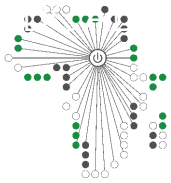


Universidade Virtual Africana

INFORMÁTICA APLICADA: ITI 4302

# ARMAZENAMENTO E MINERAÇÃO DE DADOS

Marcelo Correia

# Prefácio

A Universidade Virtual Africana (AVU) orgulha-se de participar do aumento do acesso à educação nos países africanos através da produção de materiais de aprendizagem de qualidade. Também estamos orgulhosos de contribuir com o conhecimento global, pois nossos Recursos Educacionais Abertos são acessados principalmente de fora do continente africano.

Este módulo foi desenvolvido como parte de um diploma e programa de graduação em Ciências da Computação Aplicada, em colaboração com 18 instituições parceiras africanas de 16 países. Um total de 156 módulos foram desenvolvidos ou traduzidos para garantir disponibilidade em inglês, francês e português. Esses módulos também foram disponibilizados como recursos de educação aberta (OER) em oer.avu.org.

Em nome da Universidade Virtual Africana e nosso patrono, nossas instituições parceiras, o Banco Africano de Desenvolvimento, convido você a usar este módulo em sua instituição, para sua própria educação, compartilhá-lo o mais amplamente possível e participar ativamente da AVU Comunidades de prática de seu interesse. Estamos empenhados em estar na linha de frente do desenvolvimento e compartilhamento de recursos educacionais abertos.

A Universidade Virtual Africana (UVA) é uma Organização Pan-Africana Intergovernamental criada por carta com o mandato de aumentar significativamente o acesso a educação e treinamento superior de qualidade através do uso inovador de tecnologias de comunicação de informação. Uma Carta, que estabelece a UVA como Organização Intergovernamental, foi assinada até agora por dezenove (19) Governos Africanos - Quênia, Senegal, Mauritânia, Mali, Costa do Marfim, Tanzânia, Moçambique, República Democrática do Congo, Benin, Gana, República da Guiné, Burkina Faso, Níger, Sudão do Sul, Sudão, Gâmbia, Guiné-Bissau, Etiópia e Cabo Verde.

As seguintes instituições participaram do Programa de Informática Aplicada: (1) Université d'Abomey Calavi em Benin; (2) Université de Ougagadougou em Burkina Faso; (3) Université Lumière de Bujumbura no Burundi; (4) Universidade de Douala nos Camarões; (5) Universidade de Nouakchott na Mauritânia; (6) Université Gaston Berger no Senegal; (7) Universidade das Ciências, Técnicas e Tecnologias de Bamako no Mali (8) Instituto de Administração e Administração Pública do Gana; (9) Universidade de Ciência e Tecnologia Kwame Nkrumah em Gana; (10) Universidade Kenyatta no Quênia; (11) Universidade Egerton no Quênia; (12) Universidade de Addis Abeba na Etiópia (13) Universidade do Ruanda; (14) Universidade de Dar es Salaam na Tanzânia; (15) Universite Abdou Moumouni de Niamey no Níger; (16) Université Cheikh Anta Diop no Senegal; (17) Universidade Pedagógica em Moçambique; E (18) A Universidade da Gâmbia na Gâmbia.

Bakary Diallo

O Reitor

Universidade Virtual Africana

# Créditos de Produção

## Autor

Marcelo Correia

## Par revisor(a)

Martina Barros

## UVA - Coordenação Académica

Dr. Marilena Cabral

## Coordenador Geral Programa de Informática Aplicada

Prof Tim  Mwololo Waema

## Coordenador do módulo

Robert Oboko

## Designers Instrucionais

Elizabeth Mbasu

Benta Ochola

Diana Tuel

## Equipa Multimédia

Sidney McGregor

Barry Savala

Edwin Kiprono

Kelvin Muriithi

Victor Oluoch Otieno

Michal Abigael Koyier

Mercy Tabi Ojwang

Josiah Mutsogu

Kefa Murimi

Gerisson Mulongo

# Direitos de Autor



# Apoiado por

Projeto Multinacional II da UVA financiado pelo Banco Africano de Desenvolvimento.

# Índice

# Descrição Geral do Curso

Com este módulo pretende se mostrar aos alunos e profissionais da área as técnicas e forma de gerir, armazenar e minerar os dados, para que as empresas possam criar vantagem competitivas.

## Pré-requisitos

Para este curso é suposto que os (as) estudantes disponham de: conhecimento básico de sistemas de informação e do seu funcionamento: tenham n ações básicas de base de dados e

domínio das ferramentas de internet

### *Volume horário/Tempos*

Este módulo deve ser estudado em120 horas repartidas entre leituras, actividades práticas, trabalhos dirigidos e avaliações formativas e sumativas.

Para o estudo das 4 unidades são programadas 20 horas. Para as actividades práticas, 20 horas. Para as consultas dos links e recursos, 20 horas. Para os trabalhos dirigidos são 20 horas e, para as avaliações formativas e sumativa, 40 horas.

## Materiais

Os materiais necessários para completar este curso incluem:

1. CD-Rom
2. Livros
3. E-books
4. Tutoriais
5. Computadores
6. Internet
7. Vídeo aulas

Não obstante os (as) estudantes podem recorrer a outros materiais ou softwares suplementares como forma de reforçar a compreensão e realizar simulações.

## Objectivos do Curso

Após este curso o(a) estudante deverá ser capaz de: compreender o funcionamento e de desenvolver soluções de:

- Implementar um plano de armazenamento e mineração de dados;Definir conceitos e enumerar os fundamentos relacionados com a tecnologias Data Warehouse;
- Identificar Técnicas de extracção de informação.
- Desenvolver soluções para Desenhar e implementar um Data warehouse
- Identificar os componentes de um DW;
- Indicar e descrever os componentes da arquitectura de DW;
- Diferenciar DW de Data Smart;
- Diferenciar Data warehouse de Data mining
- Caracterizar os sistemas de Data mining;
- Diferenciar e utilizar as técnicas e tarefas associadas a mineração de dados:
- Diferenciar as Técnicas de consultas as sistemas transacionais e Analitica (OLAP)

## Unidades

**Unidade 0: fundamentos data warehouse**

Com o avanço das novas tecnologias de informação e comunicação as informações passaram a ser armazenadas em diferentes meios e tornaram-se mais volumosas e heterogéneas, de forma que se produziu um caos informacional. Por isso é de extrema importância conhecer os conceitos associados ao processo de armazenamento e mineração de dados.

**Unidade 1: desenvolvimento do data warehouse**

Nessa unidade é de extrema importância que os alunos conseguam projectar um modelo dimensional para armazenar e minerar dados.

**Unidade 2: projecto de carregar dados em um data warehouse**

Depois de ter projectado e criado um Data warehouse os (as) estudantes devem estar preparados para carregar os dados oriundos de fontes heterogéneos numa base de dados centralizada, que se denomina de Data Warehouse.

**Unidade 3: extraindo informações do data warehouse**

Os dados povoados no armazém devem ser extraídos, para ajudar os decisores na tomada de decisão. Para isso deve ser utilizado ferramentas que permitam análise dos dados agregados, ferramentas OLAP.

**Unidade 4: Mineração de Dados**

Associado aos dados estão informações ocultas que precisam de ser utilizadas usando as técnicas e tarefas de mineração, para que possa ser analisadas as tendências associados a esses dados e transforma-la em conhecimento, que é considerado o melhor elemento da competitividade.

## Avaliação

Em cada unidade encontram-se incluídos instrumentos de avaliação formativa a fim de acompanhar o progresso do(a)s estudantes.

No final de cada módulo são apresentados instrumentos de avaliação sumativa, tais como testes e trabalhos finais, que compreendem os conhecimentos construídose as competências desenvolvidas ao estudar este módulo.

A implementação dos instrumentos de avaliação sumativa fica ao critério da instituição que oferece o curso. A estratégia de avaliação sugerida é a seguinte:

| | | |
|---|---|---|
| 1 | Teste | 35% |
| 2 | Teste 2 | 30% |
| 3 | Fichas | 35% |

## Calendarização

| **Unidade** | **Temas e Actividades** | **Estimativa do tempo** |
|---|---|---|
| FUNDAMENTOS DATA WAREHOUSE | Definição característica e estrutura<br>Sistema de Apoio a Decisão<br>Visão geral dos componentes de Data Warehouse<br>Data Warehouse e Data Mart | 10 h |

| | | |
|---|---|---|
| DESENVOLVIMENTO DO DATA WAREHOUSE | Ciclo de Vida do Data Warehouse<br>PROJECTO DATA WAREHOUSE<br>Definição dos Requisitos de Negócio PARA DW;<br>GRANULARIDADE<br>DATA WAREHOUSE DISTRIBUIDODesenho da Arquitetura DE DATA WAREHOUSE;<br>COMPONENTES E SUAS CARACTERISTICAS METADADOS:GERENCIA,ARMAZENAMENTO E INTEGRAÇÃO<br>Seleção dos Produtos e Ferramentas; | 4h |
| PROJECTO DE CARREGAR DADOS EM UM DATA WAREHOUSE | Princípios de modelagem<br>Esquema de estrela e floco de neve<br>Extracção, transformação e carga de dados em data warehousing- ETL<br>Qualidade de dados armazenados<br>Ferramenta para extracção, transformação e carga de dados | |
| EXTRAINDO INFORMAÇÕES DO DATA WAREHOUSE | Potencial de informações num data warehouse<br>Extraindo e transformando dados;<br>Análise Multi-dimensional e OLAP<br>Modelos Multi-Dimensionais de Dados<br>Construção de Cubos Multi-Dimensionais<br>Interrogação Multi-Dimensional de Dados<br>Ferramentas e operações OLAP<br>MODELOS OLAP<br>MOLAP<br>ROLAP | 10h |

| | | |
|---|---|---|
| MINERAÇÃO DE DADOS | Conceitos de mineração e descoberta de conhecimento<br>Técnicas de mineração de dados e algoritmo<br>Análise de associação<br>Classificação e previsão de dados<br>Segmentação e análise de cluster<br>Aplicação de mineração de dados<br>Ferramentas de mineração de dados | |

## Leituras e outros Recursos

- CD-Rom
- Livros
- E-books
- Tutoriais
- Computadores
- Internet

Não obstante os alunos podem recorrer a outros materiais ou softwares suplementares como forma de reforçar a compreensão e realizar simulações.

**Unidade 0**

*Leituras e outros recursos obrigatórios:*


- The Data Warehouse Toolkit, 3rd Edition ,Ralph Kimball and Margy Ross.
- The Data Warehouse Lifecycle Toolkit, 2nd Edition, Ralph Kimball and the Kimball Group
- The Data Warehouse ETL Toolkit Ralph Kimball.
- Como Construir o Data Warehouse - Inmon, W (8535201416),Editora: CAMPUS
- Building the Data Warehouse,W. H. Inmon
- Gerenciando Data Warehouse,W. H. Inmon editora: Makron Books,Ano: 1999
- Tecnologia e Projeto de Data Warehouse, Felipe Nery Rodrigues Machado, 2007.
- Extração de Conhecimento de Dados João Gama, Ana Carolina Lorena, Katti Faceli, André Ponce de Leon Carvalho, Márcia Oliveira. Edições Sílabo, 2012. ISBN: 9789726186984.

- Data Mining: Practical Machine Learning Tools and Techniques, Ian H. Witten, Eibe Frank, Mark A. Hall, 3rd Edition, Prentice Hall, 2011, ISBN 0123748569.
- Data Warehouses and OLAP: Concepts, Architectures and Solutions, Robert Wrembel, Christian Koncilia, IGI Publishing, 2006, ISBN 1599043645.
- Predictive Data Mining: A Practical Guide, Sholom M. Weiss, Nitin Indurkhya, Morgan Kaufman, 1997, ISBN 1558604030.
- Data mining: um guia pratico /Emanuel Passos Rio de Janeiro 2005
- Introdução a mineração de dados/ Luis Paulo Vieira Braga 2ª Edição Revista e ampliada Rio de Janeiro 2005


*Leituras e outros recursos opcionais:*

# Unidade 0. Fundamentos Data Warehouse

## Introdução à Unidade

O propósito desta unidade é confrontar aos alunos com os conceitos relacionados com o processo de armazenamento de dados e avaliar o grau de compreensão dos conhecimentos que possui relacionados com este curso.

Os alunos devem saber usar, entender o significado, definir os conceitos relacionados com o mundo da armazenamento de dados sobretudo a distinguir o sistemas transaccional de sistema Data warehouse, passando pela sua história, os processos, as etapas de armazenamento de dados. Qualquer profissional que lida com o processo de armazenamento de dados tem de entender esses conceitos de forma que lhe ajuda na planificação e implementação dos projectos de armazenamento de dados.

## Objectivos da Unidade

Após a conclusão desta unidade, deverá ser capaz de:

1. Definir o conceito Dados
2. Definir o conceito Informação
3. Diferenciar Dados de Informação
4. Caracterizar a estrutura de armazenamento
5. Distinguir sistemas transaccionais de sistemas legados
6. Analisar e interpretar o processo de recuperação de informação
7. Identificar os

## Termos-chave

**Matriz:** Uma estrutura que contém uma coleta ordenada de elementos de dados em que cada elemento pode ser referenciado por sua posição ordinal na coleta. Todos os elementos em uma matriz têm o mesmo tipo de dados.

**Metadados:** Dados que descrevem as características de dados; dados descritivos.

**Classificação da informação:** Processo que permite agrupar as informações com as características  e propriedades idênticas, facilitando assim o seu tratamento e uso.

**Data warehouse:** Armazém de dados. É um sistema que guarda e organiza todas as informações que estão espalhadas por vários sistemas dentro de uma empresa. Com ele, os executivos podem obter informações sobre tudo e todos.

**Extração de Informação:** os termos considerados relevantes nos documentos são extraídos e convertidos em dados afim de que possam ser utilizados durante o processo de mineração.

**Filtragem de informação:** Sistema de RI que indexa perfis de informação que correspondem a necessidades de informação e compara com os documentos dum fluxo fazendo chegar aos utilizadores os documentos considerados relevantes pelo respectivo perfil.

**Mineração da Dados:** assim que a informação é armazenada de forma estruturada, a descoberta de informação é feita através da mineração sobre o banco de dados criado.

**Pesquisas:** (“queries”) são feitas por um único termo ou por composição de termos utilizando-se conectores lógicos (and, or, not), operadores relacionais (>, <, =) e meta-caracteres (*, ?)

**Recuperação de Informação:** localização e recuperação de documentos que podem ser relevantes a uma pesquisa. É necessário um sistema para filtrar esses documentos especificados pelo utilizador e indexar as palavras-chave encontradas.

# Armazenamento e mineração de dados

Desde dos tempos primórdios o homem preocupava-se em armazenar dados como forma de transmitir informações á geração vindoura. Os dados foram armazenados em formato de papéis mas com a evolução das tecnologias de informação vários são as fontes de dados.

Avanços na coleta de dados científicos (por exemplo, sensores remotos e satélites espaciais), processamento de código de barras e transações governamentais têm aumentado em muito o volume de dados. Aliados aos avanços na área de armazenamento, ao uso extensivo de sistemas de gerenciamento de banco de dados e tecnologia de data warehousing, a magnitude dos dados tem evoluído drasticamente. O banco de dados tem atingido dimensões astronómicos, produzindo terabytes de dados.

O processo de armazenamento ganhou uma nova dimensão com a introdução do conceito de data warehouse que permite integrar dados de fontes heterogéneos e fazer analise em diferentes perspectivas. Mas isso ficou ainda mais completo com a nova forma de mineração de dados – Data Mining que permite que os dados armazenados sejam analisados como formar de detectar as suas tendências.

## Dados

Numa primeira abordagem dado pode ser definido como INFORMAÇÃO BRUTA.

