Figura 13 apresenta os resultados obtidos, a partir destes resultados, torna-se possível escolher quais campos serão utilizados pelo programa de mineração.



Frequência de idade dos assinantes		
idade	frequencia	%
19	5159	43,85
32	301	2,56
39	295	2,51
42	278	2,36
25	275	2,34
23	270	2,29
38	270	2,29
41	268	2,28
28	268	2,28
27	264	2,24
40	263	2,24
35	262	2,23
43	261	2,22
24	260	2,21
21	255	2,17
31	255	2,17
33	254	2,16
37	250	2,12
34	248	2,11
30	247	2,10
22	243	2,07
29	240	2,039949001
26	238	2,022949426
20	234	1,988950276
44	233	1,980450489
36	230	1,954951126
45	144	1,223969401

Figura 13 - Estatísticas da Base de Dados

3.4.3 - Seleção dos Atributos

A identificação das variáveis chaves é a entrada para o estágio de transformação dos dados em informações uteis usando a Mineração de Dados. Conforme mencionado em Bartolomeu (2002), “deve-se reavaliar as variáveis selecionadas anteriormente, baseando-se nos resultados obtidos com as análises estatísticas. Feito isso, será possível identificar as variáveis que se mostraram mais relevantes, para que possam ser levadas às próximas etapas [...]”.

Os atributos escolhidos foram:

- Sexo;
- Profissão;
- Idade;
- Endereço;
- Descrição do Plano;
- Horário.

Os atributos que não foram utilizados na mineração de dados foram muito úteis na realização das consultas estatísticas.

3.4.4 – Escolha do Método de Mineração de Dados

O método escolhido para a mineração de dados, foi o de Classificação utilizando algoritmos de Árvore de Decisão, no caso o algoritmo C4.5 que, na ferramenta Weka, é conhecido como J48. Conforme foi apresentado na parte teórica, os algoritmos de árvores de decisão, mostram-se muito eficientes na:

- **Classificação de dados:** a partir do endereço dos assinantes, e da sua profissão, foi possível classificá-los de acordo com os planos oferecidos pela empresa;
- **Previsão de valores:** com os resultados obtidos na mineração de dados, conseguimos, de acordo com a profissão dos clientes cadastrados, descobrir tendências para as contratações dos novos clientes;
- **Segmentação de dados:** separar os assinantes em grupos de classes sociais levando em conta o endereço cadastrado para os mesmos;
- **Descrição dos atributos:** descrever o que acontece com mais frequência na base de dados, como ocorreu neste estudo, onde percebemos que o grupo de mulheres apresentavam maior frequência na contratação de planos de menor valor.

Para o desenvolvimento deste estudo utilizou-se a técnica de Classificação por ser julgada a mais adequada, de acordo com os exemplos supracitados, para o cenário proposto. Convém salientar, que o software implementado neste trabalho possui a opção de exportar os dados consultados para o formato de arquivo ARFF[3], podendo este ser empregado na utilização de qualquer método presente na ferramenta Weka, conforme o desejo do usuário.

3.4.5 – Processo de Implementação

Para a parte de implementação da Mineração de Dados, a idéia inicial do projeto era a de utilizar a linguagem de programação C#, pelo fato dela se comunicar melhor com o SQL Server. Contudo, durante a implementação foram encontradas diversas dificuldades, sendo a principal a de que o Weka não possui muitos dos seus métodos de mineração implementados em *.Net*. Portanto, era necessário procurar outra linguagem que se adequasse ao Weka.

A solução encontrada foi a de utilizar a linguagem de programação Java, pelo fato do Weka ter sido desenvolvido principalmente nesta linguagem. Outro motivo foi o grande suporte do Weka para a linguagem Java, não só com o manual com diversos exemplos nesta linguagem, como também com a presença de diversos *forums* disponíveis.

A IDE (ambiente integrado para desenvolvimento) utilizada foi o NetBeans 7.3.1 por ser maior o conhecimento sobre a utilização deste ambiente.

Para a comunicação entre o Java e o SQL Server, foi utilizado o driver jdbc disponível no próprio site da Microsoft. Depois de implementados os métodos de consulta dos dados, teve início a utilização da API do Weka, consultando tanto os materiais disponibilizados no site oficial, como em sites de comunidades do Java.

Logo após a conclusão do código de mineração, foram visualizadas as regras geradas pelo mesmo. Em algumas consultas, as regras geradas eram genéricas demais e apresentavam uma quantidade muito elevada de erro. Para resolver tais situações, optou-se pela desativação da opção de poda da árvore, o que gerou mais regras porém, mais específicas.

