UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA
ESPECIALIZAÇÃO EM BANCO DE DADOS

DANILO BONIFÁCIO TELES
FABRICIO NOGUEIRA DOS SANTOS
LEANDRO PEDROSA

# A Mineração de Dados

## Aplicada à Inteligência de Negócios

Goiânia
2014

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA
ESPECIALIZAÇÃO EM BANCO DE DADOS

# AUTORIZAÇÃO PARA PUBLICAÇÃO DE MONOGRAFIA EM FORMATO ELETRÔNICO

Na qualidade de titular dos direitos de autor, **AUTORIZO** o Instituto de Informática da Universidade Federal de Goiás – UFG a reproduzir, inclusive em outro formato ou mídia e através de armazenamento permanente ou temporário, bem como a publicar na rede mundial de computadores (*Internet*) e na biblioteca virtual da UFG, entendendo-se os termos "reproduzir" e "publicar" conforme definições dos incisos VI e I, respectivamente, do artigo 5º da Lei nº 9610/98 de 10/02/1998, a obra abaixo especificada, sem que me seja devido pagamento a título de direitos autorais, desde que a reprodução e/ou publicação tenham a finalidade exclusiva de uso por quem a consulta, e a título de divulgação da produção acadêmica gerada pela Universidade, a partir desta data.

**Título:** A Mineração de Dados – Aplicada à Inteligência de Negócios

**Autor(a):** Danilo Bonifácio Teles
Fabricio Nogueira dos Santos
Leandro Pedrosa

Goiânia, 17 de Abril de 2014.

Danilo Bonifácio Teles
Fabricio Nogueira dos Santos
Leandro Pedrosa
– Autor

Edmundo Spoto – Orientador

Leandro Luís Galdino de Oliveira – Co-Orientador

DANILO BONIFÁCIO TELES
FABRICIO NOGUEIRA DOS SANTOS
LEANDRO PEDROSA

# A Mineração de Dados

## Aplicada à Inteligência de Negócios

Monografia apresentada ao Programa de Pós–Graduação do Instituto de Informática da Universidade Federal de Goiás, como requisito parcial para obtenção do Certificado de Especialização em Computação.

**Área de concentração:** Banco de Dados.

**Orientador:** Prof. Edmundo Spoto

**Co-Orientador:** Prof. Leandro Luís Galdino de Oliveira

Goiânia
2014

DANILO BONIFÁCIO TELES
FABRICIO NOGUEIRA DOS SANTOS
LEANDRO PEDROSA

# A Mineração de Dados

## Aplicada à Inteligência de Negócios

Monografia apresentada no Programa de Pós–Graduação do Instituto de Informática da Universidade Federal de Goiás como requisito parcial para obtenção do Certificado de Especialização em Computação, aprovada em 17 de Abril de 2014, pela Banca Examinadora constituída pelos professores:

---

**Prof. Edmundo Spoto**
Instituto de Informática – UFG
Presidente da Banca

---

**Prof. Leandro Luís Galdino de Oliveira**
Universidade Federal de Goiás – UFG

---

**Prof. Sergio T. Carvalho**
Universidade Federal de Goiás – UFG

---

**Profa. Daiane Dias da Silva**
Universidade Federal de Goiás – UFG

Aos familiares e amigos que nos incentivaram e apoiaram nossas decisões.

# Agradecimentos

Primeiramente aos nossos familiares pelo apoio e compreensão nos momentos difíceis. Ao Prof. Dr. Edmundo Sérgio Spoto pela orientação, encaminhamento dado a este trabalho. A todos os professores que contribuíram para nosso crescimento pessoal e profissional. Nossos sinceros agradecimentos.

A maneira como você coleta, gerencia e utiliza as informações determina se você vai vencer ou perder.

**Bill Gates**,

.

# Resumo

Bonifácio Teles Danilo,Dos Santos Nogueira Fabricio, Pedrosa Leandro. **A Mineração de Dados**. Goiânia, 2014. 34p. Monografia de Especialização. Especialização em Banco de Dados, Instituto de Informática, Universidade Federal de Goiás.

Técnicas sobre a Inteligência de Negócio têm sido um importante tópico de estudo. A descoberta de conhecimento em banco de dados e o processo de mineração de dados destaca-se por ser um conjunto de processos que une o uso do poder de processamento das máquinas atuais, com avançados algoritmos de inteligência artificial, projetados para analisar uma grande massa de dados com a finalidade de se descobrir informações que até então eram desconhecidas ou desconsideradas. O objetivo deste estudo foi identificar as diferentes técnicas de utilizaçã e manipulaão de grandes massas de dados para a geração de conhecimento com ganhos reais para indivíduos e corporações, com tempo e custo adequados. A metodologia empregada é um estudo bibliográfico, onde a busca por periódicos relevantes, livros e revistas científicas sobre o tema foi realizada. O principal resultado deste estudo é identificar as ferramentas que permitem a descoberta do conhecimento, após análise de grande volume de dados.

**Palavras–chave**

Inteligência de negócios. Descoberta de conhecimento em banco de dados. Mineração de dados.

# Abstract

Bonifácio Teles Danilo,Dos Santos Nogueira Fabricio, Pedrosa Leandro. **T**. Goiânia, 2014. 34p. Monografia de Especialização. Especialização em Banco de Dados, Instituto de Informática, Universidade Federal de Goiás.

Techniques on Business Intelligence has been an important topic of study. knowledge discovery in database and the process of Data Mining stands out for being a tool that combines the use of the processing power of today's machines, with advanced artificial intelligence algorithms designed to analyze a large body of data for the purpose of discovering information that hitherto were unknown or disregarded. The aim of this study was to identify the different techniques to use and manipulate large amounts of data to generate knowledge with real gains for individuals and corporations , with adequate time and cost. The methodology used was a literature study , where the search for relevant books and journals on the subject journals was performed . The main result of this study was to identify the tools that enable knowledge discovery , after analyzing large volumes of data.

**Keywords**

Business Intelligence. knowledge discovery in database. Data Mining.

# Sumário

# Lista de Figuras

# Lista de Tabelas

CAPÍTULO 1

# Introdução

A necessidade humana fez com que o homem se destacasse das outras espécies em nosso planeta por entenderem o meio em que vivem e desenvolverem formas de facilitar sua sobrevivência. Isso tem, ao longo do tempo, proporcionado o acúmulo de conhecimento e a geração de tecnologias que evoluem a passos largos, principalmente nos dois últimos séculos.

Essa evolução cada vez mais acelerada trouxe consigo o aumento da quantidade de dados que já não podem ser mais simplesmente transmitidos entre pessoas ou armazenados em meios que dificultem o seu acesso. Esta necessidade fez com que sistemas fossem projetados para estes fins. Assim, toda essa massa de dados não era trabalhada de forma a gerar o principal: O conhecimento. Tal conhecimento pode ser extraído por meio de técnicas de Inteligência de Negócio, as quais serão apresentadas ao longo deste documento, tendo como foco principal o processo de Descoberta em conhecimento em banco de dados.