Definimos dado como uma sequência de símbolos quantificados ou quantificáveis. Portanto, um texto é um dado. De fato, as letras são símbolos quantificados, já que o alfabeto por si só constitui uma base numérica. Também são dados imagens, sons e animação, pois todos podem ser quantificados a ponto de alguém que entra em contacto com eles ter eventualmente dificuldade de distinguir a sua reprodução, a partir da representação quantificada, com o original. É muito importante notar-se que qualquer texto constitui um dado ou uma sequência de dados, mesmo que ele seja ininteligível para o leitor. Como são símbolos quantificáveis, dados podem obviamente ser armazenados em um computador e processados por ele.
Em nossa definição, um dado é necessariamente uma entidade matemática e, desta forma, puramente sintáctica. Isto significa que os dados podem ser totalmente descritos através de representações formais e estruturais.

## TIPOS DE DADOS

Existem diferentes tipos de dados como é ilustrado na Figura 1:

Quais os tipos de dados que temos hoje?

- Dados Estruturados
- Dados Semi-Estruturados
- Dados não-estruturados

### Dados estruturados

Dados organizados em blocos semânticos (relações),numa estrutura plana (tabelas)

- √ Dados de um mesmo grupo possuem as mesmas descrições (atributos)
- √ Descrições para todas as classes de um grupo possuem o mesmo formato (esquema)
- √ Dados mantidos em um SGBD são chamados de Dados Estruturados por manterem a mesma estrutura de representação (rígida), previamente projetada (esquema).

### Dados semi estruturados

Devido à heterogeneidade dos dados, muitos dados não são mantidos no SGBD

- √ Dados Web, por exemplo, apresentam uma organização bastante heterogênea.
- √ A alta heterogeneidade dificulta as consultas a estes dados
- √ Assim, estes dados são classificados como semi-estruturados
  - Não são estritamente tipados
  - Não são completamente não-estruturados


Figura 1: Estrutura de Dados

A tabela que se segue ilustra as principais diferenças entre os diferentes tipos de dados

| Dados Estruturados | Dados Semiestruturados | Dados Não Estruturados |
|---|---|---|
| Esquema predefinido | Nem sempre há esquemas | Não há esquemas |
| Estrutura regular | Estrutura irregular | Estrutura irregular |
| Estrutura independente dos dados | Estrutura imbutida do dados | Pode não ter estrutura alguma |
| Francamente evolutiva | Fortemente evolutiva (estrutura modifica com frequência) | Fortemente evolutiva (estrutura modifica com frequência) |
| Prescritivas (esquemas fechados e restrições de integridade | Estrutura descritiva | Estrutura descritiva |
| Distinção entre estrutura e dados é clara | Distinção entre estrutura e dados não é clara | Distinção entre estrutura e dados não é clara |
| Estrutura reduzida | Estrutura extensa (particularidade de cada dado, visto que cada um pode ter uma organização própria) | Estrutura extensa (particularidade de cada dado, visto que cada um pode ter uma organização própria) |

## Definição da informação

Numa primeira abordagem "informação"pode ser definida como sendo dados contextualizados.

Hoje conta-se com um grande número de informações, segundo (CRESTANI,1991 apud EDUARDO LIQUIO TAKAO,2001) elas podem ser textuais, visuais ou auditivas. Como consequência os bancos que armazenam tais informações estão se tornando cada vez maiores.

Informação é uma abstração informal (isto é, não pode ser formalizada através de uma teoria lógica ou matemática), que representa algo significativo para alguém através de textos, imagens, sons ou animação. Note que isto não é uma definição - isto é uma caracterização, porque «algo», «significativo» e «alguém» não estão bem definidos; assumimos aqui um entendimento intuitivo desses termos. Não é possível processar informação diretamente em um computador. Para isso é necessário reduzi-la a dados. A representação da informação pode eventualmente ser feita por meio de dados. Nesse caso, pode ser armazenada em um computador.

Mas, atenção, o que é armazenado na máquina não é a informação, mas a sua representação em forma de dados. Essa representação pode ser transformada pela máquina - como na formatação de um texto - mas não o seu significado, já que este depende de quem está entrando em contato com a informação. Por outro lado, dados, desde que inteligíveis, são sempre incorporados por alguém como informação, porque os seres humanos (adultos) buscam constantemente por significação e entendimento.
Uma distinção fundamental entre dado e informação é que o primeiro é puramente sintático e o segundo contém necessariamente semântica (implícita na palavra "significado" usada em sua caracterização). É interessante notar que é impossível introduzir semântica em um computador, porque a máquina mesma é puramente sintática (assim como a totalidade da matemática). Se examinássemos, por exemplo, o campo da assim chamada "semântica formal" das "linguagens" de programação, notaríamos que, de fato, trata-se apenas de sintaxe expressa através de uma teoria axiomática ou de associações matemáticas de seus elementos com operações realizadas por um computador (eventualmente abstrato).

## Definição de Data Warehouse:

O Data Warehouse (BARQUIM, 1997; CHAUDHURI, 1997; COREY, 2001) é um banco de dados que possui uma quantidade de dados muito grande que contribui para o sistema de suporte a decisão da empresa. Esse grande banco de dados se baseia nos banco de dados dos vários sistemas da empresa. Ele é responsável por armazenar as informações de maneira a interpretar os dados conforme um determinado padrão .

Imon (1991) define um Data Warehouse (DW) como sendo: "Uma coleção de dados orientados a assunto, detalhados, integrados, não-voláteis e que variam temporalmente tendo como objetivo dar suporte ao processo de tomada de decisões gerenciais da empresa."

Outro grande pensador á volta de Data warehouse Ralph Kimball forneceu uma definição muito mais simples de um armazém de dados (DW). Como declarado em seu livro, "The Data Warehouse Toolkit", um armazém de dados é "uma cópia de dados de transação (dados transacionais dos sistemas de origem), especificamente estruturado para consulta e análise". Esta definição fornece menos perspicácia e profundidade que a do Sr. Inmon, mas não é menos precisa.

Um DW então pode ser entendido como uma base de dados centralizada, separada do ambiente transacional dos sistemas de onde são extraídos os dados para análise. O DW possui uma estrutura otimizada para fornecer um ambiente analítico de alta performance para as aplicações de Business Intelligence(BI).

A armazenagem de dados é, essencialmente o que se precisa fazer a fim de integrar dados de fontes heterogéneos num único armazém de dados (DW) para posteriormente ser analizados utilizandos ferramentas OLAP. O problema fundamental reside na necessidade do utilizador final acessar informações que estão distribuídas em vários sistemas da organização.

Com base nessas definições pode se concluir que um Data warehouse também é uma base de dados não transaccional, ou fora do ambiente da produção.

## Característica de Data Warehouse

Baseado nas definições dos diferentes autores, um Data warehouse pode ser caracterizado com base na figura a seguir:


Figura 2:Carateristica DW

**Dados Integrados:**

Os dados que são reunidos no armazém de dados (DW) a partir de uma variedade de origens heterogéneas e fundidos em um todo de forma coerente.

O DW é alimentado por diversos sistemas de origem (sistemas legados), que na maioria das vezes não foram projetados para serem integrados, logo, os dados entram de maneira inconsistente no DW e através de diversos processos de transformação, formatação, sumarização, dentre outros, esses dados devem apresentar uma única "aparência" física a nível corporativo, ou seja, eles devem apresentar um formato comum consistente.

**Dados Não-voláteis:**

Os dados em sistemas transaccionais estão constantemente a serem actualizados, uma vez, que sofrem frequentemente operações como insert,update e Delete. Normalmente registo a registo com uma freqüência regular. Num DW os dados são carregados normalmente em massa, acessados, porém, não são atualizados da mesma maneira que no ambiente transacional. Os dados são estáveis em um armazém de dados (DW). Mais dados são adicionados, mas nunca removidos. Isto confere ao gerenciamento, uma visão consistente dos negócios.

**Orientado a Assunto:**

Em contraste com os sistemas transacionais que são organizados ao redor das atividades do negócio, o DW é composto de dados que representam áreas de atuação ou de interesse do negócio. Por exemplo produtos, pedidos, vendedores etc. Os dados traduzem informações sobre um assunto particular em vez de sobre operações contínuas da empresa.

**Tempo-variante:**

Outra característica marcante é a presença de elementos que representam tempo e variação no tempo dentro dos dados do DW. Essa variação indica que cada conjunto de dados no DW são contextualizados num dado intervalo de tempo, de modo que os dados armazenados ganham marcações de tempo para poder analizar a sua evolução ao longo do tempo. Todos os dados no armazém de dados são identificados com um período de tempo particular.

## Outros características  importantes para Metodologia.

**Granularidade**

Além dos tipos de cargas (incremental ou total), devemos também decidir sobre qual será a granularidade das tabelas Fatos. A granularidade é uma das mais importantes definições na modelagem de dados do Data Warehouse (DW) e requer atenção.  O grão é o menor nível da informação e é definido de acordo com as necessidades identificadas no início do projeto. Ele é determinado para cada tabela Fato, já que normalmente os Fatos possuem informações e granularidades distintas. É importante entender o relacionamento existente entre o detalhamento e a granularidade. Quando falamos de menor granularidade, ou granularidade fina, significa maior detalhamento (menor sumarização) dos dados. Maior granularidade, ou granularidade grossa, significa menor detalhamento (maior sumarização). Assim podemos notar que a granularidade e o detalhamento são inversamente proporcionais. A granularidade afeta diretamente no volume de dados armazenados, na velocidade das consultas e no nível de detalhamento das informações do DW. Quanto maior for o detalhamento, maior será a flexibilidade para se obter respostas. Porém, maior será o volume e menor a velocidade das consultas. Já quanto menor for o detalhamento, menor será o volume, maior a sumarização dos dados e melhor será a performance. Entretanto, menor será a abrangência, ou seja, maior será as restrições das consultas às informações. A sumarização e o detalhamento do grão também podem ser compreendidos pelas operações de Drill Down e Roll Up (Drill Up).

Com o Drill Down estamos diminuindo o nível da granularidade, aumentando assim o nível de detalhes. Ao contrário disso, o Roll Up aumenta o nível da granularidade, diminuindo dessa forma, o nível de detalhamento das informações. Deve ser avaliado o equilíbrio entre detalhamento e sumarização para que a granularidade seja modelada com a melhor eficiência e eficácia para as consultas dos utilizadores, sempre levando em consideração as necessidades levantadas no começo do projeto. Nada vai adiantar deixar a granularidade alta sem que seja alcançado o grão exigido pelo negócio. Também é necessário avaliar o tipo de métrica empregue nas Fatos. No aspecto de obtenção de respostas, as Fatos com métricas aditivas terão uma melhor flexibilidade para se ter menor granularidade. As métricas semi-aditivas, como saldo, ou métricas não-aditivas, como percentuais, serão indicadas para se definir uma alta granularidade.  Portanto, devemos analisar os diversos fatores e aspectos para uma melhor definição dos grãos das tabelas Fatos. As questões de volume de dados, performance e requisitos devem ser ponderados para se chegar a uma correta decisão. Por fim, a granularidade se trata de um assunto de grande importância e enorme impacto, que se mal dimensionado, pode acarretar até mesmo na inviabilização do projeto.


Figura 3:Granularidade

A granularidade não deve ser confundida com o grau de detalhamento. Aliás a figura 4 mostra a relação entre detalhamento e a granularidade.

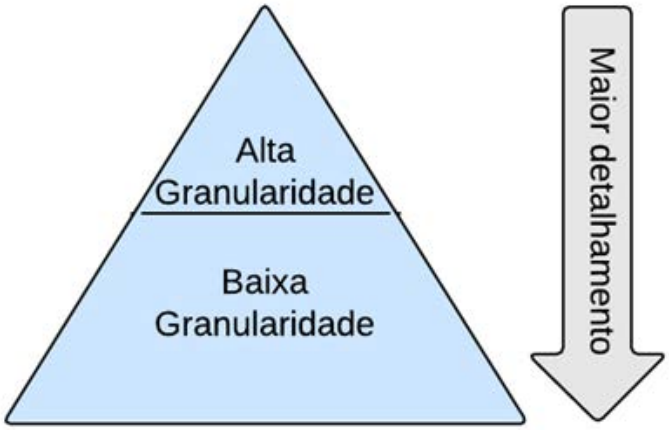



Figura 4:Granularidade e detalhamento de dados

Um outro ponto importante do DW é o nível de granularidade dos dados. Esta outra característica trata do nível de detalhamento dos dados contidos no DW. Quanto mais detalhe a consulta exigir mais baixo será o nível de granularidade, de modo inverso, quanto menos detalhe a consulta exigir mais alto será o nível de granularidade. Segundo W. H. Inmon ( apud Skorupa Parolin Erick data?), a razão pela qual a granularidade é a principal questão de projeto consiste no fato de que ela afeta profundamente na operação da extracção sobre o volume de dados que residem no DW e, ao mesmo tempo, dificulta responder às exigências do utilizador final, mediante o tipo de consulta que pode ser atendida. O volume de dados contidos no DW é balanceado de acordo com o nível de detalhamento de uma consulta.

**Níveis de Granularidade**

O DW tem dois níveis de granularidade e esta forma de trabalho geralmente se encaixa nos requisitos da maioria das empresas. A primeira camada, dos dados levemente resumidos, contém dados extraídos do armazenamento operacional. Os dados são resumidos de forma que a sua utilização fique mais fácil para os analistas e gerentes.

Na segunda camada, chamada também de nível de dados históricos, todos os detalhes vindos do ambiente operacional são armazenados. Com a criação dos dois níveis de granularidade, o analista do SAD conseguiu solucionar dois problemas no DW de uma vez, pois a maior parte do processamento do SAD utiliza-se dos dados levemente resumidos porque eles estão mais compactos e mais fácies de acessar. Porém se houver necessidade de haver um maior nível de detalhamento existe o nível de dados históricos, mas o acesso a esses dados é mais complexo. Particionamento dos Dados O principal objetivo do particionamento trata da repartição dos dados em unidades físicas menores. Esta repartição tem por objetivo facilitar o trabalho do projetista, pois permite uma maior flexibilidade no gerenciamento dos dados. Existem vários critérios para divisão dos dados, como por exemplo:

**Metadados**

Toda e qualquer informação no ambiente DW que não são os dados propriamente ditos, são chamados metadados. Estes são como uma enciclopédia para o DW. Eles estão presentes em uma variedade de formas e formatos para suportar as necessidades desiguais dos grupos de usuários técnicos, administrativos e de negócio do DW (KIMBALL apud Skorupa Parolin  Erick )

Metadados nada mais são além de dados sobre dados, e fazem parte do meio de processamento de informações há tanto tempo quanto os programas e os dados. Portanto, no mundo dos data warehouses é que os metadados assumem um novo nível de importância ao passo que é por meio deles que a utilização mais produtiva do data warehouse pode ser alcançada (INMON, 1997 apud Skorupa Parolin  Erick). Os usuários de DW precisam conhecer a estrutura e o significado dos dados do DW para poder examinar os dados, o que não ocorre em sistemas, onde os usuários interagem com as telas do sistema sem precisar conhecer como os dados são mantidos pelo banco de dados. Outra razão para a importância dos metadados é concernente ao gerenciamento do mapeamento entre o ambiente operacional e o ambiente de data warehouse. À medida que os dados passam do ambiente operacional para o ambiente de data warehouse eles são submetidos a significativas transformações  através de filtros, conversões, resumos e alterações estruturais. Essas transformações precisam manter um rigoroso acompanhamento, e os metadados do DW constituem um local ideal para isso (INMON, 1997 apud Skorupa Parolin  Erick). Mais uma tarefa dos metadados no ambiente de data warehouse é a de manter o acompanhamento das alterações das estruturas de dados ao longo do tempo. Segundo Inmon (1997 apud Skorupa Parolin  Erick), os metadados englobam o DW e mantém informações sobre "o que está aonde" no DW.

Tipicamente os aspectos sobre os quais os metadados mantêm informações são:

- √ A estrutura dos dados segundo a visão do programador;
- √ A estrutura dos dados segundo a visão dos analistas de SAD;

- √ A fonte de dados que alimenta o DW;
- √ A transformação sofrida pelos dados no momento de sua migração para o DW;
- √ O modelo de dados;
- √ O relacionamento entre o modelo de dados e o DW;
- √ O histórico das extrações de dados.