A Figura 14 mostra a situação citada acima, na parte esquerda da imagem temos uma

[3] Formato de Arquivo de Atributo-Relação, é um arquivo de texto que descreve uma lista de instâncias que compartilham um conjunto de atributos. Arquivos ARFF foram desenvolvidas pelo Projeto de Aprendizagem Máquina do Departamento de Ciência da Computação da Universidade de Waikato para uso com o software de aprendizado de máquina Weka (WEKA, 2013).

árvore não podada com 53,48% de instâncias corretamente classificadas, enquanto a árvore podada da direita apresenta apenas 33,95%.

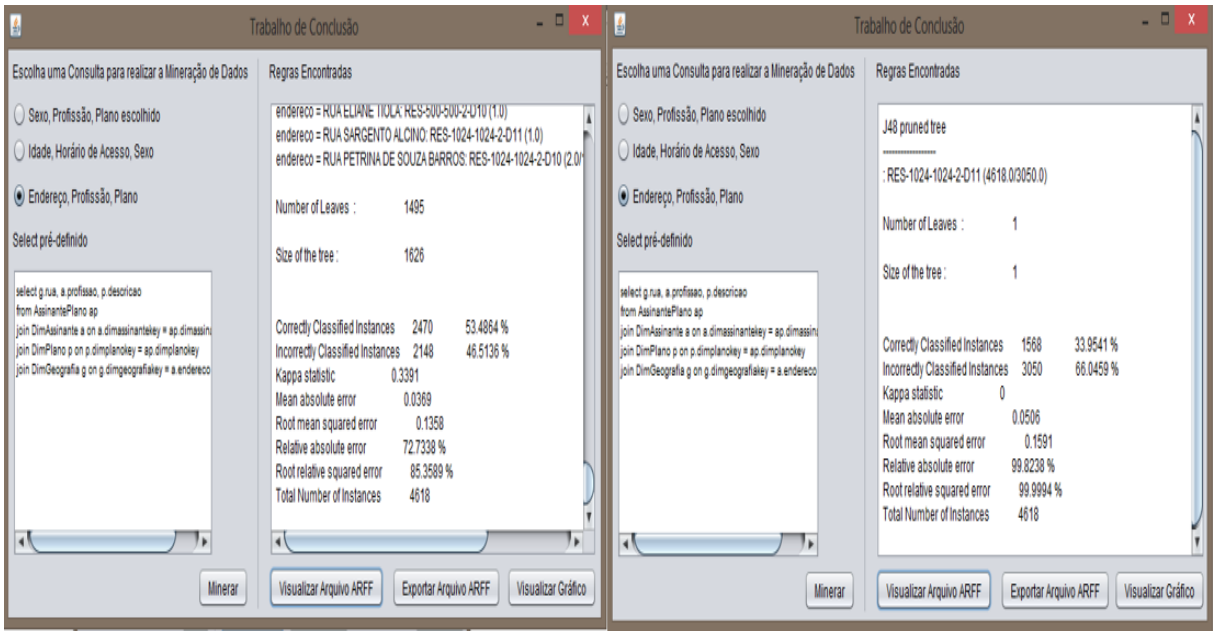


Figura 14 - Resultado de Árvores não podada e podada

3.4.6 – Análise dos Resultados Obtidos

Na Figura 15 são apresentadas algumas das telas do software implementado para a consulta de relação entre o sexo dos assinantes, idade e horário de acesso dos mesmos; à esquerda a tela com o ARFF gerado e, à direita, visualização gráfica da árvore de decisão.