A Mineração de Dados destaca-se como uma das mais interessantes e inovadoras formas de analisar e localizar padrões em uma grande massa de dados extraindo informações e gerando conhecimentos para os níveis estratégicos em diversas áreas. Ela permite que, através do uso de técnicas avançadas de inteligência artificial e poderosos recursos de hardware disponíveis, seja possível realizar uma profunda busca em grandes massas de dados por informações que, pela limitação humana, podem ser ignoradas ou até mesmo desconhecidas.

Em nossos estudos sobre banco de dados constatamos que em todo projeto de sistemas de informação o gerenciamento de dados é considerado como um dos ponto principais em sua elaboração e construção, no entanto, estes dados nem sempre recebem a devida atenção e em boa parte das corporações funcionam apenas como fonte para sistemas transacionais.

Logo, percebe-se que havendo um tratamento nestes dados de forma adequada, eles podem se tornar uma grande fonte de conhecimento, direcionando as corporações a se aperfeiçoarem em seus processos e sistemas, visando obter o máximo de ganho, seja ele tangível ou não.

Neste contexto, é possível exemplificar que através de uma simples análise da carteira de clientes de uma determina organização pode-se descobrir quais de seus atuais clientes possuem potencial para adquirir outras linhas de seus produtos e oferecê-los de forma a atender necessidades previamente manifestadas por estes clientes dentro de um perfil pré-estabelecido.

A Inteligência de Negócio e a descoberta do conhecimento em banco de dados devem entrar diretamente no processo da construção de uma solução para o negócio de médias a grandes corporações. O correto tratamento dos dados em sua forma bruta, sua contextualização e a extração do conhecimento auxiliados pelas técnicas aqui estudadas podem levar qualquer domínio de conhecimento a otimização de seus processos e às melhores tomadas de decisões.

CAPÍTULO 2

# Inteligência de negócio

> "Apresentar formas de se trabalhar os dados de forma a obter o melhor resultado dos mesmos, através de técnicas de mineração de dados é necessário estudar antes de tudo a aplicação mais ampla das informações. Ou seja, o processo de inteligência de negócio baseia-se na transformação de dados em informações, depois em decisões e finalmente em ações."[34, Efrain Turban].

### 2.0.1 Histórico

O termo inteligência de negócio, em inglês *business intelligence* sigla *BI*, foi criado pela empresa Gartner na década de 80, entretanto sua aplicação já existia há alguns séculos, por exemplo, a sociedade do Egito antigo que observava as cheias do rio Nilo e a utilizavam como referência para determinar ações que deveriam efetuadas.

> A sociedade do Oriente Médio antigo utilizava os princípios básicos de inteligência de negôcio quando cruzavam informações obtidas junto à natureza em beneficio de suas aldeias. Analisar o comportamento das marés, os períodos chuvosos e de seca, a posição dos astros, entre outras, eram formas de obter informações que seriam utilizadas para tomar decisões importantes, permitindo a melhoria de vida de suas respectivas comunidades[10].

Como exemplo é possível citar a sociedade do Egito antigo que observava as cheias do rio Nilo, e a utilizavam como referência para determinar ações que deveriam efetuadas.

Segundo o departamento de hidráulica e saneamento da universidade federal da Bahia esta observação dos ciclos que ocorriam com o rio e o mapeamento de suas variações, juntamente com seus períodos de ocorrência, proporcionou se definir as melhores épocas para plantio e até mesmo a criação de um “nilômetro” que era uma escala acompanhada exclusivamente pelos sacerdotes para a leitura do nível do rio Nilo, as variações do nível do rio nesta escala determinavam inclusive qual taxa de imposto a ser cobrada.


**Figura 2.1:** *Variações do nível das águas do rio Nilo. fonte:[30]*

## 2.1 Dado, informação, conhecimento e decisão

A inteligência de negócio aborda como será a tomada de decisão e o embasamento que será utilizado para tal, isso é feito baseado em informações que são extraídas de dados existentes na corporação. Assim, faz-se necessário definir o que é dado, informação, conhecimento e o foco principal que é a decisão.

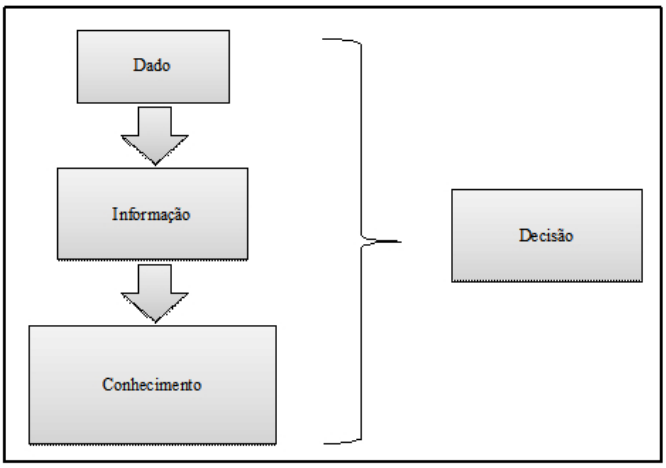


**Figura 2.2:** *Representação da relação entre dado, informação, conhecimento e decisão. fonte:[10]*

### 2.1.1 Dado

Dado é o valor bruto sem contexto; seria o que se armazena propriamente no banco de dados.

**Analogia:** Se dissermos para um grupo de pessoas: Imaginem o dado como sendo apenas o número 2, algumas pessoas desse grupo podem perguntar, mas 2 o quê? Não há nada mais para se complementar é só um dado sem um contexto.

### 2.1.2 Informação

Informação é a contextualização de um valor bruto, ou seja, é o resultado do tratamento do dado, comparado, classificado ou relacionado a outros dados.

**Analogia:** O 2 é apenas um número, quando colocamos: "A equipe de basquete da cidade fez apenas 2 pontos na partida". Com essa informação as pessoas já conseguem entender melhor sobre do que se trata o assunto, estamos acrescentando algo a mais nos dados, ele não é apenas um valor seja numérico, caracteres alfa numéricos ou datas e assim, pulamos do nível de dado para o nível de informação.

### 2.1.3 Conhecimento

Conhecimento é quando se aplica a informação recebida em algum contexto e através da contextualização dessa informação, obtemos o conhecimento e com o conhecimento gerado através das informações, temos o auxílio na tomada de decisões.

**Analogia:** Se a equipe de basquete da cidade fez apenas 2 pontos isso significa que eles não foram tão eficientes na partida, ou seja, nós tínhamos apenas o número 2, que se trata de um dado, depois obtivemos a informação de que a equipe de basquete fez apenas 2 pontos na partida, a informação foi gerada e agora, se vamos discutir o desempenho da equipe no jogo, essa informação passa a ser conhecimento, afinal de contas, aplicamos a informação dentro de um contexto. Dessa forma podemos tirar como conhecimento que a equipe não foi bem pois não jogaram em sua melhor forma física ou a equipe adversária foi totalmente superior a nossa equipe, podemos extrair também que a equipe fez apenas 1 cesta na partida que resultou em 2 pontos.