## Base dados Transacionais vs Data warehouse

Os bancos de dados Transacionais armazenam as informações necessárias para as operações do dia-a-dia da empresa. São utilizados por todos os funcionários para registar e executar operações pré-definidas e seus dados podem sofrer constantes mudanças conforme as necessidades atuais da empresa. Por não ocorrer redundância num banco de dados e as informações históricas não ficarem armazenadas por muito tempo, este tipo de estrutura não exige grande capacidade de armazenamento.

Já um DW armazena dados analíticos, tanto detalhados como resumidos, e destinados às necessidades da gerência no processo de tomada de decisões. Isto pode envolver consultas complexas que necessitam acessar um grande número de registos, por isso é importante a existência de muitos índices criados para acessar as informações da maneira mais rápida possível. Um DW armazena informações históricas de muitos anos e por isso deve ter uma grande capacidade de processamento e de armazenamento.

Por isso é de extrema importância diferenciar Data warehouse e Base de dados transaccionais.

| Características | Bancos de dados Operacionais | Data Warehouse |
|---|---|---|
| Objetivo | Operações diárias do negócio | Analisar o negócio |
| Uso | Operacional | Informativo |
| Utilizadores | Informáticos/funcionários Espec. | Gestores/Administração |
| Dominio | Tarefas Rotineiras e operacionais | Decisões Estratégicas |
| Tipo de processamento | OLTP | OLAP |
| Unidade de trabalho | Inclusão, alteração, exclusão | Carga e consulta |
| Número de utilizadors | Milhares | Centenas |
| Tipo de utilizador | Operadores | Comunidade gerencial |
| Interação do utilizador | Somente pré-definida | Pré-definida e ad-hoc |
| Condições dos dados | Dados operacionais | Dados Analíticos |

| | | |
|---|---|---|
| Volume MB – gigabytes | Gigabytes – | terabytes |
| Histórico | 60 a 90 dias | 5 a 10 anos |
| Granularidade | Detalhados | Detalhados e resumidos |
| Redundância | Não ocorre | Ocorre |
| Estrutura | Estática | Variável |
| Estrutura de dados | Aplicacional | Orientado a temas |
| Manutenção desejada | Mínima | Constante |
| Acesso a registos | 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 registo | Aperfeiçoar |
| consultas | | |
| Padrão de uso | Previsível | Difícil de prever |
| Orientado | a aplicações | Orientado a assunto |
| Organização dos dados | Detalhamento Alto | Sumarizado |
| Valores | Valores atuais e voláteis | Valores históricos e imutáveis |

## Arquitectura Data warehouse

### Data Warehouse Bus Architecture

A arquitectura em BUS ou em série,é uma ferramenta concetual que permite decompor em varias fases a tarefa de desenvolvimento de DW.

Definindo um barramento padrão para o ambiente de DW, data marts separados podem ser implementados por grupos diferentes em tempos diferentes. Todos os processos da cadeia de valores da organização criarão uma família de modelos dimensionais que compartilham um conjunto completo de dimensões comuns e conformadas.

Essa arquitectura facilita o desenvolvimento faseado de Data mart e a sua posterior agragação num DW consistente e coerente

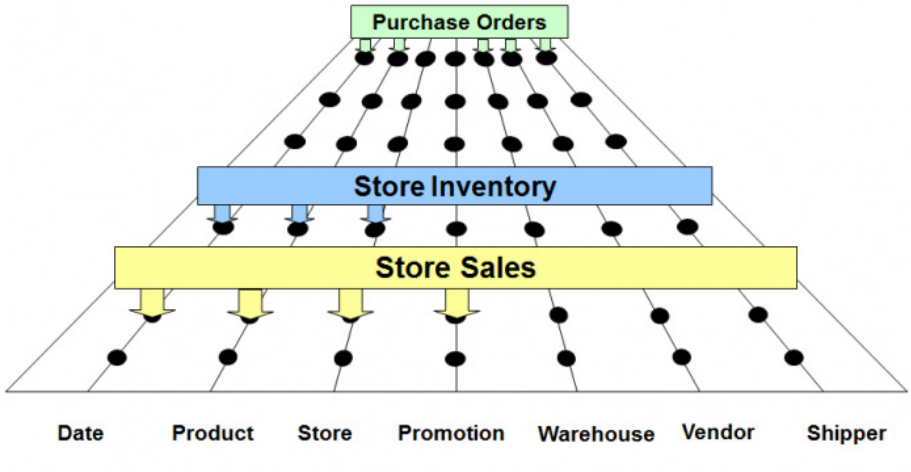



Figura 5:Bus Arquitetura

## Data Warehouse Bus Matrix

É uma ferramenta utilizada para criar, documentar e comunicar a arquitectura de Data warehouse.

A matriz é um quadro em que as linhas representam o processo directamente associados ao fluxo da informação da empresa. Para aplicações matriciais não deve ser considerado processos com suportes departamentais ou sectoriais.

As linhas da matriz em bus vão traduzir directamente em Data Marts com suporte as actividades fundamentais da empresa  e as colunas a dimensões conformadas. A matriz é a ferramenta usada para criar, documentar, gerenciar e comunicar a arquitetura de barramento. Segundo Kimball, é o artefato de análise mais importante do desenvolvimento de um DW. É uma ferramenta híbrida, que serve para design técnico, para gerência de projeto e como forma de comunicação organizacional.

A matriz em bus é como o próprio indica,um quadro com função primordial de servir de suporte ao desenvolvimento faseado, mas consistente do Data Warehouse

| | COMMON DIMENSIONS | | | | | | |
|---|---|---|---|---|---|---|---|
| BUSINESS PROCESSES | Date | Product | Warehouse | Store | Promotion | Customer | Employee |
| Issue Purchase Orders | X | X | X | | | | |
| Receive Warehouse Deliveries | X | X | X | | | | X |
| Warehouse Inventory | X | X | X | | | | |
| Receive Store Deliveries | X | X | X | X | | | X |
| Store Inventory | X | X | | X | | | |
| Retail Sales | X | X | | X | X | X | X |
| Retail Sales Forecast | X | X | | X | | | |
| Retail Promotion Tracking | X | X | | X | X | | |
| Customer Returns | X | X | | X | X | X | X |
| Returns to Vendor | X | X | | X | | | X |
| Frequent Shopper Sign-Ups | X | | | X | | X | X |

Figura 6:Data Warehouse Bus Matrix

## Componentes Do Data Warehouse

Data warehouse é um sistema composto por um conjunto de subsistema que trabalham em conjunto para atingir um objectivo comum. Objectivo esse que é de integrar os dados num único armanzém como forma de aceder com facilidade. Para que possa entender o principio de funcionamento de um Data Warehouse, precisa conhecer os seus componentes para conhecer a sua arquitectura

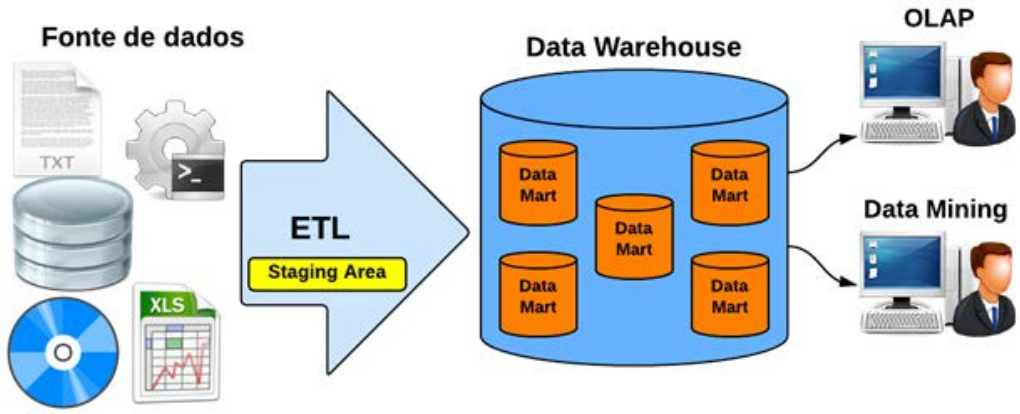



Figura 7: Componente Data warehouse

- **Fonte de dados ou aplicações transacionais**: abrange todos os dados de origem que irão compor as informações do DW. Compreende os sistemas OLTP (Online Transation Processing), ficheiros em diversos formatos (XLS, TXT, etc), sistemas de CRM, ERP, entre vários outros. Pode ser considerado a fonte de dados que abastece o Data Warehouse,isto é,um Data Warehouse é povoado por fontes de dados diversificados.
- **ETL:** o ETL, do inglês Extract, Transform and Load, é o principal processo de condução dos dados até o armazenamento definitivo no DW. É responsável por todas as tarefas de extração, tratamento, limpeza dos dados, e carregamento de dados transacionais no Data ware house.
- **Staging Area/ área de publicação dos dados:** a Staging Área é uma área de armazenamento temporário de dados e um conjunto de processos conhecidos como ETL (Limpeza,Transformação e carga de dados).Faz a ponte entre aplicações transaccionais e o data warehouse . Auxilia a transição dos dados das origens para o destino final no DW.
- **Data Warehouse:** essa é a estrutura propriamente dita de armazenamento de dados. Apenas os dados com valor para a gestão corporativa serão povoados no DW.
- **Data Mart**: o Data Mart é uma estrutura similar ao do DW, porém com uma proporção menor de informações. Trata-se de um subconjunto de informações do DW que podem ser identificados por assuntos ou departamentos específicos. O conjunto de Data Marts em conformidade dentro da organização compõe o DW.
- **OLAP:** o OLAP, do inglês On-line Analytical Processing, na arquitetura de um DW se refere as ferramentas com capacidade de análise em múltiplas perspectivas das informações armazenadas.

- **Data Mining:** Data Mining ou Mineração de Dados, se refere as ferramentas com capacidade de descoberta de conhecimento relevante dentro do DW. Encontram correlações e padrões dentro dos dados armazenados.

O fluxo das atividades nessa arquitetura se inicia com a extração dos dados das origens. Esses dados são então armazenados temporariamente na Staging Area, onde são tratados com as regras e padrões predeterminados para então prosseguir para a etapa de carga (Load), em que os dados são carregados no DW. Por fim, essas informações são normalmente consultadas através de ferramentas de análises (OLAP) ou ferramentas de mineração (Data Mining) para encontrar, assim as respostas e insights necessários para a tomada de decisão.

Cada um desses componentes pode ser considerado um subsistema que deve obdecer uma ordem de processamento, que tem de ser respeitada para que se possa atingir o resultado final com sucesso. A falha de um desses componentes ou o não seguimento da ordem pode comprometer o resultado final

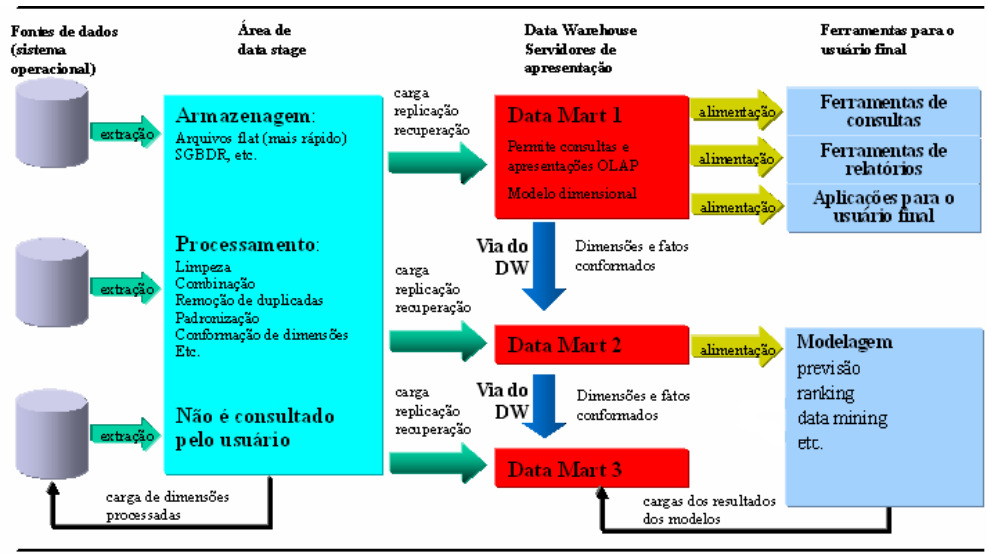



Figura 8: Funções dos componentes DW

## Sistema de Apoio à Decisão

O processo de tomada de decisão tem sido transformado a partir de sua inserção em sistemas de informações capazes de gerarem possibilidades e reproduzirem cenários de acordo com premissas e dados estabelecidos. Esses sistemas não produzem apenas informações gerenciais, mas dão suporte à tomada de decisão dos gestores.

Sistema de apoio a decisão não pode ser montado sem um armazém de dados uma vez que os modelo para tomada de decisão deve analizado com base na evolução dos dados de um determinado assunto.

Um sistema de informação é parte integrante das organizações, pois transformando informação em conhecimento auxilia no cumprimento desde tarefas rotineiras e simples até às não-rotineiras e complexas.

Estas últimas, geralmente, são de competência dos gestores que ditam os rumos das organizações através de suas decisões, ficando claro que o processo decisório continua sendo um dos papéis mais desafiadores de qualquer gerente ou administrador.

√ Visando satisfazer essa prerrogativa, dentre outras, a tecnologia em software evoluiu até o surgimento dos Sistemas de Apoio à Decisão. Alguns autores, como Turban (2004 apud Akyria Bolonine et al) denominam esse sistema de Sistema de Apoio à Decisão (SAD) e outros como Laudon (2001) de Sistema de Suporte à Decisão (SSD). Importante é saber que esses softwares trabalham com sistemas interativos que, seguindo premissas, oferecem informações e modelos para a solução de questões de cunhos tático e estratégico.

O Sistema de Apoio à Decisão (SAD) é um sistema baseado em computadores que através de informações e modelos especializados ajudam a resolver problemas organizacionais, sua função é apoiar o processo de tomada de decisão em áreas de planejamento estratégico, controle gerencial e controle operacional, sendo isso o que o diferencia dos demais tipos de sistemas de informações.

Sua demanda surgiu diante do crescimento competitivo das organizações, pois o SAD é desenvolvido através de dados históricos e experiências individuais que são incorporados como informações úteis possibilitando melhores condições para a tomada de decisão e aumentando as vantagens obtidas pela empresa.

Muitas empresas estão utilizando o SAD para melhorar o processo decisório. As razões citadas pelos gerentes são:

- Necessidades de informações novas e mais precisas;
- Necessidade de Ter informações mais rapidamente;
- O monitoramento das inúmeras operações de negócios da empresa estava cada vez mais difícil;
- A empresa estava operando em uma economia instável;
- A empresa enfrentava maior concorrência nos mercados interno e externo;
- Os sistemas instalados na empresa não apoiavam adequadamente os objetivos de maior eficiência, rentabilidade e ingresso em mercados lucrativos;
- O departamento de sistemas de informação não conseguia mais atender à diversidade de necessidades imediatas da empresa e de seus executivos e não havia funções de análise de negócio embutidas nos sistemas existentes." (TURBAN, 2004, apud Akyria Bolonine et al).

É importante que os conceitos do SAD retratem a cultura organizacional, não servindo apenas para atender às necessidades específicas do utilizador, mas que seja orientado para pessoas que tomam decisões, devendo ser flexível na busca, acesso e manipulação das informações, utilizando-se de uma interface o mais amigável possível para satisfazer as necessidades gerais das organizações.

## Características

Os SADs possuem várias características, sendo algumas delas:

- Trabalhar com diversas fontes de dados;
- Variedade nos Relatórios;
- Análise de Sensibilidade, Simulação e Análise de Tomada de Decisão.

Utilizando um SAD é possível aos tomadores de decisão buscar informações em bancos de dados diferentes, mesmo que estejam em lugares distintos. É possível também acessar a outras fontes de dados pela Internet ou por uma Intranet da organização. O processo de tomada de decisão necessita que se tenha informações específicas sobre o determinado problema, para que, desta maneira, o gerente possa analisá-lo e suprir suas necessidades.