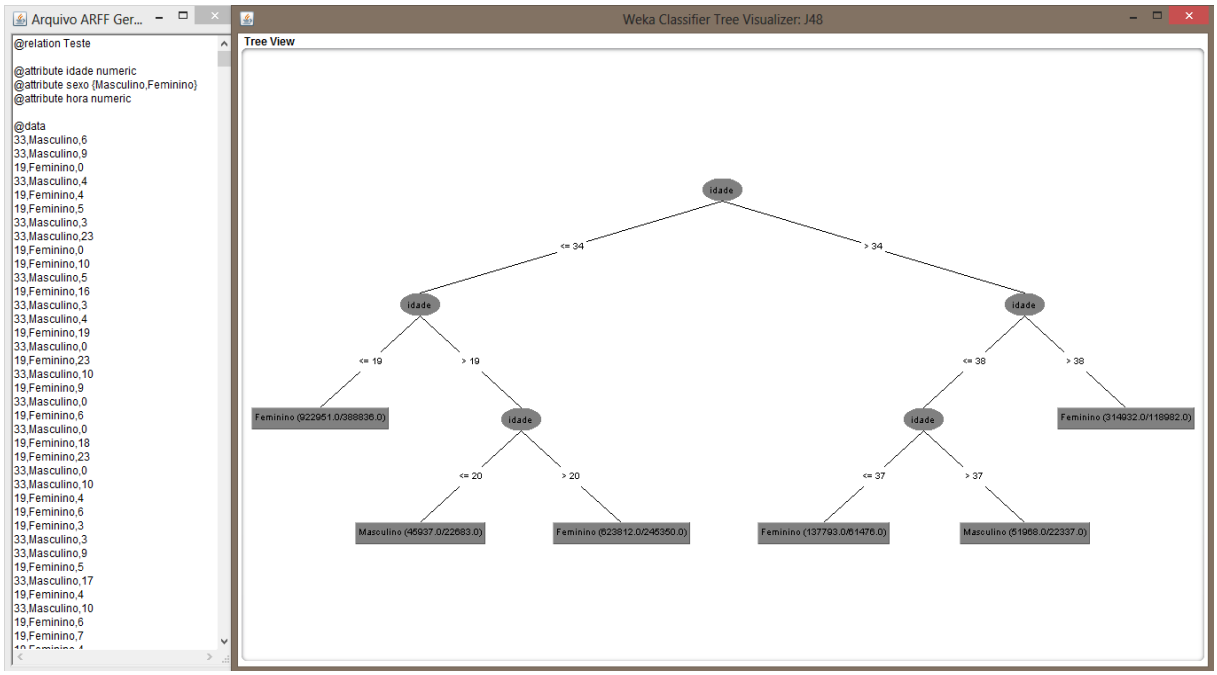


Figura 15 - Telas do Sistema Implementado

Depois de implementada a parte de mineração de dados, foi feita a análise de algumas das regras geradas, constatou-se que algumas delas eram condizentes com a realidade. Por exemplo: “profissão = APOSENTADO então RES-500-500-2-D11 (451.0/287.0)”, onde RES-500-500 é plano de Internet de baixa velocidade, e (451/287) significa que essa regra foi correta para 451 registros e incorreta para 287 outros registros. Pode-se concluir com esses valores que os aposentados contratam planos de menor velocidade, portanto mais baratos, justamente por não precisarem de altas velocidades. Os aposentados comumente utilizam a Internet para ler notícias, receitas, envio de e-mail, tarefas estas que não exigem muita banda.

Aqui se encerra a parte do estudo de caso proposto. No capítulo seguinte são elencadas as conclusões obtidas e, por fim, sugerimos trabalhos futuros.

Conclusão

Por meio da análise bibliográfica obteve-se os conhecimentos necessários para a elaboração não só de um *Data Mart*, mas também o estudo das características, das vantagens e desvantagens das principais técnicas de mineração, o que possibilitou a escolha da técnica mais adequada para o cenário em questão.

Com a criação do *Data Mart*, as informações obtidas nas bases de dados foram limpas e tratadas, tornando o *Data Mart* uma fonte de dados confiável para a empresa, sendo possível sua utilização tanto em relatórios empresariais, como também pelo software responsável pela mineração de dados que exige elementos sem redundância, sem erros de escrita, em um formato padronizado.

Com a utilização do ambiente criado neste trabalho as empresas de telecomunicação irão obter, de forma mais simples e rápida, as informações analíticas necessárias para o processo de tomada de decisão.

Ao final da implementação deste estudo, obteve-se um software que:

- Consulta uma base de dados;
- Minera as informações obtidas;
- Permite a exportação dos dados em arquivos ARFF para posterior utilização na ferramenta Weka;
- Lista as regras geradas pela mineração em forma de texto;
- Apresenta de forma gráfica as regras encontradas.

Como sugestão para trabalhos futuros, propõem-se:

- Acrescentar campos para número de dependentes e, se possível, a idade e sexo dos mesmos para maior precisão na mineração de dados;
- O desenvolvimento de mais métodos da ferramenta Weka na linguagem C#, visto que a maioria está disponível em Java;
- O estudo e utilização de outras regras de negócio disponíveis no sistema da empresa Alfa;
- A utilização de outras técnicas de mineração de dados para o software desenvolvido.

Referências Bibliográficas

Almeida, A. M. DE. **Proposição de Indicadores Para Avaliação Técnica de Projetos de Datawarehouse: Um Estudo de Caso no Datawarerouse da Plataforma Lattes**. Universidade Federal de Santa Catarina, 2006.

Barbosa, D. M. **Aplicação de Data Webhousing para Monitoramento de Acessos a Sites Web de Grupos de Pesquisa e Desenvolvimento: Um Estudo de Caso**. Universidade Federal de Santa Catarina, 2003.