Estes três itens são a base para atingir o principal deles: A tomada de decisão. É justamente neste ponto que a correta análise dos dados pode ser o diferencial para se ter uma informação de qualidade, fidedigna e integra.

CAPÍTULO 3

# Armazém de dados

Um armazém de dados, em inglês *data mart* sigla *DM*, pode ser definido como um conjunto de dados destinado a auxiliar em decisões de negócios. Seus dados não são voláteis, pois uma vez carregados não podem mais sofrer alterações.

Cada conjunto de dados, ao ser carregado em um armazém de dados representa um determinado tempo que o identifica dentre os demais e fica associado a uma visão instantânea e sumarizada dos dados operacionais que corresponde ao momento de carga dos dados, para se realizar análise e tendências.

Para se definir uma forma mais objetiva é importante fazer uma comparação com o conceito tradicional de banco de dados, um banco de dados que, nada mais é do que uma coleção de dados operacionais armazenados e utilizados pelo sistema de aplicações de um domínio específico, podem ser chamados de dados "operacionais"ou "primitivos".

Os bancos de dados operacionais armazenam as informações necessárias para as operações diárias do domínio, são utilizados para registrar e executar operações pré-definidas, por isso seus dados podem sofrer constantes mudanças conforme as necessidades atuais. Por não ocorrer redundância nos dados e as informações históricas não ficarem armazenadas por muito tempo, este tipo de banco de dados não exige grande capacidade de armazenamento.

Com base nestes conceitos pode-se concluir que o armazém de dados não é um fim, mas sim um meio que as empresas dispõem para analisar informações históricas, podendo utilizá-las para a melhoria dos processos atuais e futuros.

Resumindo as principais características de um armazém de dados tem-se:

- **Orientado a assuntos**: por exemplo, vendas de produtos a diferentes tipos de clientes, atendimentos e diagnósticos de pacientes, rendimento de estudantes;
- **Integrado**: diferentes nomenclaturas, formatos e estruturas das fontes de dados precisam ser acomodadas em um único esquema para prover uma visão unificada e consistente da informação;
- **Séries temporais**: o histórico dos dados por um período de tempo superior ao usual em BD's transacionais permite analisar tendências e mudanças.

- **Não volátil**: Os dados de uma armazém de dados não são modificados como em sistemas transacionais (exceto para correções), mas somente carregados e acessados para leituras, com atualizações apenas periódicas.

Observe na tabela 3.1 uma relação de funcionalidades de um armazém de dados e um banco de dados operacional e suas diferenças:

**Tabela 3.1:** *Comparação entre DW e Bancos de dados fonte: [10]*

| **Características** | **Banco de dados Operacionais** | ***Data Warehouse*** |
|---|---|---|
| Objetivo | Operações diárias do negócio | Analisar o negócio |
| Uso | Operacional | Informativo |
| Tipo de processamento | OLTP | OLAP |
| Unidade de trabalho | Inclusão, alteração e exclusão | Carga e consulta |
| Números de usuários | Milhares | Centenas |
| Tipo de usuários | Operadores | Comunidade gerencial |
| Interação do usuário | Somente pré-definida | Pré-definida e ad-hoc |
| Volume | Megabytes – Gigabytes | Gigabytes – Terabytes |
| Histórico | 60 a 90 dias | 5 a 10 anos |
| Granularidade | Detalhados | Detalhados e resumidos |
| Redundância | Não ocorre | Ocorre |
| Estrutura | Estática | Variável |
| Manutenção desejada | Mínima | Constante |
| Acesso de registros | Dezenas | Milhares |
| Atualização | Contínua (Tempo real) | Periódica (*Em batch*) |
| Integridade | Transação | A cada atualização |
| Número de índices | Poucos/Simples | Muitos/Complexos |
| Intenção dos índices | Localizar um registro | Aperfeiçoar consultas |

Para organizar os dados em um armazém de dados, são necessários novos métodos de armazenamento, estruturação e novas tecnologias para a geração e recuperação dessas informações, essas tecnologias diferem dos padrões operacionais de sistemas de banco de dados em três maneiras:

- Disponibilizam visualizações informativas, pesquisando, reportando e modelando capacidades que vão além dos padrões de sistemas operacionais frequentemente oferecidos;
- Armazenam dados frequentemente em formato de cubo *On-line Analytical Processing* (OLAP)[1] multidimensional, permitindo rápida agregação de dados e detalhamento das análises (*drilldown*, *drill trought* etc.);

[1]O OLAP (On-line Analytical Processing) oferece uma alternativa diferente. Voltado para a tomada de decisões, proporciona uma visão dos dados orientados à análise, além de uma navegação rápida e flexível. O OLAP recebe dados do OLTP para que possa realizar as análises e essa carga de dados acontece conforme a necessidade da empresa e sendo um sistema de tomada de decisão não realiza transações (INSERT, UPDATE, DELETE) [25].

- Dispõem de habilidade para extrair, tratar e agregar dados de múltiplos sistemas operacionais em mercado de dados ou armazéns de dados separados.

A ferramenta de extração dos dados é uma parte muito importante do projeto do armazém de dados, mas apenas uma pequena parcela de um conjunto bastante complexo de soluções de *hardware* e *software*, e somente depois de definido e projetado o escopo do projeto e depois de construído o repositório de dados, é que se deve chegar às ferramentas de interface com o usuário responsáveis pelo meio de campo entre as bases de dados e os usuários finais da área executiva.

## 3.1 *Mercado de dados*

O mercado de dados é um sub-conjunto do armazém de dados, ou seja, o agrupamento ou sumarização dos dados em assuntos específicos e relacionados. Numa visão comparativa dos dados, onde considera-se os requisitos, escopo, integração, tempo, agregação, análise e dados voláteis, percebe-se que a diferença está no escopo, pois enquanto o armazém de dados é feito para atender um domínio como um todo, o mercado de dados é criado para atender um sub-conjunto desse domínio. Observe que atender um sub-conjunto pode significar reunir dados de outros setores, já que, na prática, raramente um único setor possui ou gera toda informação que precisa.

### Arquitetura de mercado de dados independente

Uma arquitetura de mercado de dados independente consiste em uma série de mercado de dados independentes que são controlados por um grupo particular específico e são construídos especificamente para atender necessidades específicas, ou seja, neste tipo de arquitetura, são categorizados de forma que os dados existentes em cada um deles é referente somente ao interesse do setor a que ele se destina. Este tipo de implementação tem um impacto mínimo no tempo gasto e na alocação de recursos, porém a integração mínima e a falta de uma visão mais global dos dados podem representar algum tipo de empecilho.