"Enquanto os outros sistemas de informação disponibilizam basicamente relatórios de formato fixo, os SSDs possuem uma variedade maior de formatos." (REYNOLDS, 2002, p. 316 apud Akyria Bolonine et al). A flexibilidade que o SAD oferece ao disponibilizar os relatórios facilita o gestor, de modo que ele tenha somente as informações que necessita, visto que a variedade de problemas e necessidades dos tomadores de decisão é muito ampla.

"A análise de sensibilidade constitui o processo de introduzir mudanças hipotéticas nos dados do problema e observar o impacto nos resultados." (REYNOLDS, 2002, p.317 apu Akyria Bolonine et al). Dessa forma, é permitido que o gerente planeje a decisão que tomará, pois é possível modificar hipoteticamente os dados franqueando uma visão do que acontecerá se aquela decisão for tomada.

A simulação é outra característica importante num SAD, pois demonstra a probabilidade de algo acontecer através de cenários construídos a partir de decisões tomadas, possibilitando ao gestor uma maior segurança para solucionar o problema.

Cabe, ainda, mencionar a análise de tomada de decisão que é um processo conduzido pelo SAD. Pois, permite ao gerente fornecer os dados de um problema e obter o resultado fornecido pelo SAD como sua solução, desse modo, conseguindo visualizar o alcance de uma determinada meta.

Vale lembrar que algumas decisões são tomadas em grupos abrangendo diversas visões sobre um mesmo tema. Para atender a essa situação foram desenvolvidos os Sistemas de Apoio à Decisão em Grupo (SADG) que convergem diferentes pontos de vista em uma solução comum. Uma grande vantagem desse sistema é a participação de vários gerentes de diversas filiais em cidades diferentes no processo decisório, utilizando-se de ferramentas como: Rede Local de Decisões, Sala de Decisões, Rede Remota de Decisões e Teleconferência.

## Estrutura

"Os componentes de SAD incluem um banco de dados usado para consulta e análise, um sistema de software com modelos, data mining e outras ferramentas analíticas e uma interface com o utilizador." (OLIVEIRA, 2003, p.198 317  apu Akyria Bolonine et al). Os principais componentes são:

- O banco de dados SAD que é uma coleção de dados atuais e históricos de uma variedade de sistemas ou grupos. Pode ser um pequeno banco de dados em um computador isolado, coletando dados externos e corporativos, combinando-os para auxiliar na tomada de decisão. Ou ele pode ser um poderoso data warehouse continuamente atualizado por dados organizacionais.
- O sistema de software que pode conter várias ferramentas OLAP, ferramentas de data mining ou uma coleção de modelos matemáticos ou analíticos que podem ser facilmente acessados pelo utilizador do SAD.
- A interface do SAD que permite ao utilizador interagir com o sistema de software. Geralmente, seus utilizadors são executivos ou gerentes de corporações e não possuem muita perícia no uso da tecnologia, levando essa interface a ser amigável ao extremo para que o se possa tirar o máximo proveito da ferramenta.

Um modelo de SAD pode ser físico, matemático ou verbal, visto que cada SAD é construído para um propósito, ele poderá fazer diferentes coleções de modelos disponíveis na organização dentro da realidade do propósito desejado. Os modelos mais conhecidos e utilizados são:

- Modelos estatísticos;
- Modelos de otimização;
- Modelos de previsão;
- Modelos de biblioteca e
- Modelos de análise de sensibilidade (OLIVEIRA, 2003).

Teoricamente, um SAD pode ser aplicado em qualquer área do conhecimento. Alguns exemplos seriam: diagnósticos médicos, preparo do solo para plantio, uso na meteorologia, na produção de aviões e para controle de irrigação de um solo, analisando o tipo de cultura e solo para determinar o tipo de irrigação a ser implantado.

Ao longo dos tempos vários autores mostraram que os SAD (Sistemas de Apoio a Decisão), são sistemas interativos, baseados em computadores, que têm como objetivo principal ajudar nas decisões e utilizar os dados e modelos para identificar e resolver problemas. Os objetivos dessas aplicações é ceder informações para facilitar uma melhor tomada de decisão. Os sistemas de apoio à decisão não substituem o decisor, no geral esses sistemas usam são desenvolvidos utilizando um processo evolutivo e iterativo, com uma interface que facilita a aprendizagem. São capazes de apoiar todos os níveis de gestão, desde o nível estratégico até o operacional.

Claramente esses sistemas são uma poderosa ferramenta e esta se tornando essencial para apoiar os gestores, aumentando a capacidade de processamento de grandes volumes de informação ao longo do processo de tomada de decisão. Um SAD traz como beneficio uma vantagem competitiva ou estratégica sobre os concorrentes, pois encoraja o decisor na exploração e descobertas do mesmo. Um SAD tem a função de gerar informação, utilizando ferramentas sofisticadas de análise, banco de dados internos e externos, para propiciar ao decisor soluções para as questões essenciais ao funcionamento da empresa, auxiliando assim a tomada de decisão. Um SAD eficiente permite fácil interação com o utilizador do sistema, para que este possa acessar tranquilamente seu banco de dados e modelos e absorver de forma natural as informações e sugestões armazenadas, obtendo vantagem competitiva no mercado em que atua. A eficácia de um SAD vai depender dos sistema que lhe dão suporte. Entre estes sistemas temos Data Warehouse, que funciona como um "armazém" de dados que dá suporte ao processo de decisão, são orientado a assuntos onde os dados relacionam-se a temas específicos, variável com o tempo, que dizem respeito a períodos de tempo específicos como dia de pagamento, mês, férias, semestre. O acesso dessas informações podem ser feitas de 3 formas básicos:

- √ Virtual - o utilizador final tem acesso direto à informação;
- √ Central - acesso à informação se dá ao em um único ponto ;
- √ Distribuída - informação distribuída em alguns pontos da estrutura da empresa, o que exige alimentação e manutenção distribuída dos dados. Alguns equívocos na hora são comuns na hora de implementar um Data Warehouse, o principal deles é o de escolher um gerente orientado para tecnologia, muitas empresas comentem esse tipo de erro. Focar o sistema em dados tradicionais internos orientados a registo e ignorar o valor potencial dos dados textuais e dados externos, gerar expectativas que não serão atendidas. A utilização de um SAD proporciona um auxílio significativo ao processo de tomada de decisão, onde as informações fornecidas por esse sistema são incorporadas às nossas experiências individuais. É importante que um SAD retrate a cultura de uma organização, tornando-se parte dela, de forma que não atenda às necessidades apenas de uma única pessoa, razão pela qual empresas estão adotando cada vez mais essa tecnologia.

## Avaliação da Unidade

XIV. Actividades de aprendizagem

**Verifique a sua compreensão!**

"Data Warehouse é um sistema que extrai, limpa, trata e entrega dados de várias fontes em um modelo dimensional e suporta/implementa consultas e análises para o processo de tomada de decisão."
Acerca de Data Warehouses, assinale a afirmação incorreta.

1. Uma tabela de dimensão registra medidas de negócios que serão analisadas enquanto a tabela de fatos contém as descrições textuais do negócio. Os fatos são os aspectos pelos quais se pretende observar as métricas relativas ao processo que está sendo modelado.
2. A tabela fato é composta por duas ou mais chaves estrangeiras que se conectam com as chaves primárias da tabela dimensão. Sua chave é usualmente chamada de chave composta ou concatenada. Em um modelo dimensional, toda tabela que expressa uma relação de muitos-para-muitos deve ser uma tabela de fatos. Todas as outras tabelas são tabelas de dimensão.
3. Para suprir a necessidade de registrar um fato quando não há medidas, são utilizadas tabelas de fatos sem fatos, capturando apenas o relacionamento entre as chaves envolvidas.
4. Também conhecidas como chaves sem significado, chaves inteiras, chaves não naturais, as chaves substitutas são, basicamente, um valor inteiro sequencial atribuído a cada registro inserido, conforme a necessidade. Uma das razões para se utilizar estas chaves é que o data warehouse deve se manter isolado das regras operacionais para gerar, atualizar, excluir, reciclar e reutilizar os códigos utilizados nos sistemas transacionais e não pode ficar vulnerável a problemas de sobreposição de chaves, no caso de aquisição ou consolidação de dados. Outra razão é o melhor desempenho no acesso às informações. Para manter o relacionamento com a tabela de fatos, muito espaço em disco é desperdiçado por causa do tamanho destas chaves.
5. Os atributos mais comuns em uma tabela de fatos são valores numéricos, que são, em sua maioria, aditivos. As métricas aditivas são as que permitem operações como adição, subtração e média de valores por todas as dimensões, em quaisquer combinações de registros. Métricas aditivas são importantes porque normalmente as aplicações de data warehouse não retornam uma linha da tabela de fatos, mas sim centenas, milhares e até milhões. Existem também métricas não-aditivas e métricas semi-aditivas. As métricas não aditivas são valores que não podem ser manipulados livremente, como valores percentuais ou relativos, e as métricas semi-aditivas são valores que não podem ser somados em todas as dimensões.

2-Objectivos de Data warehouse

a. Integrar sistemas de múltiplas fontes

b. Facilitar o processo de análise sem impacto para o ambiente de dados operacionais

c. Obter informação em quantidade e qualidade

d. Ser flexível e ágil para atender novas análises

3-Definição Data warehouse

Data Warehouse "Uma colecção de dados...

a) Orientados ao assunto

b) Integrados

c) Voláteis

d) Variantes no tempo e espaço

... Para fornecer suporte ao processo de tomada de Decisões na organização" [Inmon, 92]

Indique, para cada uma das alíneas tipo de processamento (OLPT ou OLAP) correspondente:

| | Tipo de Processamento (OLPT/ OLAP) |
|---|---|
| a) Transacções pontuais (1 registo por vez)<br>b) Velocidade e automação de funções "repetitivas"<br>c) Actualizações e consultas em grande número<br>d) Trabalha com alto nível de detalhe<br>e) Situação corrente<br>f) "Pequeno" número de consultas "variáveis"<br>g) Centenas, milhares, de registos por consulta<br>h) Diversas fontes de dados<br>i) Diferentes perspectivas<br>j) Operações de agregação e cruzamentos<br>k) Actualização quase inexistente, apenas novas inserções<br>l) Dados históricos são relevantes | |

# Unidade I. Desenvolvimento Do Data Warehouse

## Introdução à Unidade

O propósito desta unidade é verificar a compreensão dos conhecimentos que possui relacionados com este curso.

É de extrema importância ter em mente as etapas para o desenvolvimento de um Data warehouse. Os alunos devem diferenciar as actividades a serem executadas em cada Etapa até a ter os dados armazenados no armazém.

## Objectivos da Unidade

Após a conclusão desta unidade, deverá ser capaz de:

1. Conhecer ciclo de Vida do Data Warehouse
2. Projectar data warehouse
3. Identificar dos Requisitos de Negócio PARA DW;
4. Apontar os aspectos importantes para o Desenho da Arquitetura DE DATA WAREHOUSE;
5. Componentes e suas caracteristicas

Seleção dos Produtos e Ferramentas;

### Termos-chave

**Banco de dados relacional (RDB):** um banco de dados que pode ser considerado um conjunto de tabelas e manipulado de acordo com o modelo relacional de dados. Cada banco de dados inclui um conjunto de tabelas de catálogo do sistema que descrevem a estrutura lógica e física dos dados, um arquivo de configuração que contém os valores dos parâmetros alocados para o banco de dados e um registo de recuperação com transações contínuas e transações que podem ser colocadas em arquivo.

**Cardinalidade:** O número de linhas em um banco de dados ou o número de elementos em uma matriz.

**Chave:** Uma coluna ou uma coleção ordenada de colunas que é identificada na descrição de uma tabela, índice ou limitação referencial. A mesma coluna pode fazer parte de mais de uma chave.

**Desnormalização:** A duplicação intencional de colunas em várias tabelas para aumentar a redundância dos dados. Às vezes, a desnormalização é utilizada para aprimorar o desempenho.

**Banco de dados (BD)**: Uma coleta de itens de dados relacionados ou independentes que são armazenados juntos para atender a um ou mais aplicativos.

## Requisitos

## Ciclo de Vida do Data Warehouse

**Projeto de Armazém de dados e Criação:**

O processo de projetar o armazém de dados, deve ser analizado cuidadosamente como forma de responder às perguntas para as quais o armazém será definido. Isto envolve um esforço que exige uma compreensão do esquema de banco de dados a ser criado, e muita interação com O CLIENTE FINAL. O projeto é freqüentemente um processo interativo e deve sofrer várias alterações durante o tempo de definição, antes do modelo SER CONSIDERADO FINAL. Grande cuidado deve ser tomado nesta fase, porque uma vez que o modelo seja povoado com grandes quantidades de dados, sendo alguns desses dados muito difíceis de recriar, o modelo não poderá ser facilmente modificado.

Esse processo passa por uma série de etapas, que deve ser analisado e construído de forma separado.


Figura 1:Ciclo de Vida de um DW

Frente à realidade do mercado no gerenciamento de informações estratégicas, Kimball et al (1998 APUD RAMOS PEREIRA,MOISÉS data?) promoveu, desde meados dos anos 80, um modelo contendo as principais directrizes de desenvolvimento do DW, distribuídas em 7 fases. Esse modelo afirma que os projetos do DW devem incidir sobre as necessidades do negócio e que os dados devem ser unidimensionais quando apresentados aos clientes. Cada projeto no DW deverá ter um ciclo finito com início e fim bem definidos.

A sua primeira fase consiste do Planejamento de Projeto, seguida pelas fases de Definição dos Requisitos de Negócio, Design Técnico de Arquitetura, Seleção e Instalação de Produtos, Modelagem Dimensional, Design Físico, Design e Desenvolvimento da Data Staging Área, Especificação e Implementação da Aplicação Analítica, Implantação, Manutenção e Crescimento (KIMBALL, 2002). Este modelo foca o desenvolvimento do DW para um ambiente de bancos de dados relacionais.

**Etapas para construção do DW:**

**1-Levantamento das necessidades OU DEFINIÇÃO DOS REQUISITOS:** devemos antes de tudo fazer o levantamento de todas as informações desejadas pelo utilizador. Nesse primeiro momento fazemos o cruzamento de Dimensões e Fatos necessários para alcançar os anseios dos gestores. Nesse primeiro momento trabalha em O QUÊ o DW terá, e não O COMO, por isso não devemos nos preocupar com a existência efetiva dos dados mas sim com os desejos requisitados.

Como um data warehouse não se de pode comprar, mas sim tem de ser desenvolvido, isto quer dizer, que deve ser desenhado de forma a responder algumas inquietações dos decisores. Desta forma que que a perspectiva de analise de requisitos  é bem diferente da metodologia utilizada no desenho do sistema operacional.

Na construção de um DW a preocupação é optimizar o modo como os utilizadores percebem os dados e não evitar a redundância.

Dai é de extrema importância o levantamento de requisitos para que se possa identificar as métricas e as dimensões que descrever os registos a serem medidas.

**2.Mapeamento dos dados:** nessa etapa fazemos o mapeamento dos dados, identificando a fonte e como chegar até eles. Aqui vemos a viabilidade dos desejos elencados na primeira etapa, verificando a existência ou não dos dados para o alcance das necessidades solicitadas.

**3.Construção da Staging Area:** após o mapeamento, construímos a estrutura chamada Staging Area, que se trata da área de transição dos dados dosistema transaccional para o DW. Nessa área os dados são copiados e desacoplados dos sistemas de operação (OLTP) e recebem o devido tratamento para as futuras cargas nas tabelas de Fatos e Dimensões.

**4.Construção das Dimensões:** construímos nessa etapa a estrutura das Dimensões que farão parte do DW. Definimos também a historicidade que os dados irão possuir nas Dimensões.

**5.Construção ds(s) Fato(s):** construímos nessa etapa (após a construção das Dimensões) a(s) estrutura(s) da(s) Fato(s). Aqui é avaliado e definido a granularidade da informação que será armazenada em cada Fato. Avaliamos também a expectativa de crescimento e de armazenamento que serão utilizados.

**6.Definição do processo geral de carga:** após concluirmos as etapas anteriores, precisamos criar o motor para que tudo seja carregado, atualizado, orquestrado e processado de forma automática e ordenada. Por isso, a necessidade do processo geral de carga que é o "cérebro" do DW.