Bartolomeu, T. A. **Modelo de Investigação de Acidentes do Trabalho Baseado na Aplicação de Tecnologias de Extração de Conhecimento**. Universidade Federal de Santa Catarina, 2002.

Beckmann, M. **Algoritmos Genéticos como Estratégia de Pré-Processamento em Conjuntos de Dados Desbalanceados**. Universidade Federal do Rio de Janeiro, 2010.

Brito, J. J. **Processamento de Consultas SOLAP Drill-Across e com Junção Espacial em Datawarehouses Geográficos**. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2012.

Clemes, M. **Data Warehouse como Suporte ao Sistema de Informações Gerenciais em Uma Instituição**. Universidade Federal De Santa Catarina, 2001.

Fayyad, U.; Shapiro, G.; Smyth, P. **From Data Mining to Knowledge Discovery: An Overview**, AAAI Press / MIT Press, 1996.

Garcia, S. C. **O Uso de Árvores de Decisão na Descoberta de Conhecimento na Área da Saúde**. Universidade Federal Do Rio Grande Do Sul, 2003.

Gonçalves, E. **Mineração de dados na prática com Weka API - Revista SQL Magazine 107.** Disponível em: <http://www.devmedia.com.br/mineracao-de-dados-na-pratica-com-weka-api-revista-sql-magazine-107/26841>. Acesso em 5 de Dezembro de 2013.

Haisten, M. **The Real-Time Data Warehouse: The Next Stage in Data Warehouse Evolution**. DM Review. Disponível em: < https://www.db.bme.hu/sites/default/files/history_of_dw.pdf>. Acesso em 20 de Junho de 2013, 1999.

Halmenschlager, C. **Um algoritmo para Indução de Árvores e Regras de Decisão**. Universidade Federal Do Rio Grande Do Sul, 2002.

Inmon, W. H. **Building the Data Warehouse**. 4ª Edição ed. Indianapolis: Wiley Publishing, 2005.

Jesus, A. P. De. **Data Mining Aplicado à Identificação do Perfil dos Usuários de Uma Biblioteca para a Personalização de Sistemas Web de Recuperação e Disseminação de Informações**. Universidade Federal de Santa Catarina, 2004.

Kimball, R.; Ross, M. **The Data Warehouse Toolkit-The Complete Guide to Dimensional Modeling**. 2. ed. New York: John Wiley and Sons, Inc., 2004. p. 447

Moraes, A. F. De. **Um Modelo Representativo de Conhecimento para Aplicação da Mineração de Dados no Cadastro Técnico Urbano**. Universidade Federal De Santa Catarina, 2003.

Ninho, C. V. Dos S. **Aplicação de Mineração de Dados nas Transações de Compras em Empresa do Segmento de Petróleo e Gás**. Universidade Federal do Rio de Janeiro, 2011.

Ponniah, P. **Data Warehousing Fundamentals for IT Professionals, Second Edition**. 2. ed. Hoboken: JohnWiley & Sons, 2010. p. 602.

Reeves, L. L. **A Manager ' s Guide to Data**. 1. ed. Indianapolis: Wiley Publishing, 2009. p. 482.

RNP. **Introdução a Gerenciamento de Redes TCP/IP**. Disponível em: < http://www.rnp.br /newsgen /9708/n3-2.html>. Acesso em 7 de Outubro de 2013, 2004.

Sarka, D. et al. **Exam 70-463: Implementing a Data Warehouse with Microsoft® SQL Server® 2012.** 1. Ed. California: O'Reilly Media, Inc, 2012, p. 600.

Sarkis, L. C. **Data Warehouse: O Processo De Migração De Dados.** Universidade Federal De Santa Catarina, 2001.

Seixas, J. A. De. **Integração De Mineração De Dados E Visualização De Informações Geográficas Aplicados À Saúde Pública.** Universidade Federal do Rio de Janeiro, 2011.

Souza, N. De. **Ambiente De Apoio À Decisão para o Programa de Avaliação Institucional: Uma Aplicação na Universidade do Vale do Itajaí – Univali.** Universidade Federal de Santa Catarina, 2002.

Thearling, K. **An Introduction to Data Mining**. Disponível em: < http://www.thearling.com /text/dmwhite/dmwhite.htm>. Acesso em 3 de Setembro de 2013, 2010.

Wagner, C. A. **Estudo Para Implantação de Um Data Warehouse em um Ambiente Empresarial.** Universidade Federal De Santa Catarina, 2003.

WEKA. **Machine Learning Group**. Disponível em: < http://www.cs.waikato.ac.nz /ml/index.html>. Acesso em 7 de Outubro de 2013.