### Arquitetura de mercado de dados integrados

Parecem com a arquitetura global, porém, mercado de dados são separados por departamento, gerando os seus próprios dados e conectados através de interfaces de integração que possibilitam a troca de informações entre eles. Normalmente possuem alguns cadastros duplicados, como, por exemplo, o cadastro de clientes que acaba sendo duplicados.

No formato global isso não ocorreria, pois estaria tudo centralizado, mas isso pode ser contornado tratando de forma clara as rotinas de integração.

## 3.2 Extração, Transformação e Carga

O processo de extração, transformação e carga, em inglês *Extract, Transform, Load* sigla *ETL*, é utilizado para efetuar a carga dos dados em um armazém de dados. A primeira parte desse processo é a extração de dados dos sistemas que podem partir de diferentes fontes, podem ser bases de dados relacionais em ambientes heterogêneos, arquivos de texto, planílhas eletrônicas, bases de dados não relacionais etc., ou seja, qualquer fonte de dados que atenda o domínio.

O estágio de transformação aplica uma série de regras ou funções aos dados extraídos para derivar os dados a serem carregados. Algumas fontes de dados necessitarão de pouca manipulação de dados. Em outros casos, podem ser necessários um ou mais de um dos seguintes tipos de transformação:

- Seleção de apenas determinadas colunas para carregar (ou a seleção de nenhuma coluna para não carregar);
- Tradução de valores codificados (se o sistema de origem armazena 1 para sexo masculino e 2 para feminino, mas o Data Warehouse armazena M para masculino e F para feminino, por exemplo), o que é conhecido como limpeza de dados;
- Codificação de valores de forma livre (mapeando “Masculino”,“1” e “Sr.” para M, por exemplo);
- Derivação de um novo valor calculado (montanteVendas iqual à quantidade mutiplicado pelo preço Unitário, por exemplo);
- Junção de dados provenientes de diversas fontes;
- Resumo de várias linhas de dados (total de vendas para cada loja e para cada região, por exemplo);
- Geração de valores de chaves substitutas (*surrogate keys*);
- Transposição ou rotação (transformando múltiplas colunas em múltiplas linhas ou vice-versa);
- Quebra de uma coluna em diversas colunas (como por exemplo, colocando uma lista separada por vírgulas e especificada como uma cadeia em uma coluna com valores individuais em diferentes colunas).

O processo de carga de um armazém de dados varia amplamente, dependendo das necessidades da organização. As informações existentes de alguns armazéns de dados podem ser substituídas semanalmente, com dados cumulativos e atualizados. No entanto outros armazém de dados (ou até mesmo partes do mesmo armazém de dados) podem ter

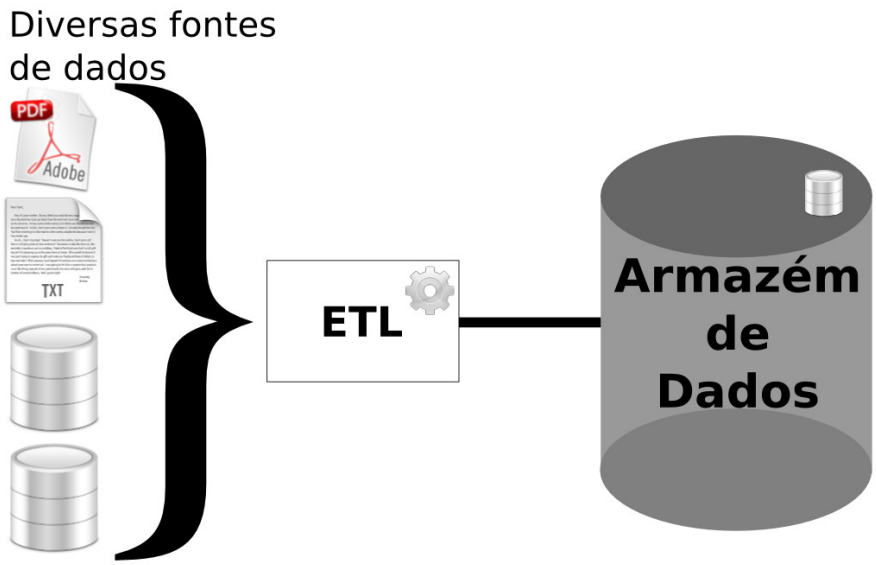


**Figura 3.1:** *Fluxo geral de ETL. fonte: [37]*

dados adicionados a cada hora. A temporização e o alcance de reposição ou acréscimo constituem opções de projeto estratégicas que dependem do tempo disponível e das necessidades de negócios. Sistemas mais complexos podem manter um histórico e uma pista de auditoria de todas as mudanças sofridas pelos dados.

## 3.3 Tratamento de dados para um armazém de dados

Os dados são a principal fonte para a mineração de dados, entretanto estes dados podem ser de diferentes tipos e origens, levando com que ocorram distorções na mineração. Sendo assim para se trabalhar com os dados são necessários algumas etapas para o preparo dos mesmos para o processo de mineração a primeira etapa é a de pré-processamento.

> A etapa de pré-processamento, no processo de descoberta de conhecimento em banco de dados compreende a aplicação de várias técnicas para captação, organização, tratamento e a preparação dos dados. É uma etapa que possui fundamental relevância nesse processo. [17, GPEA].

Compreende desde a correção de dados errados até o ajuste da formatação dos dados para os algoritmos de mineração de dados que serão utilizados.

**Algumas das principais funções da etapa de pré-processamento dos dados**

- **Seleção de atributos** A seleção de atributos é uma etapa da fase de pré-processamento do processo de descoberta de conhecimento em banco de dados. Como o próprio nome já diz, o objetivo é escolher um subconjunto de atributos (também conhecidos como variáveis) ou criar outros atributos que substituam um conjunto deles a fim de reduzir a dimensão do banco de dados. Com essa redução de dimensão, reduz-se a complexidade do banco de dados e assim o tempo de processamento para extrair dele algum conhecimento. Além disso, atributos desnecessários podem causar ruído no resultado final e isto pode ser evitado com a aplicação de técnicas de seleção de atributos.
- **Limpeza dos dados** Abrange qualquer tratamento realizado sobre os dados selecionados de forma a assegurar a qualidade (completude, veracidade e integridade) os fatos por eles representados. Informações ausentes, errôneas ou inconsistentes nas bases de dados devem ser corrigidas de forma a não comprometer a qualidade dos modelos de conhecimento a serem extraídos ao final do processo.
- **Discretização** Alguns algoritmos de mineração de dados, especialmente os algoritmos de classificação, requerem que os dados estejam na forma de atributos categorizados. Assim, muitas vezes é necessário transformar um atributo contínuo em categórico.
- **Binarização** Algoritmos utilizados para descoberta de padrões de associação requerem que os dados estejam na forma de atributos binários. Assim, muitas vezes tanto os atributos contínuos quanto os discretos necessitam ser transformados em um ou mais atributos binários.
- **Construção de atributos** Essa operação consiste em gerar novos atributos a partir dos atributos existentes. A importância desse tipo de operação é justificada pois novos atributos, além de expressarem relacionamentos conhecidos entre atributos existentes, podem reduzir o conjunto de dados simplificando o processamento dos algoritmos de mineração de dados.
- **Transformação de variáveis** Se refere a uma transformação que seja aplicada a todos os valores de um atributo. Em outras palavras, para cada objeto, a transformação é aplicada ao valor do atributo para aquele objeto. Uma transformação que pode-se citar é a normalização dos dados, que consiste em ajustar a escala dos valores de cada atributo de forma que os valores fiquem em pequenos intervalos, tais como de -1 a 1 ou de 0 a 1. Tal ajuste se faz necessário para evitar que alguns atributos, por apresentarem uma escala de valores maior que outros, influenciem de forma tendenciosa determinados métodos de mineração de dados.