**7.Criação dos metadados:** por fim, precisamos desenvolver toda a documentação dos metadados, que incluem o processo de construção e o dicionário de dados. Os metadados fornecem apoio importante para a gestão do conhecimento.

Lembrando que o Data Mart é a divisão do DW em subconjuntos de informações organizadas por assuntos específicos. Logo, todas essas etapas, com exceção do "levantamento das necessidades" (que deve ser realizada, preferencialmente, uma única vez), devem ser repetidas a cada novo Data Mart criado.

Segue o fluxo do processo de construção, que pode ser cíclico até que todos os Data Marts sejam desenvolvidos:

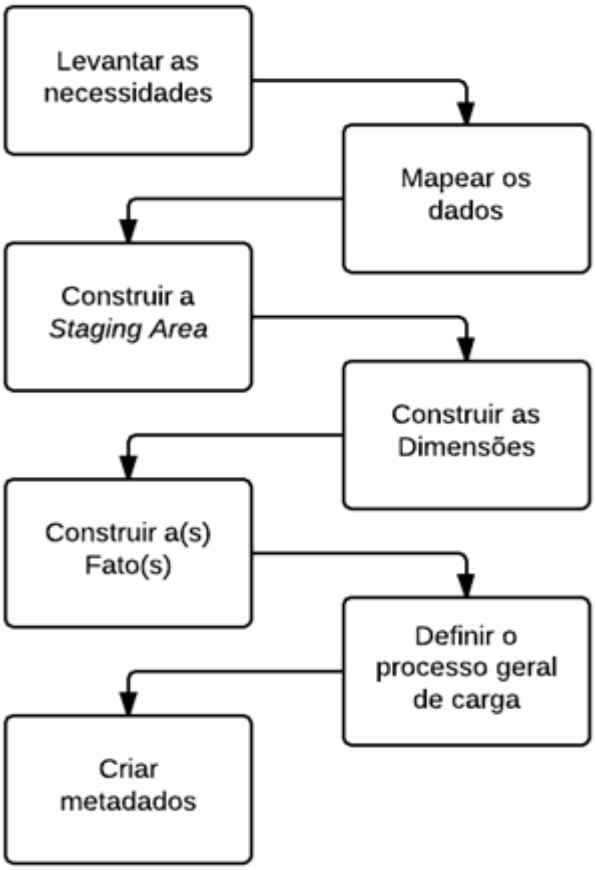



Figura 2:Etapas para construção DW

É importante respeitar a sequência dessas etapas, pois elas possuem dependência de término para início. Ou seja, a etapa sucessora só deve ser iniciada quando a sua predecessora for concluída.

Por fim, com a devida atenção dada a cada uma dessas atividades e a conclusão de cada uma delas com êxito fará com que as chances de sucesso no projeto de construção do DW seja praticamente garantido. Dessa forma teremos, efetivamente, um repositório que armazenará as informações que auxiliarão a organização na tomada de decisão.

## Atividade:

1. Exemplifique as tarefas que são executadas em cada etapas.
2. Será que levantamentos de requisitos feito para um modelo tradicional serve para Data Warehouse? Justifique a sua resposta.
3. Será que pode ser definido o ciclo de vida de DW igual a ciclo de Vida de um Software,em que a sequência pode ser definido por membros da equipa?Instruções

# Unidade 2: Projecto De Carregar Dados Em Um Data Warehouse

- Princípios de modelagem
- Construção de Cubos Multi-Dimensionais
- Esquema de estrela e floco de neve
- Extração, transformação e carga de dados em data warehousing-etl
- Qualidade de dados armazenados
- Ferramenta para extração, transformação e carga de dados

## Objectivos:

Nesta unidade os alunos devem apreender a diferenciar e desenhar o modelo para o desenho de DW e preparar os dados para serem carregados. Para isso devem compreender o conceito de modelação dimensional , e as suas vantagens. Mas devem diferenciar os dois modelos existente ,de estrele e floco de neve e saber quando utilizar um ou outro modelo.

### Termos-chave

**Cubo:** É uma estrutura de dados multidimensional que organiza os dados de um Data Mart em dimensões, métrica e cálculos, promovendo respostas muito rápidas para análises de dados.

**KPI (Key Performance Indicator) – Indicador Chave de Performance:** É um subconjunto de métricas de negócio gerenciáveis que reflete de maneira consistente e apurada toda a performance do negócio. O conjunto exato de métricas varia de negócio para negócio. O cálculo que relaciona um indicador de performance a metas pré-estabelecidas, retornando o status deste indicador com relação ao que foi planejado. Normalmente são três status: vermelho (para indicador com baixa performance), amarelo (para indicador com performance razoavel) e verde (para indicador com performance dentro do planejado).

**Métrica:** Objeto ao qual se permite atribuir um valor mensurável: venda, custo, etc.

**ODS ou Operational Data Store:** Banco de dados operacional.

**Modelo Estrela:** Um tipo de esquema de banco de dados relacional que é composto de um conjunto de tabelas constituído de uma única tabela central de fatos rodeada por tabelas de dimensão. Consulte também tabela de dimensão, junção star.

**DIMENSÃO** Também chamadas Dimensões de Negócios, são o núcleo de componentes ou categorias de negócio, ou seja, tudo que se quiser analisar nos relatórios ou visões.

**Dimensão:** Entidade relacional (tabela) utilizada em *Data Warehouses*, que contém dados cadastrais: clientes, produtos, geografia, tempo, etc., que são utilizados para criar visões de análise do fato a ser estudado.

**Multidimensional Data:** São dados que podem ser analisados de acordo com muitos critérios, por exemplo, vendas por tipo de produto, por região, por tipo de consumidor etc.

**Tabela Fato:** Entidade relacional (tabela), componente do *Data Warehouse*, que contém os dados transacionais que se deseja estudar. Normalmente fato é algum assunto especifico que ocorreu, como venda, por exemplo. A tabela de fato contém, na maioria dos casos, indicadores numéricos que, agregados, agregam valor ao negócio.

**OLAP (Online Analytical Processing) – Processamento Analítico Online:** É uma abordagem para análise e relatório que permite ao usuário, de maneira fácil e seletiva, extrair e visualizar dados de diferentes pontos de vista baseados em uma estrutura de dados multidimensional chamada Cubo.

**OLTP (Online Transaction Processing) – Processamento de Transações Online:** É uma aplicação que facilita e gerencia o processamento de transações, tipicamente para entrada e recuperação de dados. Por exemplo: sistema de transações bancárias registra todas as operações efetuadas em um banco.

**Surrogate Key:** Número inteiro sequencial. No caso das Datas, pode ser usado ou não. Em datas pode-se usar, por exemplo, o tipo date ou um numérico representando a data como YYYYMMDD, por exemplo.

**ETL :** Sistemas ETL (Extract, transform, and load), como o WebSphere DataStage, fazem o trabalho de extração de dados de vários sistemas operacionais, transformando-os para atender às necessidades dos aplicativos de negócios e carregando-os nos bancos de dados de destino, como armazéns de dados e data marts.

**Extract, load, and transform (ELT):** O processo de extração de dados de uma ou mais origens, carregando-os diretamente em um banco de dados relacional e depois utilizando o mecanismo de banco de dados para executar transformações de dados. Consulte também extract, transform, and load.

**Gerenciador de banco de dados:** Um programa que gerência dados fornecendo controle centralizado, independência de dados e estruturas físicas complexas para acesso, integridade, recuperação, controle de simultaneidade, privacidade e segurança eficientes. Consulte também servidor de dados.

**Tabela:** Em um banco de dados relacional, um objeto de banco de dados que consiste em um número específico de colunas e que é utilizado para armazenar um conjunto não ordenado de linhas. Consulte também tabela de base, tabela temporária, visualização.

**Dimensão:** Uma categoria de dados, como hora, contas, produtos ou mercados. Em um esboço de banco de dados multidimensional, as dimensões representam o nível de consolidação mais alto.

**Hierarquia:** Um relacionamento definido entre um conjunto de atributos que são agrupados por níveis na dimensão de um modelo de cubo. Estes relacionamentos entre níveis geralmente são definidos com uma dependência funcional para aprimorar a otimização. Várias hierarquias podem ser definidas para uma dimensão de um modelo de cubo.

**Nível:** Um conjunto de atributos que trabalham juntos como uma etapa lógica na ordenação de uma hierarquia. Um nível contém um ou mais atributos que estão relacionados e podem funcionar em uma ou mais funções no nível. O relacionamento entre os atributos em um nível normalmente é definido com uma dependência funcional.

**Medida:** Uma entidade de medida utilizada em objetos de fatos. É possível calcular medidas com uma expressão SQL ou mapeá-las diretamente para um valor numérico em uma coluna. Uma função de agregação resume o valor da medida para análise dimensional.

**Tabela de armazenamento em cluster multidimensional (tabela MDC):** Uma tabela cujos dados são organizados fisicamente em blocos ao longo de uma ou mais dimensões, ou chaves de cluster, especificada na cláusula ORGANIZE BY DIMENSIONS.

**Tabela de fatos:** Uma tabela relacional que contém fatos, como unidades vendidas ou custo de mercadorias e chaves estrangeiras que ligam a tabela de fatos a cada tabela de dimensão. Consulte também junção star.

## Modelagem de Dados Multidimensional

Modelagem multidimensional é uma técnica de concepção e visualização de um modelo de dados, de um conjunto de medidas que descrevem aspectos comuns de negócio. Sua utilização ajuda na sumarização e reestruturação dos dados e apresenta visões que suportam a análise dos valores destes dados (MACHADO,2004).

De acordo com Oliveira (2002), um banco de dados multidimensional não armazena os dados como registos em tabelas, ele armazena os dados em arrays multidimensionais, possuindo um número fixo de dimensões.

Três elementos básicos compõem o modelo multidimensional: (MACHADO, 2004)

- Fatos: é uma coleção de itens de dados composta de medidas. É utilizado para analisar o processo de negócio de uma empresa, refletindo assim a evolução dos negócios do dia-a-dia desta empresa. Ele é implementado em tabelas denominadas tabelas de fato (fact tables) e representado por valores numéricos;
- Dimensões: são os elementos que participam de um fato e que determinam o contexto de um assunto de negócios. As dimensões podem ser compostas por membros que podem conter hierarquias. Membros são as possíveis divisões ou classificações de uma dimensão. Por exemplo, a dimensão tempo, pode ser dividida nos seguintes membros: ano, trimestre e mês, e a dimensão localização em: cidade, estado e país;

– Medidas (variáveis): são os atributos numéricos que representam um fato, ou seja, representam o desempenho de um indicador de negócios relativo às dimensões que participam desse fato. Uma medida é determinada pela combinação das dimensões que participam de um fato e estão localizados como atributos de um fato. Por exemplo, o valor em reais das vendas, o número vendido de unidades de produtos e a quantidade em estoque.

## Tipo de FATOS

No desenvolvimento de um Data Warehouse (DW) devemos estar atentos aos tipos de métricas que iremos trabalhar. Métricas são resultantes do processo de negócio ou eventos que pretendemos medir. Muitas vezes os analistas engana-se,porque pensam que métrica é tudo igual, e que seu comportamento é o mesmo de todas as outras e que tem de ser numerico. As ações efetuadas através das ferramentas OLAP (On-line Analytical Processing) para a geração das análises são distintas a depender do tipo de métrica existente. Elas podem ter regras de agregações (sumarização) como somatório, média, mínimo, máximo, etc. Para entendermos melhor, vamos conhecer os tipos de métricas que normalmente podem está contidos dentro de um DW.

São três os tipos: métricas aditivas, semi-aditivas e não-aditivas.

- Geralmente são numéricos e aditivos
- Pode ser não numérica
- Pode ser Sem medidas (Tabelas sem fatos).

As métricas aditivas são aquelas que podem ser sumarizadas independente das dimensões utilizadas. Este tipo de métrica pode ser utilizada sem quase nenhuma restrição ou limitação e são flexíveis o suficiente para gerar informações em qualquer perspectivas. Por exemplo, métricas como quantidade e valores de determinados itens podem ser, em geral, sumarizados por data (dia, mês ou ano), local, clientes, entre outras dimensões, sem perder a consistência da informação.

As métricas semi-aditivas são aquelas que podem ser sumarizadas em alguns casos. Isso porque a depender da situação empregada à métrica, ela pode perder sentido para a análise caso seja agregada. Neste caso a sumarização só fará sentido com algumas dimensões específicas. Por exemplo, a métrica saldo bancário. O saldo é um valor que reflete a situação atual da conta, que pode ter o saldo credor ou devedor. Faria sentido, por exemplo, somar os saldos de todos os dias de um mês para uma determinada conta bancária? Claro que não. Pois se um dia o saldo for de -1000 e no dia seguinte ter os mesmos -1000, a soma irá me devolver um saldo negativo de -2000, o que não é verdade. Mas há casos onde a métrica semi-aditiva adquire característica de aditiva. Se por acaso somar os saldos de várias contas bancária em um determinado dia, poderemos ver o saldo geral, o que tem total sentido e utilidade para uma instituição bancária, por exemplo.

As métricas não-aditivas são aquelas que não podem ser sumarizadas ao longo das dimensões. Essas métricas não podem ter agregações pois perdem a veracidade do valor. Percentuais são

exemplos de valores armazenados nas métricas que não permitem sumarizações. Por exemplo, não faz sentido algum somar o percentual de vendas de um item "A" que teve 50% de saída com um item "B" que teve 60%. A soma resultaria em um valor agregado de 110%. O que isso nos diz? Nada! Como muitas vezes as métricas semi-aditivas e não-aditivas são derivadas de métricas aditivas, recomendo, se possível, que sejam armazenadas as métricas brutas para o cálculo em tempo de execução. A métrica semi-aditiva saldo, por exemplo, pode ser calculada em tempo de execução com as métricas aditivas do valor de crédito e débito. Portanto devemos ficar atentos a essas diferenças, para que no desenvolvimento do DW possamos efetuar o tratamento adequado em cada um desses casos. Lembrando que quanto menos flexível for a utilização das métricas, mais complexo será a utilização pelos utilizadores, o que pode ser um fator crítico de sucesso para o projeto. Sempre que possível devemos gerar as agregações em tempo de execução para as métricas semi-aditivas e não-aditivas, facilitando a utilização e deixando transparente aos utilizadores os cálculo efetuados.

## Agregação

Agregação é um mecanismo de cálculo que a partir da tabela fatos com um alto grau de atomicidade dos atributos produz um resultado sumariado. Através da agregação dos dados no nível de detalhe desejado para as operações de consulta, estas oferecem um desempenho muito superior comparado as consultas nos dados que não estão sumarizados. A processo de agregação requer a definição de uma função (sum() ou avg() por exemplo) que será empregada sobre os atributos que caracterizam cada instância gerando um valor que irá representar todo o conjunto. Os atributos dos fatos que serão agregados podem ser aditivos, semi-aditivos ou não aditivos. Normalmente neste processo é utilizado pelo menos um atributo de uma das dimensões em relação á tabela Fatos.

## Classificação de FATOS

Com base nessas característica os factos podem ser classificadas em :

- √ Fatos Aditivos
- √ Não aditivos
- √ Semi aditivos
- √ Fatos aditivos são aqueles que podem ser somados relativamente a todas as dimensões.

Ex: Se for considerado cubo de venda o atributo Quantidade, Valor na Tabela Itens • São númericos e podem ser somados em relação às dimensões existentes

Ex: quantidade e valor podem ser somados ao longo de qualquer dimensão (Produto, Promoção, Atendente, Loja e Data)

Indicadores para Fatos aditivos aparece sempre que aparece um campo de dados numérico na modelação

Em geral, fatos aditivos representam medidas de atividade do negócio, ligadas ao seus indicadores de desempenho (KPI – Key performance indicators).

## Fato Semi-Aditivo

São aqueles que podem ser somados relativamente a algumas dimensões ou mesmo a uma única dimensão. Ex. Quantidade na Tabela Estoque • Também são numéricos, mas não podem ser somados em relação a todas as dimensões existentes (a semântica não permite) – Ex: quantidade em estoque só pode ser somada ao longo da dimensão Produto. Nas dimensões Loja e Data, a soma não faria muito sentido (especialmente nesta última, nenhum sentido)

Em geral, fatos semi-aditivos representam leituras medidas de intensidade do negócio. – São snapshots destas leituras que entram no DW. – O valor atual já leva em consideração valores passados. • Fatos semi-aditivos típicos: Níveis de Estoque, Saldos, Fechamento diário/mensal de conta, etc...