Um conjunto de dados muitas vezes pode ser visto como uma coleção de objetos de dados, outros nomes para um objeto de dados são registros, ponteiros, vetores, padrões, eventos, casos, exemplos, observações ou entidades. Por sua vez, objetos de

dados são descritos por um número de atributos que capturam as características básicas de um objeto, como a massa de um objeto físico ou o tempo no qual um evento tenha ocorrido. Conforme KUMAR et al. (2009) outros nomes para um atributo são variável, característica, campo, recurso ou dimensão.

> Exemplo: Muitas vezes, um conjunto de dados é um arquivo, no qual os objetos são registros (ou linhas) no arquivo e cada campo (ou coluna) corresponde a um atributo. Por exemplo, a seguir mostra um conjunto de dados que consiste de informações sobre alunos. Cada linha corresponde a um aluno e cada coluna é um atributo que descreve algum aspecto de um aluno, como a média (GPA) ou número de identificação (ID). (KUMAR et al., 2009, p. 3). [40].

**Tabela 3.2:** *Dados de exemplo contendo informações de alunos. Fonte: [40]*

| **ID Aluno** | **Ano** | *Medida GPA* | ⋮ |
|---|---|---|---|
| | ⋮ | | |
| 1034262 | Terceiro | 3,24 | ⋮ |
| 1052663 | Segundo | 3,51 | ⋮ |
| 1082246 | Primeiro | 3,62 | ⋮ |
| ⋮ | ⋮ | ⋮ | ⋮ |

Embora conjuntos de dados baseados em registros sejam comuns, tanto em arquivos horizontais ou quanto em sistemas de bancos de dados relacionais, há outros tipos importantes de conjuntos de dados e sistemas para armazenamento de dados. Neste sentido, será abordado alguns dos tipos de conjuntos de dados que são comumente encontrados na mineração dedados.

### 3.3.1 Atributos e medidas

O que é um atributo? Segundo KUMAR et al. [40] “Um atributo é uma propriedade ou característica de um objeto que pode variar, seja de um objeto para outro ou de tempo para outro”.

Por exemplo, a cor dos olhos varia de pessoa para pessoa, enquanto que a temperatura de um objeto varia com o tempo. Observe que a cor dos olhos é um atributo simbólico com um pequeno número de valores possíveis marrom, preto, azul, verde, castanho, etc., enquanto que a temperatura é um atributo numérico com um número potencialmente ilimitado de valores. No nível mais básico, os atributos não se relacionam

com números ou símbolos. Entretanto, para discutir e analisar com maior precisão as características de objetos atribuímos números ou símbolos a eles.

Com esta fundamentação, torna-se possível discutir o tipo de um atributo, um conceito importante para determinar se uma técnica de análise de dados específica é consistente com um determinado tipo de atributo.

### 3.3.2 O tipo de um atributo

A determinação do tipo de um atributo é importante para que se possa determinar qual ação será mais adequada a ele e até mesmo se ele deve fazer parte de uma ação, conforme afirma KUMAR et al. (2009) a seguir:

> O tipo de atributo nos informa quais propriedades do mesmo são refletidas nos valores usados para medi-lo. Conhecer o tipo de um atributo é importante porque nos informa quais propriedades dos valores medidos são consistentes com as propriedades correspondentes do atributo e, portanto, evita que executemos ações tolas, como calcular a média das ID's dos funcionários. Observe que é comum se referir ao tipo de um atributo como o tipo de uma escala de medição. [40].

Exemplo: (Idade e Número de ID do Funcionário). Dois atributos que poderiam ser associados a um funcionário são sua ID e idade (em anos), ambos os atributos podem ser representados como números inteiros. Entretanto, embora seja razoável falar em média de idade de um funcionário, não faz sentido falar em média de ID do funcionário. De fato, o único aspecto dos funcionários que pretende-se capturar com o atributo ID é que eles são distintos e consequentemente, a única operação válida para IDs de funcionários é testar se são iguais.

Para o atributo idade, as propriedades dos números inteiros usadas para representar a idade são bem as propriedades do atributo. Mesmo assim, a correspondência (relação de inteiros e idades) não é completa, já que, por exemplo, idades têm um máximo, enquanto que números inteiros não.

### 3.3.3 Tipos de conjunto de dados

**Dados de registro:** Grande parte do trabalho de mineração de dados supõe que o conjunto de dados seja uma coleção de registros (objetos de dados), cada um dos quais consistindo de um conjunto fixo de campos de dados (atributos). Dados em registros geralmente são armazenados em arquivos horizontais ou em bancos de dados relacionais. Bancos de dados relacionais são certamente mais do que uma coleção de registros, porém a mineração de dados muitas vezes não usa algumas das informações

adicionais disponíveis em um banco de dados relacional, ao invés disso, o banco de dados serve como um lugar conveniente para encontrar registros.

**Dados de transação:** É um tipo especial de dados em registros, onde cada registro (transação) envolve um conjunto de itens. Considere uma mercearia, o conjunto de produtos comprados por um cliente durante uma ida à mercearia constitui uma transação, enquanto que os produtos individuais que foram comprados são os itens.

Este tipo de dado é chamado de dado de cesta de mercado porque os itens em cada registro são os produtos em uma "cesta de mercado"de uma pessoa. Dados de transação é uma coleção de conjuntos de itens, mas podem ser vistos como um conjunto de registros cujos campos são atributos assimétricos.

Com maior frequência, os atributos são binários, indicando se um item foi comprado ou não, porém, mais comumente, os atributos podem ser discretos ou contínuos, como o número de itens comprados ou a quantia gasta nesses itens.

**Tabela 3.3:** *Dados em registros. Fonte: [40]*

| TID | Reembolso | Estado Civil | Renda tributável | Tomador de empréstimo em dívida |
|---|---|---|---|---|
| 1 | Sim | Solteiro | 125K | Não |
| 2 | Não | Casado | 100K | Não |
| 3 | Não | Solteiro | 70K | Não |
| 4 | Sim | Casado | 120K | Não |
| 5 | Não | Divorciado | 95K | Sim |
| 6 | Não | Casado | 60K | Não |
| 7 | Sim | Divorciado | 220K | Não |
| 8 | Não | Solteiro | 85K | Sim |
| 9 | Não | Casado | 75K | Não |
| 10 | Não | Solteiro | 90K | Sim |

**Tabela 3.4:** *Dados de transação. Fonte: [40]*

| TID | Itens |
|---|---|
| 1 | Pão, Refrigerante, Leite |
| 2 | Cerveja, Pão |
| 3 | Cerveja, Refrigerante, Fralda, Leite |
| 4 | Cerveja, Pão, Fralda, Leite |
| 5 | Refrigerante, Fralda, Leite |

CAPÍTULO 4

# Mineração de dados

## 4.1 Descoberta de conhecimento em banco de dados

A mineração de dados, em inglês *data mine* sigla *DM*, é um processo de busca em grandes massas de dados por padrões que possam ser utilizados para a geração de conhecimento, entretanto ele por si só não pode ser definido como responsável pelo processo de geração deste conhecimento. O processo como um todo é conhecido como descoberta de conhecimento em banco de dados, em inglês *knowledge discovery in database* sigla *KDD*.

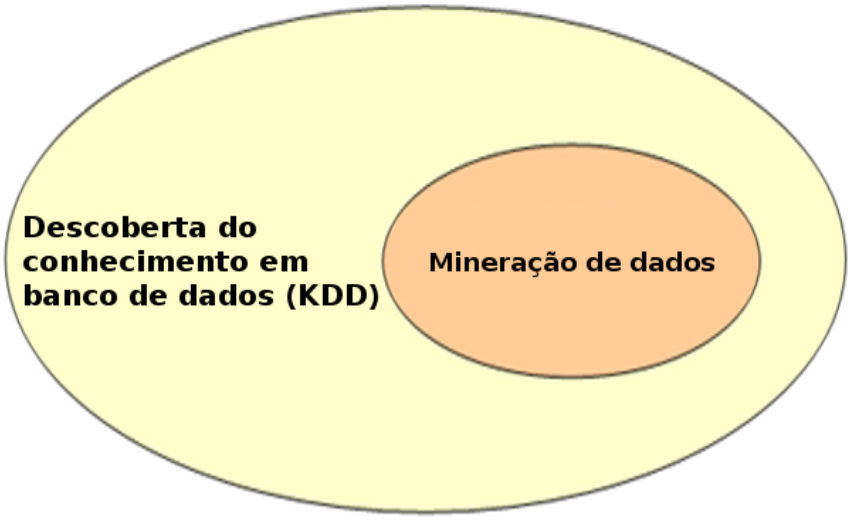


**Figura 4.1:** *Relação entre KDD e Mineração de dados. fonte: O autor*

A mineração de dados veio da impossibilidade do ser humano de analisar manualmente todas as informações geradas por uma grande massa de dados. Tem por objetivo, obter informações que não são muito óbvias, que não sejam possíveis de se analisar manualmente. Podemos associar a mineração de dados a eficiência e a vantagem competitiva. Qualquer instituição que consiga aplicar técnicas de mineração de dados irá ter uma grande vantagem competitiva em relação ao seu concorrente. Não envolve

somente corporações na concepção de negócio, mas também grandes redes de pesquisas ligadas a universidades, medicina e outras instituições governamentais.

O processo de descoberta de conhecimento em banco de dados poder ser aplicado a qualquer área que tenha um volume de dados que possa ser explorado, entretanto nem todas as tarefas de descoberta da informação podem ser consideradas mineração de dados, como por exemplo uma simples busca em bancos de dados por mecanismos rotineiros, é nada mais do que tarefas simples de recuperação de dados que são suportadas por sistemas de informação simples.

**São alguns exemplos de mineração de dados**

O governo dos Estados Unidos da América se utiliza do mineração de dados há bastante tempo para identificar padrões de transferência de fundos internacionais que se parecem com lavagem de dinheiro do narcotráfico.

Vendas cruzadas podem ser realizadas com facilidade se um banco de dados com informações sobre o passado do cliente existir. Sabendo as necessidades e gostos do cliente, novos produtos podem ser oferecidos pela empresa mantendo a fidelidade do cliente que não precisa ir buscar o produto em outro local.

Devido à competição empresarial, clientes mudam de empresa com facilidade. A mineração de dados pode ser usado para verificar porque os clientes trocam uma empresa por outra e oferecer serviços, vantagens e ofertas que evitem esta fuga de clientes. É mais fácil manter um cliente do que adquirir um novo. Com o processo de mineração de dados, pode-se localizar que oferta fazer a que cliente para mantê-lo na empresa ou mesmo localizar os clientes que podem sair da empresa sem representar prejuízo.

Na medicina já é possível a criação e manutenção de grandes bancos de dados com informação sobre sintomas, resultados de exames, diagnósticos, tratamentos e curso das doenças para cada paciente. A mineração destes dados pode fornecer conhecimento novo como, por exemplo, a relação entre algumas doenças e certos perfis profissionais, sócio culturais, hábitos pessoais e local de moradia. Estas relações são utilizadas para melhor entendimento das doenças e seus tratamentos.

Outras áreas que por característica própria de seu estudo, como a astronomia e a geologia, geram e acumulam enormes quantidades de dados já estão utilizando intensamente o mineração de dados para descobrir conhecimento novo que a olho nu não seriam facilmente percebidos.

## 4.2 Processo de Mineração de dados

A mineração de dados é o processo de descoberta automática de informações úteis em grandes depósitos de dados. As técnicas de mineração de dados são organiza-

das para agir sobre grandes bancos de dados com o intuito de descobrir padrões úteis e recentes que poderiam de outra forma permanecer ignorados. Elas também fornecem capacidade de previsão do resultado de uma observação futura. Trata-se de um conjunto de técnicas reunidas da estatística e da inteligência artificial com o objetivo de descobrir conhecimento novo que por ventura esteja escondido em grades massas de dados armazenados em bancos de dados.

Apesar de algumas tarefas de descoberta da informação serem importantes e envolver o uso de algoritmos e estruturas de dados sofisticadas, a explicação é que estas tarefas se baseiam em técnicas tradicionais da tecnologia da informação e em recursos óbvios dos dados para criar estruturas de índice para organizar e recuperar de forma eficiente as informações. Contudo, a mineração de dados tem sido usada para melhorar sistemas de recuperação de informações, e segundo [40, KUMAR et al.], a mineração de dados é uma parte integral da descoberta de conhecimento em bancos de dados, é o processo geral de conversão de dados brutos em informações úteis.

CAPÍTULO 5

# Passos do processo de descoberta em banco de dados

O processo de minerção de dados é divido basicamente em sete passos, em uma ponta temos a entrada de dados e na outra, a saída, temos o conhecimento.