## Fatos não Aditivos

São aqueles que não podem ser somadas em relação a nenhuma dimensão.

Mas também existem casos em que textos podem ser considerados eventualmente Fatos. Ex: Data warehouse de registo de acidente transito ou casos com percentagens e os Rácios.

**Tipo de Tabelas FATOS**

As tabelas de fatos encontram se dividas em dois grandes grupos:

- As que contem regista medidas ou propriamente ditas Fatos
- As que registam acontecimento que também são denominadas de tabelas sem fatos

**Tipos de Dimensões**

**Dimensões Degeneradas**

Dimensão sem tabela, ela fica em uma chave na tabela de fatos, ela tem a mesma cardinalidade/granularidade da tabela de fatos.

**Dimensões Lixo (Junk)**

É um conjunto de códigos aleatórios, flags, atributos de texto, baixa cardinalidade, consequentemente removendo flags da tabela de fatos.

**Dimensões com Vários Papéis (Conformada)**

Quando uma mesma dimensão aparece várias vezes na mesma tabela de fatos. Cada role da dimensão é representado por uma tabela lógica. O exemplo mais falado nesse cado é a dimensão tempo.

**Mini-Dimensões**

Subconjuntos de uma dimensão grande, por exemplo Cliente, que são divididos em dimensões artificiais menores para controlar o crescimento explosivo de uma dimensão grande de mudança rápida.

**Dimensões Multivaloradas (Bridge Table)**

São tabelas pontes, associativa, tabela ajuda, é no caso da modelagem (ainda irei abordar) um floco do modelo "Floco de Neve/Snow Flacke".

*Dinâmica das Dimensões*

Slowly Changing Dimensions

- √ Tipo 1 – Atualiza o valor já existente. Não funciona para bases Históricas por não preservar o histórico.
- √ Tipo 2 – Adicionar uma nova linha com o novo valor do atributo, mantendo os outros valores.
- √ Tipo 3 – Adicionar uma coluna com o novo valor do atributo, mantendo os outros valores.

O modelo de dados multidimensional é caracterizado por valores, geralmente contínuos, distribuídos em uma ou mais dimensões, como por exemplo, o total de vendas de um produto agrupado por mês e por loja. Para um produto, em um determinado mês e uma loja específica, existe apenas um total de vendas. Pode-se imaginar estes dados através de um cubo tridimensional onde uma dimensão representa os produtos, a outra representa as lojas e a terceira representa o mês/ano.

A visão multidimensional irá auxiliar as empresas no aprimoramento de seus indicadores através dos ajustes de seus sistemas internos, além de promover uma nova cultura de análise e gestão da informação nos diversos níveis da organização.

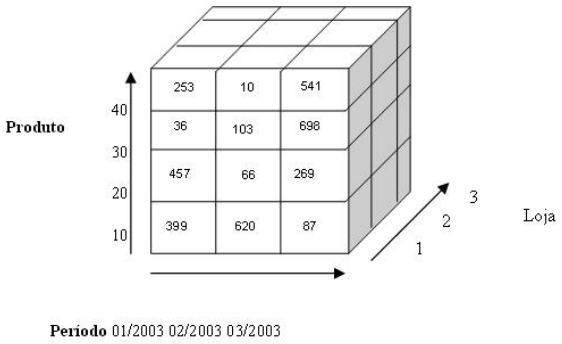



Figura 1:Cubo Tridimensional com o Total de Vendas

Fonte: CAMPOS, 2005.

Para a estruturação do data warehouse, existem dois tipos de modelos que podem ser utilizadas, a Star Squema, ou Modelo Estrela, e a Snowflake.

As Dimensões são os descritores dos dados oriundos da Fato. Possui o caráter qualitativo da informação e relacionamento de "um para muitos" com a tabela Fato. É a Dimensão que permite a visualização das informações por diversos aspectos e perspectivas.

Segundo KIMBALL, as tabelas de dimensão não devem ser normalizadas pois:

1. Não há atualização freqüente nas bases;
2. O espaço em disco economizado é relativamente pequeno;
3. Esse ganho de espaço não justifica a perda de performance na realização de consultas por conta das junções necessárias em caso de normalização

A vantagem de utilizar dimensões é que os fatos podem ser vistos sob diferentes perspectivas. Neste exemplo das vendas, o total de vendas pode ser apresentado por produto, por cliente, por loja ou por vendedor. O interessante também dos dados multidimensionais é que as dimensões podem ser cruzadas: por exemplo, comparar a idade do cliente com o preço do produto. Tal tipo de cruzamento nos dará informações que não poderiam ser vistas no modelo relacional

A estrutura relacional diferencia-se da estrutura multidimensional principalmente devido a normalização, pouca redundância e a frequência de atualizações suportadas. A estrutura multidimensional possui, normalmente, desnormalização de tabelas, alta redundância e suporta periodicidade de atualizações de dados muito menor do que uma estrutura relacional convencional.

No modelo dimensional visa armazenar uma imensa quantidade de dados. Um modelo que vai permitir uma análise histórica dos dados, desempenho nas consultas e facilidade no desenvolvimento das mesmas

A modelagem dimensional possui dois modelos: o modelo estrela (star schema) e o modelo floco de neve (snow flake). Cada um com aplicabilidade diferente a depender da especificidade do problema. As Dimensões do modelo estrela são desnormalizados, ao contrário do snow flake, que parcialmente possui normalização.

Segundo Caldeira, Carlos Pampulim as Etapas no desenho do modelo de Dados Dimensional são:

1. Escolha dos processos de negócio;
2. Declaração do grão ou nível de detalhe;
3. Escolha das dimensões:
4. Identificação dos factos ou parametros medidos

Os Processos são actividades que produzem resultados susceptíveis de serem armazenados em tabela fatos. Os fatos podem ser qualitativos ou quantitativos. Num processo pode estar associado diferentes métricas a serem analisados, com determinadas características de dimensionalidade e granularidades que têm de ser colocadas á disposição do utilizador como forma de lhe ajudar a analisar os Fatos.

## Modelo de Estrela

O modelo estrela é composto por uma tabela central, a tabela fato, com múltiplas junções com outras tabelas chamadas de tabelas de dimensão. Cada uma das tabelas dimensão possui apenas uma junção com a tabela fato. O modelo estrela, mostrado na Figura 12, tem a vantagem de ser simples intuitivo e de fácil compreensão.

A Tabela de Fato contém as métricas que vão serem analizados. Possui um caráter quantitativo das informações descritivas armazenadas nas Dimensões. É onde estão armazenadas as ocorrências do negócio e possui relacionamento de "muitos para um" com as tabelas periféricas (Dimensão).

A estrutura dimensional normalmente é desenhada no formado do esquema estrela (star schema). Nesse modelo, as tabelas de Dimensões são ligadas diretamente a tabela Fato. Outra característica marcante é que os dados são desnormalizados, pois a redundância resultante gera benefícios para a otimização das consultas e navegação das informações.

A imagem a seguir permite uma visualização simples do que seria um esquema estrela na modelagem de dados multidimensional:

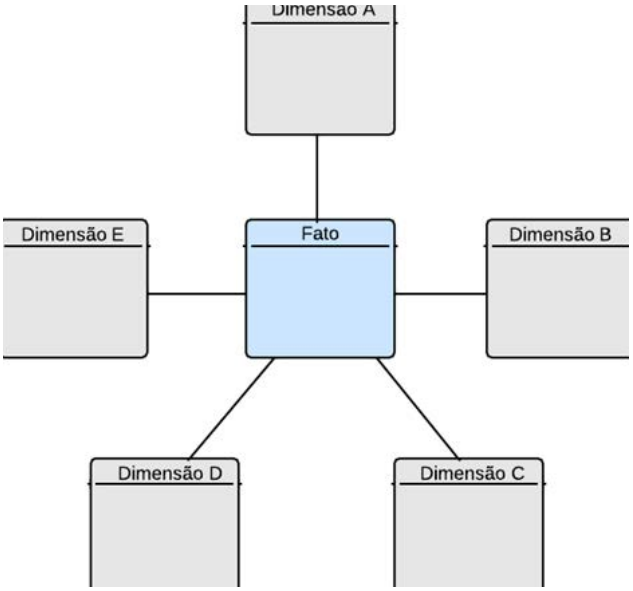



Figura 2:Modelo de Estrela

Perceba que a Fato fica centralizada no modelo e as Dimensões (tabela auxiliares) ficam na zona periférica, fazendo assim a analogia com o formato estelar.

A Fato possui característica quantitativa dentro do DW. A partir dela são extraídas as métricas que são cruzadas com os dados das Dimensões, concebendo, assim, informações significativas para a análise do utilizador. A Fato armazena as medições necessárias para avaliar o assunto pretendido. O conteúdo histórico no DW, contendo longo período de tempo, ficam depositadas na Fato.

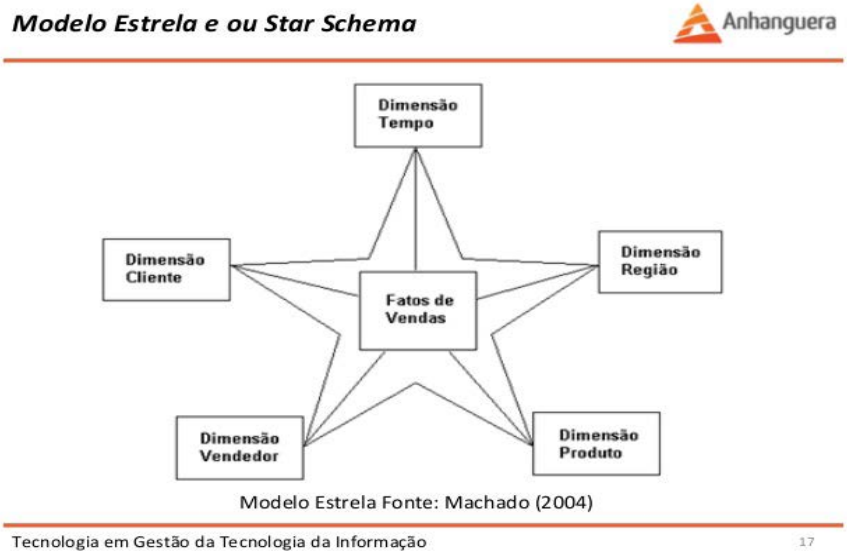



Figura 3:Modelo de Estrela

As Dimensões e Fatos são componentes complementares e dependentes entre si. Em um modelo dimensional é obrigatório a existência de ambos. Sem um desses elementos, a compreensão e análise das informações ficam comprometidas no modelo dimensional, ou até mesmo inviabilizadas, porque na tabela FATO serão carregadas as métricas, isto é, o que vai ser medido enquanto que as tabelas dimensões são carregadas os atributos que são utilizadas para descrever os dados na tabela fato.

Exemplo de DW

Vamos imaginar que o problema seja a falta de informações para gerar estatística das lojas que mais vendem, dos funcionários mais empenhados, dos produtos mais vendidos e até mesmo do período que mais tem movimento.

Na situação hipotética, teríamos que ter no mínimo quatro Dimensões: Funcionário, Produto, Loja e Data. As Dimensões normalmente possuem muitos campos descritores, porém, neste exemplo, iremos supor que foram solicitados uma quantidade restrita de campos, apenas para a melhor compreensão do modelo.

No levantamento das necessidades dos utilizadores, por exemplo, foram solicitados os seguintes campos para a análise das informações:

*Dimensão Funcionário*

Nome, cargo e setor do funcionário.

*Dimensão Produto*

Nome, descrição e marca do produto.

*Dimensão Loja*

Nome, cidade, estado, data de inauguração e CEP da loja.

*Dimensão Data*

Dia, mês, ano, bimestre, trimestre, semestre, feriados e fim de semana da data do registo da venda.

*Fato Vendas*

Quantidade e valor de itens vendidos.

Supondo também que todos os campos existem nos sistemas operacionais da empresa, logo o modelo dimensional resultante teria a seguinte estrutura:

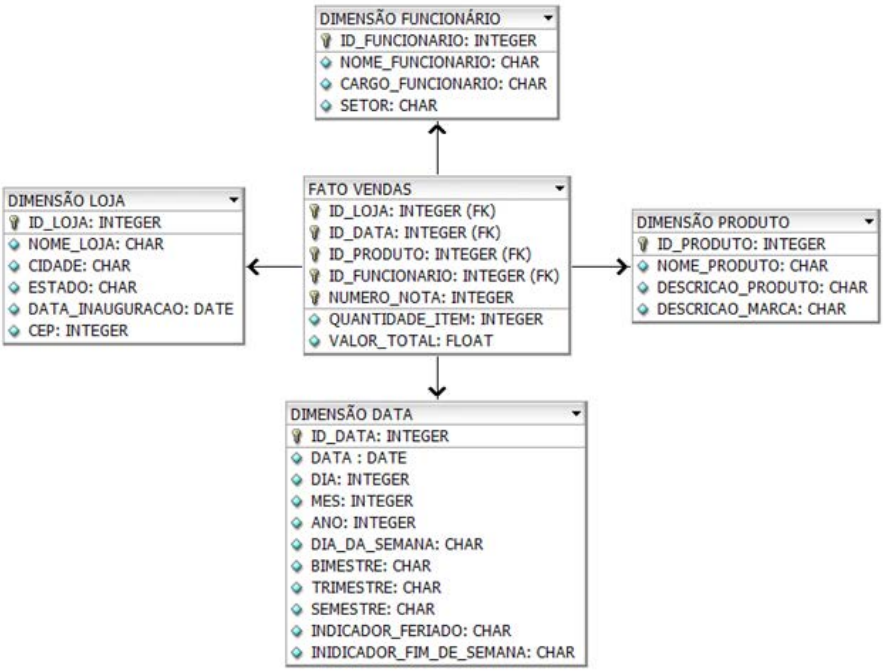



Figura 4:Modelo Estrela do negócio

O campo "NUMERO_NOTA" (que seria o número da transação) na Fato Vendas se trata de uma Dimensão Degenerada (Degenerate Dimension), devido ao campo não possuir tabela de Dimensão correspondente. Ela também serve como parte integrante da chave da Fato Vendas e permite o agrupamento dos produtos vendidos em um único documento (pedido), pois apesar de uma compra poder ter vários itens, os dados de cada produto são armazenados separadamente na Fato.

O modelo do exemplo possui estrutura básica para responder as solicitações hipotéticas do utilizador, como também fazer outras perguntas como:

- Qual a loja que vendeu mais bebidas no último ano?
- Qual o trimestre que são vendidos mais produtos?
- Qual é o produto mais vendido no período de maior movimento?

Além disso, esse modelo pode fazer inúmeros outros cruzamentos que serão definidos a partir da visão analítica do utilizador, possibilitando a geração de valiosos insights para a organização. Tudo isso de forma intuitiva e com a velocidade adequada exigida pelos negócios empresarias.

## Floco de Neve

No modelo, Snowflake, há a decomposição de uma ou mais dimensões do modelo estrela que possuem hierarquias entre seus membros. Machado [1] explica que os dois modelos acessam o mesmo fato, mas são desenvolvidos de forma diferente e a diferença está no fato do modelo estrela possuir as mesmas dimensões, porém estão em uma só tabela e o modelo snowflake é separado em hierarquias, uma tabela para cada nível. O fato é uma coleção de dados que representam um evento de negócio da empresa e é utilizado para analisar o processo de negócio de uma empresa. As dimensões são as possíveis formas de se visualizar os dados e a medida é um atributo numérico que representa um fato. O termo extração de dados, segundo Silberchatz, Korth e Sudarshan, refere-se à busca, no sentido de garimpar ou minerar, de informações relevantes, ou à descoberta de conhecimento, a partir de um grande volume de dados.