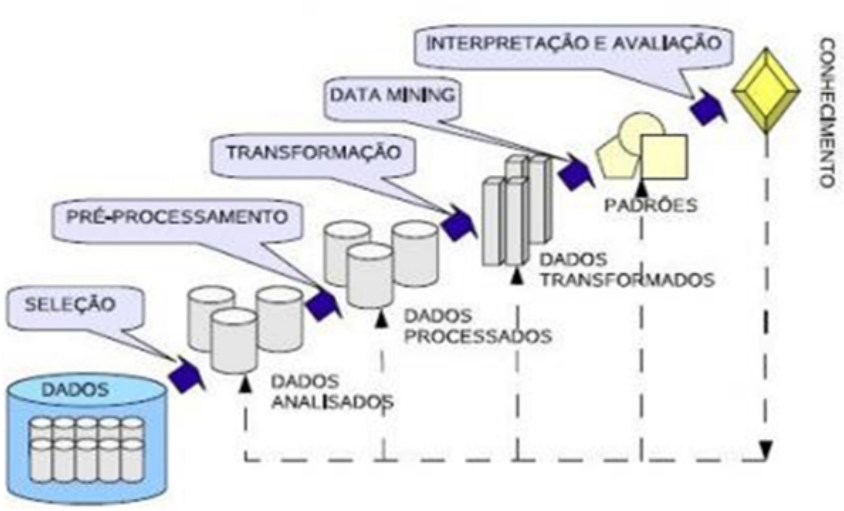


**Figura 5.1:** *Fluxo de funcionamento do processo de descoberta do conhecimento.[29]*

1. Dados
2. Seleção
3. Pré-processamento
4. Transformação
5. Mineração de dados
6. Interpretação e avaliação
7. Conhecimento

Esses passos não são lineares, é um processo adaptativo. É muito importante o trabalho de base: Seleção, pré-processameto, e transformação antes de entrarmos nos algoritmos propriamente dito na mineração de dados. A maioria dos algoritmos, dependo do volume de dados que será trabalhado, geralmente e quase sempre, a unidade de medida de tempo é dada em dias devido a grande quantidade de dados a serem analisados,

dificilmente se trabalhara com uma unidade de tempo menor do que dias. Isso destaca mais ainda a importância dos passos anteriores ao da mineração de dados em si. Em todas essas atividades é importante lembrar que em cada uma delas não existe apenas um algoritmo que resolve o problema, vc tem um conjunto de algoritmos, e as vezes em uma determinada aplicação um algoritmo é melhor que outro dentro de uma mesma atividade.

## 5.1 Dado

Valor bruto, como visto no capítulo 2 seção 2.1.1, são a matéria prima do processo de descoberta do conhecimento.

## 5.2 Seleção

Como o próprio nome diz Selecionar dados que serão analisados, algumas perguntas devem ser feitas para se saber quais dados serão selecionados e o que deve ser encontrado nesses dados. Ele pega informações de diversas fontes, arquivos texto, planilhas eletrônicas etc. e os concentra em uma única base. Como resultado: Dados selecionados.

## 5.3 Pré-processamento

Nada mais é do que reduzir o tamanho do banco de dados. Exitem diversos dados que não são necessários para a mineração. É importante que antes do inicio do processo seja estabelecido qual o objeto da mineração. Que esteja dentro de um contexto, ter uma visão geral do que se quer encontrar. O processo de mineração responde a uma série de perguntas pré estabelecidas. E como temos um contexto, nessa fase se eliminam os dados que não estão dentro do contexto estabelecido. Os algoritmos de mineração geralmente são muito demorados, se temos informações que não atendem ao contexto, o processo será mais lento e o desempenho será muito baixo. Como resultado temos os dados processados.

## 5.4 Transformação

Nesse passo podemos usar vários tipos de algoritmos, cada tipo tem um formato de dado de entrada. Nesse passo serão transformados os dados para o algoritmo específico que será utilizado no próximo passo. (Você transformar um quadrado em um cubo pq futuramente iremos precisar de um cubo por exemplo ) É um dos passos mais complexos no processo pq muitas vezes fazemos todos os passos apresentados até o momento, e

descobrimos no momento de executar o passo de DM que alguma coisa ficou errada lá pra trás, é onde falta o ponto e virgula da programação na hora da compilação. Dependendo do caso, é necessário voltar no incio e restartar o processo com a nova informação. Como resultado temos os dados transformados e preparados para a aplicação dos algoritmos de mineração.

## 5.5 Mineração de dados

Após sair com os dados processados e transformados, aplicamos os algoritmos para a extração do conhecimento. Exitem diversos algoritmos que podem ser aplicados. A escolha do algoritmo é em função de "perguntas"que devem ser formuladas na seleção dos dados, qual o objetivo desse data mine? Qual a pergunta que ele deverá responder? Basicamente as perguntas, que eles chamam de atividades de data mine, podem ser de dois tipos: Atividade Preditivas e Atividades descritivas. Como dizem, nessa fase é onde o show acontece. Atividades preditivas: Você tem basicamente dois tipos de algoritmos: Classificação e regressão. Classificação: são diversas categorias pré estabelecidas, vc tem um novo item e quer prever em qual categoria esse novo item se encaixa. Regressão: Quando vc quer prever uma variável numérica. Ex. vc quer prever que nota determinada pessoa daria a um filme baseando nos votos anteriores dela. Atividades descritivas São basicamente três tipos: Regras de associação: Descrevem se exitem alguma dependência entre as variáveis, associação entre dois itens Sumarização: que descreve um conjunto de dados Clusterização: divide o universo de itens em clusters, em grupos, colocar cada coisa em seu devido lugar.

## 5.6 Interpretação e avaliação

Normalmente nessa etapa, a TI trabalha associado a especialistas da área de aplicação do negócio. pode ser a melhor ou a pior etapa do processo. Nessa fase é que entregamos o produto aos patrocinadores, nela são vistos os padrões e resultados da mineração.

## 5.7 Conhecimento

CAPÍTULO 6

# Conclusão

A Mineração de dados é uma das áreas do conhecimento mais deslumbrante no contexto da Tecnologia da Informação e Comunicação. A principal contribuição deste estudo foi a identificação de ferramentas que permitam a descoberta de conhecimento e a rápida análise em um grande volume de dados, que cresce a cada dia nas organizações ou até mesmo na Internet.

Utilizando ferramentas desenvolvidas com esta finalidade e complexidade específica, junto com conhecimento do funcionamento dos principais algoritmos de Mineração de dados, acredita-se que é possível analisar e gerar conhecimento que até então estavam incobertos, junto à milhares ou milhões de registros em bancos de dados corporativos, mostrando assim novas formas de atuação a partir da descoberta feita.

Com a crescente massa de dados, torna-se evidente o uso de tecnologias mais avançadas para o fornecimento de informações para apoiarem no processo de tomada de decisão. Neste contexto, o uso de técnicas e algoritmos relacionados a Mineração de dados se fazem necessários para viabilizar a geração de conhecimentos e, dessa forma, fornecer subsídios para os gestores maximizarem seus investimentos e alocarem da melhor forma seus recursos observando o custo-benefício. Para isto, percebe-se que existem um conjunto de técnicas, ferramentas e práticas utilizadas em Inteligência de negócio e Mineração de dados, cada qual com seus conceitos, boas práticas, sistemas de computador e os algoritmos mais utilizados para estas finalidades. Assim, destaca-se a importância das organizações tomarem conhecimento da sua realidade para que tenham sucesso e alcance os resultados almejados.

# Referências Bibliográficas


[1] ABERNETHY, M. **Mineração de dados com weka**. IBM. Acesso em: 15 de dezembro de 2013.

[2] ALECRIM, E. **Redes neurais artificiais**. Infowester. Acesso em: 21 de dezembro de 2013.

[3] ALMEIDA, E. C. D. **Estudo de viabilidade de uma plataforma de baixo custo para data warehouse**. arxiv.org. Acesso em: 14 de dezembro de 2013.

[4] BERNERS-LEE, R.; SWICK, T. **Semantic web development**. www.w3.org. Acesso em: Agosto de 2013.

[5] BOGORNY, V. **Algoritmos e Ferramentas para Descoberta de Conhecimento em Bases de Dados Geográficos**. Biblioteca UFRGS, Porto Alegre, 2003.

[6] CONSULTORIA, C. **O que é data mart?** Cetax. Acesso em: 19 de dezembro de 2013.

[7] CORPORATION, I. **Ibm spss modeler**. www.pse.pt. Acesso em: 14 de dezembro de 2013.

[8] DA COSTA, W. **A empresa digital**. Wladi Fatec. Acesso em: 08 de dezembro de 2013.

[9] DA COSTA, W. **Lojas renner melhoram resultados de suas campanhas de marketing direto com análise preditiva da spss**. Dmss. Acesso em: 10 de dezembro de 2013.

[10] DA SILVA, F. V. **Decisões com B.I. (Business Intelligence)**. Ciência Moderna, Rio de Janeiro, 2008.

[11] DARIO, B. R. **Data mart**. Dataprix. Acesso em: 20 de dezembro de 2013.

[12] DE ALMEIDA; ANTÔNIO, A. L. **Sistemas de informação geográfica dicionário ilustrado editora hucitec**. www.uefs.br. Acesso em: 08 de Dezembro de 2013.

[13] DE CARVALHO, L. A. V. **DATAMINING: A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração.** Ciência Moderna, Rio de Janeiro, 2005.

[14] DE COME, G. **Os metadados no ambiente de data warehouse**. *IV SEMEAD*, 1999.

[15] DE COMPUTAÇÃO CIENTÍFICA, L. N. **Tutorial de redes neurais aplicação em bioinformática**. www.lncc.br. Acesso em: 21 de dezembro de 2013.

[16] DE JESUS, F. **Pensamento e Planejamento Estratégico: uma abordagem competitiva**. Edição do Autor, Goiânia, 2011.

[17] DE PESQUISA EM ENGENHARIA DE ALGORITMO GPEA, G. **Pré-processamento em data mining**. www.din.uem.br. Acesso em: 05 de dezembro de 2013.

[18] DO NASCIMENTO, L. A. **Data warehouse arquitetura e implantação**. Nascimento. Acesso em: 10 de novembro de 2013.

[19] E EDUARDO O C, O. M. **Sistemas de informação e sistemas de apoio á decisão**. Ebah. Acesso em: 10 de dezembro de 2013.

[20] E FREITAS (H.), P. M. **Características desejáveis de um eis enterprise information system rumo á proatividade**. www.lume.ufrgs.br. A cesso em: 10 de dezembro de 2013.

[21] F, C. **Classificação de sistemas de informação**. Tecspace. Acesso em: 06 de Janeiro de 2014.

[22] FURTADO, L. S. C. R. S. S. **Clusterização em mineração de dados**. www.ic.uff.br. Acesso em: 20 de novembro de 2013.

[23] IBL. **Extração, transformação e carga**. www.infobras.com.br. Acesso em: 15 de dezembro de 2013.

[24] INMON, W. H. **Building the Data Warehouse Third Edition**. John Wiley and Sons Inc, New York, 2002.

[25] JUNIOR, M. A. **Oltp x olap**. MarquinhosNet. Acesso em: 10 de dezembro de 2013.

[26] LORENZI, C. A. **Arquiteturas de data warehousing parte 2**. Blog do Lito. Acesso em: 30 de dezembro de 2013.

[27] MICROSOFT. **Algoritmos de mineração de dados (analysis services mineração de dados)**. MSDN Microsoft. Acesso em: 15 de dezembro de 2013.

[28] MICROSOFT. **Dados hierárquicos (sql server)**. Technet Microsoft. Acesso em: 08 de dezembro de 2013.

[29] MORELLATO, L. **Mineração de dados e web semântica**. iMasters. Acesso em: 13 de dezembro de 2013.

[30] PANAZZO E MARIA LUISA VAZ, S. **Navegando pela historia**. Aprendendo a gostar de historia. Acesso em: 16 de dezembro de 2013.

[31] PERZEPIORSKI, E. **Análise de crédito bancário com o uso de data mining: Redes neurais e árvores de decisa o**. www.ppgmne.ufpr.br. Acesso em: 20 de dezembro de 2013.

[32] RENATO. **Data warehouses: Fundamentos, ferramentas e tendâncias atuais**. www.inf.ufsc.br. Acesso em: 13 de dezembro de 2013.

[33] ROCHA, F. L. **Bussines inteligence: arquiteturas e tecnologias**. Felipelirarocha. Acesso em: 14 de dezembro de 2013.

[34] SHARDA, E. T. J. E. D. K. R. **Business Intelligence: Um enfoque gerencial para a inteligência do negócio**. Bookman, Porto Alegre, 2009.

[35] SIDEMAR, J. O. **Porque business intelligence?** iMasters. Acesso em: 08 de Janeiro de 2014.

[36] SOUZA, M. **Ferramentas ol ap**. iMasters. Acesso em: 13 de Dezembro de 2013.

[37] SYBASE. **Sybase etl**. Sybase. Acesso em: 19 de dezembro de 2013.

[38] SYBASE, S. **Sap sybase iq**. Sybase. Acesso em: 14 de dezembro de 2013.

[39] UNIPRESS. **Bam business activity monitoring: inteligência de negócios em tempo real**. Unipress. Acesso em: 10 de Janeiro de 2014.

[40] VIPIN, P.-N. M. **Introdução ao Data Mining Mineração de dados**. Ciência Moderna, Rio de Janeiro, 2009.

[41] W, R. M. G. **Princípios de Sistemas de Informação: Uma abordagem gerencial**. Pioneira Thomson Learning, São Paulo, 2006.