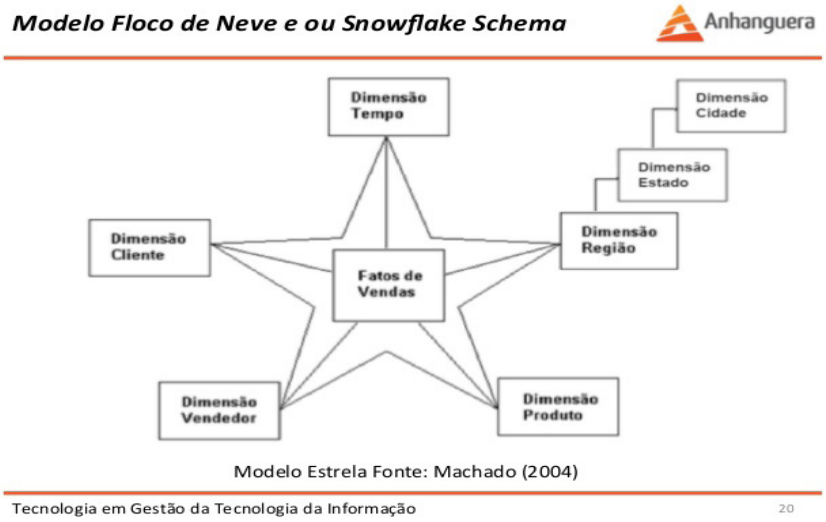



Figura 5:Modelo de Floco de Neve

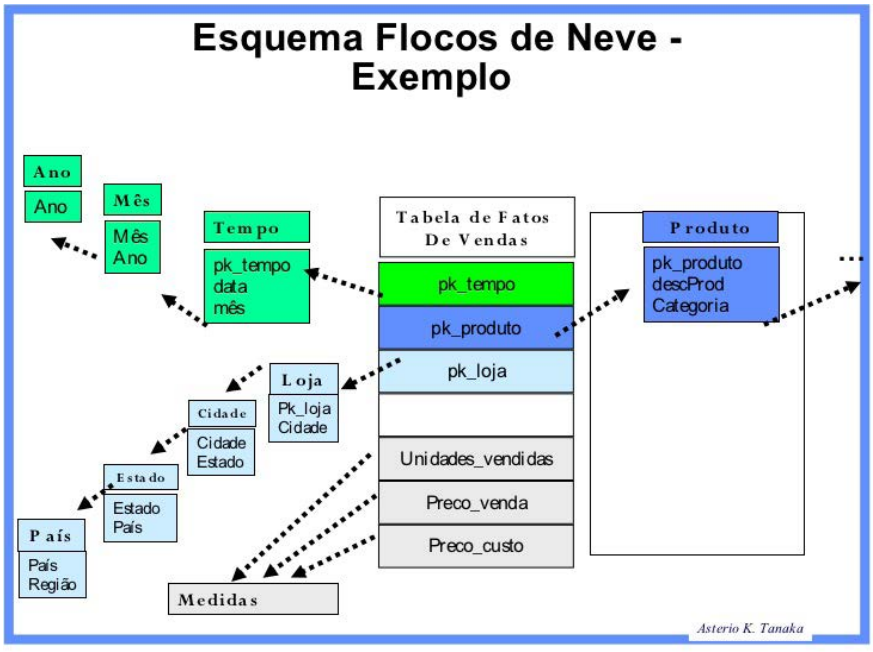



Figura 6:Modelo de Negocio Floco de Neve

A decisão relativamente á escolha do modelo a utilizar será respondida no levantamento de requisitos e na análise dos dados e do modelo transacional.

A Importância da Modelagem de Dados para Garantir a Qualidade

A modelagem de dados é um modo eficiente de entender os dados. O seu propósito é prover um registo apurado de alguns aspectos do mundo real para um contexto particular.

Através da modelagem, o projetista do banco de dados pode eliminar redundâncias, que representam algumas das fontes de informações inconsistentes e podem levar a sistemas ineficientes (COUGO, 1997).

Um modelo de dados é uma coleção de conceitos que podem ser utilizados para descrever um conjunto de dados e as operações para a sua manipulação (BATINE, CERI, NAVATHE, 1992). A não utilização de um modelo implica em um crescimento desorganizado das aplicações, promovendo altos custos, esforços para a manutenção e dificuldades no crescimento de uma aplicação. O modelo permite ao utilizador um melhor entendimento do negócio, com a vantagem de facilitar a visualização das conseqüências de qualquer ação dentro do ambiente e o impacto de qualquer mudança sobre o mesmo (DEVLIN, 1997). Ele representa a definição, caracterização e relacionamento dos dados em um determinado ambiente. Um dos diagramas mais empregados na modelagem de dados é o Diagrama de Entidade e Relacionamento (DER).

A modelagem deve sempre que possível levar em consideração a possibilidade de futuras implementações nas organizações. A capacidade de reconhecer antecipadamente estas necessidades dependerá em grande parte da experiência e conhecimento da equipe de desenvolvimento em relação às necessidades da organização. A modelagem deve levar em consideração aspectos como: aquisição dos dados; arquivamentos dos registos de dados; padronização dos dados; comunicação e integração; armazenamento e recuperação das informações e análise dos dados para levantamento(WEN, 2007).

## XIV. Actividades de aprendizagem

Uma grande distribuidora de filmes possui sistema para controle dos seus filmes.

O sistema atual controla os filmes por sala de cinema onde são exibidos, tendo informações sobre a capacidade de lotação de casa sala, localização regional no país, assim como os registros de bilheteria de cada sessão diária de cinema.

O Sistema atual

Para permitir localização rápida de um filme, o sistema controla os atores que participam do elenco de cada filme exibido, assim como o diretor do filme, que pode também participar como ator no mesmo filme.

Os filmes são classificados por gênero, por origem (país do estúdio).

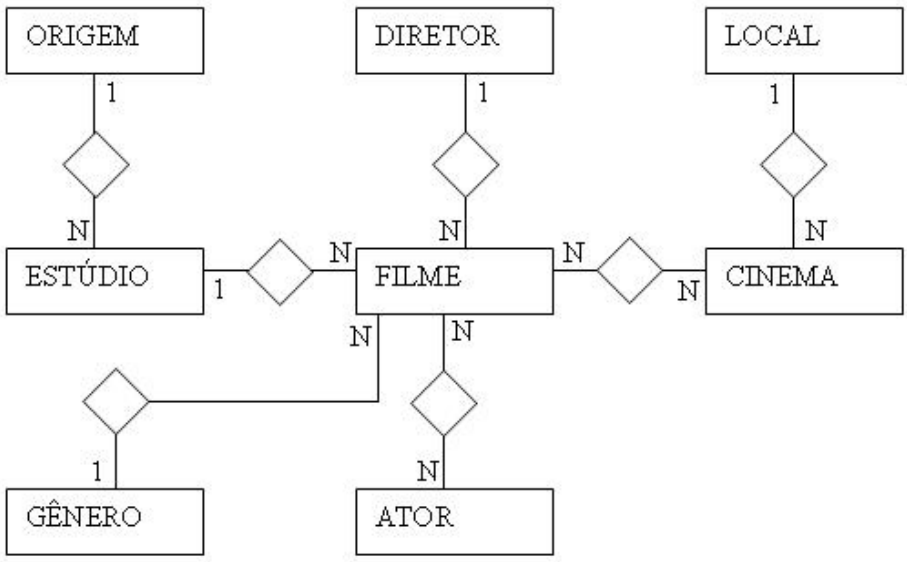
As sessões de cinema são controladas para fins de quantidade de publico e valor arrecadado de bilheteria. O cliente nos forneceu um modelo de dados do sistema atual.

As necessidades Executivas

Nas entrevistas para entendimento dos requisitos para análise estatística e criação de um Data Warehouse, foram apresentadas as seguintes necessidades:

1. Os gerentes de área da distribuidora desejam acompanhar mensalmente a evolução do público e valor arrecadado em nível de região do país, estado e cidade, classificados por gênero de filme e sala de cinema.
2. Também é necessário avaliar a evolução de filmes por ator participante, assim como por diretor.
3. Queremos saber quais os diretores que atraem maior público e em que gênero está esse público.
4. O tempo é fator fundamental de análise, pois temos de ter a visão de quais períodos do ano possuem mais publico por gênero, ator e diretor geograficamente.

**MER do sistema atual:**




1. Identifique os candidatos ás tabelas fatos e dimensões com base nas necessidades e classifique-as.
2. Com base no cenário apresentado desenho os modelos de estrela e floco de neve.

# Unidade 3: Extraindo Informações Do Data Warehouse

- Potencial de informações num data warehouse
- Extraindo e transformando dados;
- Análise Multi-dimensional e OLAP
- Interrogação Multi-Dimensional de Dados
- Ferramentas e operações OLAP
- MODELOS OLAP
  - MOLAP
  - ROLAP

## Objectivos :

1. Explicar os conceitos de ferramentas analiticas
2. Explicar a arquitectura de modelos OLAP
3. Utilizar operadores olap para extracção de informação

### Termos-chave

**OLTP:** (Online Transaction Processing) - Processamento de transações em tempo-real. São sistemas que se encarregam de registrar todas as transações contidas em uma determinada operação organizacional. Por exemplo: sistema de transações bancárias registra todas as operações efetuadas em um banco.

Base de dados - coleção de dados interelacionados logicamente, ex: agenda de telefones.

**SGBD (Sistema Gerenciador de Banco de Dados):** é um software com recursos

específicos para facilitar a manipulação das informações de um BD e o desenvolvimento de programas aplicativos. Exemplos: Oracle, Paradox, MySQL, Access, Interbase, Sybase.

**Tabelas:** Representam as estruturas de armazenamento de dados dos sistemas Gestão da Base de Dados,compostos por colunas (Atribuots) e linhas (Registo).

**Drill Down:** É o método de exploração de dados detalhados que foram usados na criação de dados sumarizados, resumidos. Os níveis de Drill down dependem da granularidade dos dados no Data warehouse.

**Drill UP:** É o método de exploração de dados para o nível acima (contrário do Drill Down), ou seja, nível menos detalhado.

**EAI** (do inglês Enterprise Application Integration): é uma referência aos meios computacionais e aos princípios de arquitetura de sistemas utilizados no processo de Integração de Aplicações Corporativas. Os procedimentos e ferramentas de EAI viabilizam a interação entre sistemas corporativos heterogêneos por meio da utilização de serviços.

**ETL (Extract Transform Load) – Extração, Transformação e Carga:** São ferramentas de software projetadas para "fundir" dados em uma única forma, aplicando regras de transformação. "Extração" se refere a remover os dados de múltiplas fontes de dados transacionais, incluindo arquivos e sistemas legados; "Transformação" se refere à conversão dos dados, incluindo limpeza, agregação e filtragem; e "Carregamento" se refere ao processo de colocar os dados convertidos e fundidos em uma nova base de dados estruturada corretamente para as atividades de consulta e análise.

**SBD (Sistema de Banco de Dados):** é um sistema de manutenção de dados envolvendo quatro componentes principais: dados, hardware, software e utilizadores.

**Gerenciadores De Banco De Dados;** Conjunto de programas (software) para gerenciar (criando, modificando e usando) um banco de dados e garantir a integridade e segurança dos dados. São exemplos de "Sistemas Gerenciadores de Banco de Dados": MYSQL, SQL Server, Oracle, DB2, ADABAS etc.

**Dado:** factos em "bruto", ou seja, a mais pequena p qp porção de dados que o computador reconhece ex:pode ser um número, um caracter,...;

**Campo :** um caracter ou conjunto de caracteres com um significado específico como por exemplo um nome; significado específico, como por exemplo um nome;

**SQL (Structured Query Language):** Linguagem de Consulta Estruturada. É uma linguagem padrão de comunicação com sistemas gerenciadores de banco de dados (SGBD), utilizada para acessar sistemas de banco de dados relacional.

## Extração de dados

A infra-estrutura de uma aplicação BI deve ser feita de forma a existir um ambiente para que os dados sejam registados e um ambiente para que de fato sejam analisados.

Para analisar os dados pode ser considerado Dois Ambientes:

- Sem Data Warehouse
- Com Data Warehouse

Por isso, precisa se de ferramentas para extracção e manipulação de dados em função do ambiente: Sendo assim fala se de OLTP e OLAP.

Ambiente Transacional e Ambiente Analítico. No Ambiente Transacional,que é a mesma que ambiente operacional, também chamado de OLTP (Online Transaction Processing) ocorrem as transações com o banco de dados,(como num banco comercial) ou seja, é nele que os dados são registados, eliminados e atualizados. No Ambiente Analítico, também chamado de OLAP (Online Analytical Processing) ocorrem as consultas e análises,isto é,um ambiente em que os gestores precisão de visão agregada dos dados.Por isso a operação OLAP é exercicida sobre um armanzem de dados (Data warehouse). Segue abaixo uma explicação detalhada de cada um dos ambientes:

Ambiente Transacional (OLTP): O termo Online Transaction Processing (OLTP) se refere a Sistemas de Informação capazes de efetuar transações, como registo,eliminar e atualizações de dados. Muitas empresas possuem sistemas para esse propósito, porém muitos deles antigos e de difícil manutenção. São os chamados Sistemas Legados. O ambiente OLTP pode ser composto por várias fontes de dados, tais como: Bancos de Dados Relacionais e até mesmo folha de calculo ou arquivos texto.

Ambiente Analítico (OLAP): O Ambiente Analítico, ou Online Analytical Processing (OLAP) tem o objetivo de dar suporte a decisões, portanto, dar suporte às funções associadas à concepção do negócio. O OLAP executa basicamente consultas, que fornecem dados relevantes ao usuário. As consultas são "ad-hoc", ou seja, podem ser montadas pelo usuário em tempo de execução, sem a necessidade da utilização de uma linguagem de busca como a SQL. O Ambiente Analítico é composto pelo Cubo OLAP,sendo assim é de extrema importância o papel de modelagem dimensional na projecção de datawarehouse, que relaciona as dimensões de análise e pelo Front-End OLAP, que oferece uma interface para que o ciente execute as consultas "ad-hoc".

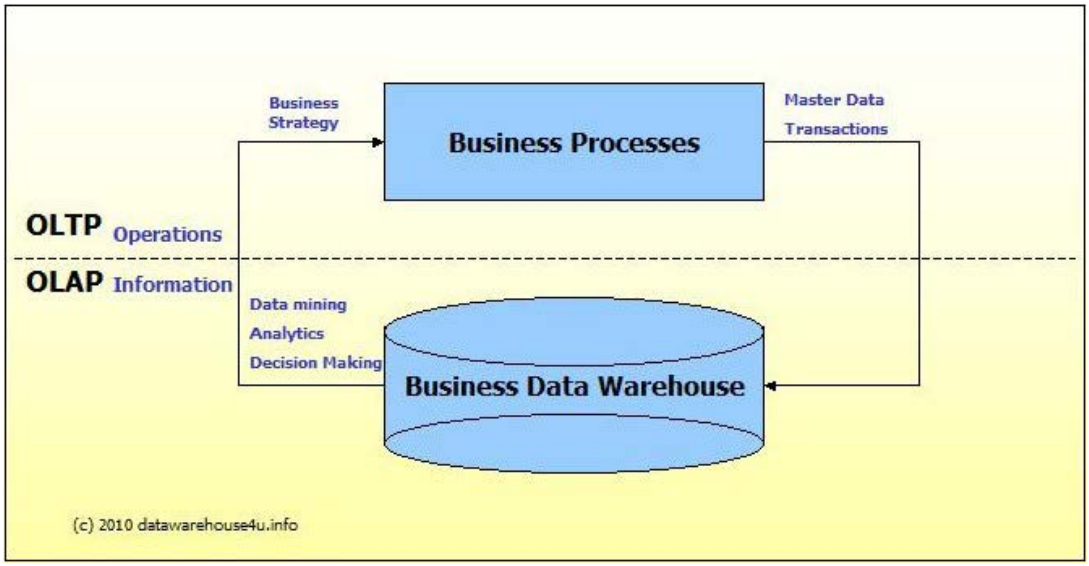



Figura 1:OLAP vs OLPT

A figura 17 mostra o funcionamento das Ferramentas . No Ambiente Transacional, temos um Banco de Dados Relacional, no qual os dados são extraídos. Os dados extraídos do Banco de Dados Relacional são levados para um banco de dados temporário, chamado Banco de Dados Staging. Estes dados são transformados através de alguns scripts e carregados em um Banco de Dados Multidimensional, conhecido como Data Warehouse. Esse processo de "extração, transformação e carga" é conhecido pela sigla ETL (Extract-Transform-Load). Depois desse processo, os dados do Data Warehouse são estruturados na forma de Cubo, no Ambiente Analítico. O Cubo OLAP alimenta o Front-End OLAP, que será de fato a interface com o usuário final.

## OLAP

Partindo dos primórdios da informatização, quando um sistema que gerava relatórios era a principal fonte de dados que os decisores disponham. Toda vez que uma análise necessitasse ser feita, eram necessários produzir novos relatórios. Estes relatórios tinham que ser produzidos pelo departamento da informática e, normalmente, precisavam de muito tempo para ficar prontos. E, também, apresentavam os seguintes problemas:

- Os relatórios eram estáticos;
- O acúmulo de diferentes tipos de relatórios num sistema gerava um problema de manutenção.

Como os decisores precisam de dados e não despoem por vezes de conhecimentos de linguagem SQL,que lhes permite manipular os dados ficam por vezes na dependência de especialista da área o que dificulta na tomada de decisão.

Então surgiu o conceito de OLAP (On-Line Analytic Processing).

OLAP é um software cuja tecnologia de construção permite aos analistas de negócios, gerentes e executivos analisar e visualizar dados corporativos de forma rápida, consistente e principalmente interativa. A funcionalidade OLAP é inicialmente caracterizada pela análise dinâmica e multidimensional dos dados consolidados de uma organização permitindo que as atividades do utilizador final sejam tanto analíticas quanto navegacionais A tecnologia OLAP é geralmente implementada em ambiente multi-utilizador e cliente/servidor, oferecendo assim respostas rápidas às consultas ad-hoc (construção de listagens, interligando a informação disponível na base de dados conforme as necessidades especificas da empresa, assim como a sua exportação, possibilitando várias simulações), não importando o tamanho do banco de dados nem sua complexidade. Hoje em dia, essa tecnologia também vem sendo disponibilizada em ambiente Web. Essa tecnologia auxilia o utilizador a sintetizar informações corporativas por meio de visões comparativas e personalizadas, análises históricas, projeções e elaborações de cenários.

Para tornar o problema ainda mais complexo, nem os profissionais de tecnologia da informação (TI) nem os utilizadores finais estão suficientemente informados sobre que produtos OLAP adquirir. O processo de avaliação de uma ferramenta OLAP envolve análise das: funcionalidades, arquitetura, e interfaces e impacto sobre a organização.

O processo de aquisição de uma ferramenta OLAP deveria envolver não só os profissionais de TI como também os grupo de utilizadores finais. Porém, nem sempre esta premissa é verdadeira. Diversas organizações relatam sérios problemas de comunicação entre as equipes envolvidas na escolha de uma ferramenta OLAP.

A escolha de uma ferramenta OLAP inadequada pode ocasionar severas conseqüências para um projeto de datawarehouse, entre as quais podemos citar:

- Falha total do projeto e conseqüente perda dos benefícios esperados para os negócios da empresa, além dos prejuízos financeiros gerados pelo alto custo da aquisição de software, serviços e treinamentos das equipes iniciais do projeto resultando benefícios ilusórios e temporários. Esta situação é muito comum em diversas organizações e sob certos aspectos gera piores resultados.
- Falha parcial do projeto onde apenas alguns módulos sobrevivem, reduzindo assim o escopo, estando comparados com o item anterior, uma vez que passam a existir menores incentivos para substituir os sistemas atuais.

*Funcionalidades da Ferramenta OLAP*

A funcionalidade de uma ferramenta OLAP é caracterizada pela análise multidimensional dinâmica dos dados, apoiando o utilizador final nas suas atividades, tais como: Slice and Dice e Drill.

## Características da Análise OLAP / Operadores Olap

**Drill Across:** O Drill Across ocorre quando o utilizador pula um nível intermediário dentro de uma mesma dimensão. Por exemplo: a dimensão tempo é composta por ano, semestre, trimestre, mês e dia. O utilizador estará executando um Drill Across quando ele passar de ano direto para semestre ou mês.

**Drill Down:** O Drill Down ocorre quando o utilizador aumenta o nível de detalhe da informação, diminuindo o grau de granularidade.

**Drill Up:** O Drill Up é o contrário do Drill Down, ele ocorre quando o utilizador aumenta o grau de granularidade, diminuindo o nível de detalhamento da informação.

**Drill Throught:** O Drill Throught ocorre quando o utilizador passa de uma informação contida em uma dimensão para uma outra. Por exemplo: Estou na dimensão de tempo e no próximo passo começo a analisar a informação por região.

**Slice And Dice:** O Slice and Dice é uma das principais características de uma ferramenta OLAP. Como a ferramenta OLAP recupera o microcubo, surgiu a necessidade de criar um módulo que se convencionou de Slice and Dice para ficar responsável por trabalhar esta informação. Ele serve para modificar a posição de uma informação, alterar linhas por colunas de maneira a facilitar a compreensão dos utilizadores e girar o cubo sempre que tiver necessidade.

**Alertas:** Os Alertas são utilizados para indicar situações de destaque em elementos dos relatórios, baseados em condições envolvendo objetos e variáveis. Servem para indicar valores mediante condições mas não para isolar dados pelas mesmas.

**Ranking:** A opção de ranking permite agrupar resultados por ordem de maiores / menores, baseado em objetos numéricos (Measures).Esta opção impacta somente uma tabela direcionada (relatório) não afetando a pesquisa (Query).

**Filtros**: Os dados selecionados por uma Query podem ser submetidos a condições para a leitura na fonte de dados. Os dados já recuperados pelo Utilizador podem ser novamente "filtrados" para facilitar análises diretamente no documento.

**Sorts:** Os sorts servem para ordenar uma informação. Esta ordenação pode ser customizada, crescente ou decrescente.

**Breaks:** Os Breaks servem para separar o relatório em grupos de informações (blocos). Por exemplo: O utilizador tem a necessidade de visualizar a informação por cidades, então ele deve solicitar um Break. Após esta ação ter sido executada, automaticamente o relatório será agrupado por cidades, somando os valores mensuráveis por cidades.

Consultas Ad-Hoc: São consultas com acesso casual único e tratamento dos dados segundo parâmetros nunca antes utilizados, geralmente executado de forma iterativa e heurística.

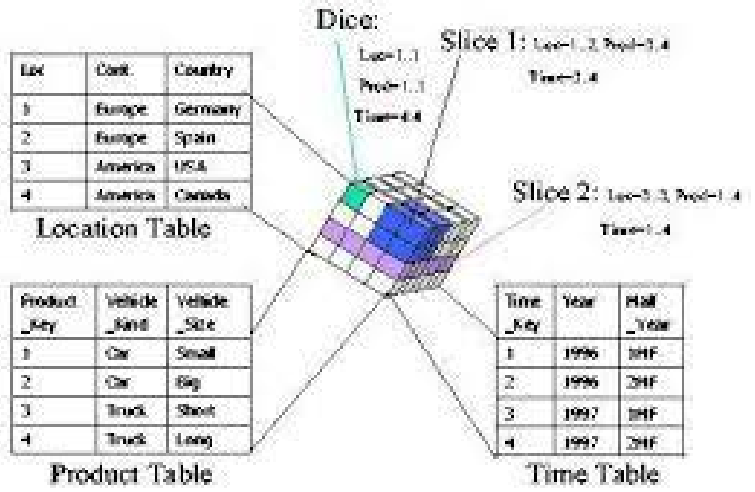



Figura 2:Slice & Dice

## Arquiteturas OLAP

As grandes quantidades de nomenclaturas que estão aparecendo, são algumas variações de estrutura de OLAP. A Tecnologia surgiu com a evolução dos Sistemas de Informação. Esses Sistemas, no seu começo, armazenavam grandes quantidades de dados, mas a recuperação dos mesmos era um tormento para os utilizadors finais e os analistas.

Os SGBD's (Sistemas Gerenciadores de Banco de Dados) foram evoluindo junto com as linguagens de programação, o que facilitava um pouco a vida dos analistas. Montagens de Dados já poderiam ser acessadas de uma maneira um pouco mais simples.

Acompanhando a evolução dos sistemas, introduziram uma nova classe de Ferramentas no mercado, que se chamava de OLAP (On Line Analitical Processing), que permitiam acesso rápido aos dados conjugado com funcionalidades de análise multidimensional dos mesmos pelos utilizadores finais. A rapidez exigida tinha de ser satisfatória. A análise deveria ser dinâmica, onde o utilizador poderia fazer a consulta que quisesse sem depender de um profissional da área das tecnologias.

A análise multidimensional é uma das grandes utilidades da tecnologia OLAP, consistindo em ver determinados cubos de informações de diferentes ângulos e de vários níveis de agregação. Os "cubos" são massas de dados que retornam das consultas feitas ao banco de dados e podem ser manipulados e visualizados por inúmeros ângulos e diferentes níveis de agregação. As ferramentas que disparam uma instrução SQL, de um cliente qualquer, para o servidor e recebem o microcubo de informações de volta para se analisado na Workstation, chamam-se DOLAP (Desktop On Line Analitical Processing).

O Slice/Dice é uma das principais características de uma ferramenta OLAP. É uma operação com responsabilidade de recuperar o micro-cubo dentro do OLAP, além de servir para modificar a posição de uma informação, alterar linhas por colunas de maneira a facilitar a compreensão dos utilizadors e girar o cubo sempre que tiver necessidade.

## Caracteristicas Do Olap X Olpt

| CARACTERÍSTICAS | APLICAÇÃO OLTP | APLICAÇÃO OLAP |
|---|---|---|
| Operação típica | Atualização | Análise |
| Telas | Não alteráveis | Definidas pelo usuário |
| Dados por transação | Poucos | Muitos |
| Nível do dado | Detalhado | Muitos |
| Idade do dado | Atual | Histórico, atual e projetado |
| Idade do dado | Registos | Vetores, séries de tempo |

## Modelos OLAP

Entre os sistemas OLAP importa ainda distinguir ROLAP e MOLAP. Embora sejam ambos ferramentas de análise, os sistemas ROLAP (Relational Online Analytical Processing) diferem dos MOLAP (Multidimensional Online Analytical Processing) na medida em que não necessitam de qualquer processamento nem armazenamento prévio dos dados a processar, em oposição ao funcionamento dos sistemas MOLAP, que é o tradicional dos sistemas OLAP – a imagem da informação armazenada, já depois de algum processamento no cubo multidimensional.

Os sistemas ROLAP são capazes de aceder aos dados numa dada base de dados relacional e gerar as queries (consultas) necessárias no momento em que o utilizador as solicita. No entanto, é claro que a base de dados deverá estar convenientemente preparada para o efeito, e este processo é mais moroso que na tecnologia MOLAP.

Os sistemas HOLAP (Hybrid Online Analytical Processing) como o acrónimo indica, são uma mistura dos dois sistemas anteriores, tentando explorar as vantagens de ambos e colmatar as fraquezas de cada um.

Assim, para informação pré-processada, o HOLAP age como os sistemas MOLAP, imitando a rapidez destes. Quando é necessária informação mais detalhada, ainda não sumariada nem pré-processada, os HOLAP imitam os sistemas ROLAP na sua habilidade para capturarem a informação através de um sistema de bases de dados relacionais.

## Avaliação da Unidade

Verifique a sua compreensão!

1. A funcionalidade pré-programada de resumir os dados, com generalização crescente, oferecida pelas aplicações por meio das ferramentas de construção de data warehouses é denominadaParte superior do formulário

   a) Roll up. b) Drill down. c) Pivot. d) Sorting. e) Slice and disse

2. No âmbito dos DWs e OLAP, o processo onde se faz a junção dos dados e transforma-se as colunas em linhas e as linhas em colunas, gerando dados cruzados, é chamado de

   a) drill-across. b) star. c) cube. d) pivot. e) cross-joinParte inferior do formulário

3. Leia as afirmações a seguir:

   I. Um Data Warehouse é um repositório de dados atuais e históricos de uma organização que possibilita a análise de grande volume de dados para suportar a tomada de decisões estratégicas, possuindo registros permanentes.

   II. O processo de Data Mining, ou mineração de dados, tem por objetivo localizar possíveis informações em um banco de dados através de comparações com dados informados pelo usuário e registros de tabelas.

   III. Um ERP, ou Sistema Integrado de Gestão Empresarial, é conhecido por integrar os dados de diferentes departamentos de uma organização, aumentando o uso de interfaces manuais nos processos.

   IV. As ferramentas OLAP (On-line Analytical Processing) são capazes de analisar grandes volumes de dados, fornecendo diferentes perspectivas de visão e auxiliando usuários na sintetização de informações.
   Está correto o que se afirma APENAS em

   I e II. b) II e III. c) I, III e IV. d) I, II e III. e) I e IV.

4. Um processo importante que ocorre em relação à formação de um data warehouse é a obtenção dos dados de uma ou mais bases de dados da origem. Deve ser rigoroso para evitar a deformação e/ou a perda dos dados quando passados da fonte original para o destino. Trata- se deParte superior do formulário

   a) MINING. b) DATA MART. c) MOLAP. d) STAR. e) ETL.Parte inferior do formulário

5. No contexto de DW, é uma categoria de ferramentas de análise denominada open-end e que permite ao usuário avaliar tendências e padrões não conhecidos entre os dados. Trata-se de

Parte superior do formulário

a) slice. b) star schema. c) ODS. d) ETL. e) data mining.

6. As consultas no star schema de um data warehouse podem ser feitas em maior ou menor nível de detalhe. Assim uma consulta mais detalhada das informações denomina-se

Parte superior do formulário

a) drill-down. b) data mart. c) data mining. d) roll-up. e) snowflake.

7. A organização dos data warehouse em tabela de fato e tabelas de dimensão relacionadas, é característica

Parte superior do formulário

a) do esquema estrela. b) da mineração de dados. c) do roll-up.

d) do processador analítico on-line. e) do drill-down.

8. A seguir são feitas algumas afirmações a respeito de data warehouses e ferramentas OLAP.

I - Os usuários finais do data warehouse, em geral, não possuem acesso à Data Staging Area.

II - Drill in, drill out, roll over e roll on são típicas operações disponibilizadas pelas ferramentas de consultas OLAP para navegar pela hierarquia de uma dimensão.

III - As rotinas de ETL muitas vezes originam solicitações de mudanças e melhorias nos sistemas OLTP e outras fontes de dados que alimentam o data warehouse, pois têm o potencial de revelar inconsistências entre os diversos sistemas corporativos.

IV- Um data warehouse, em geral, deve ser projetado para fazer junções entre fatos e dimensões através de chaves naturais, evitando chaves substitutas (surrogate keys), pois estas apenas contribuiriam para aumentar o tamanho e a complexidade do esquema sem nenhum benefício para o usuário final.

Estão corretas APENAS as afirmações

Parte superior do formulário

a) I e II. b) I e III. c) II e III. d) II e IV. e) III e IV.

9. Analise as afirmações sobre as operações realizadas em cubos OLAP a seguir.
I - A operação pivot permite ao usuário alternar as linhas e colunas em que os valores visualizados serão recalculados.

II - A operação slice é caracterizada pela seleção de determinado membro de uma dimensão, a fim de analisar as demais informações do cubo sob tal perspectiva.

III - A operação dice corresponde à seleção específica de membros de duas ou mais dimensões.

Está correto o que se afirma em

Parte superior do formulário

a) III, apenas. b) I e II, apenas. c) I e III, apenas. d) II e III, apenas. e) I, II e III.

10. Um sistema de orçamento de um órgão público federal, que apresenta as informações baseadas na tecnologia OLAP, realiza agregações dos valores na etapa de carga dos dados para o cubo. Essa pré-agregação tem como principais características o(a):

Parte superior do formulário

a) aumento do uso da memória RAM para consulta aos dados do cubo, necessidade da existência de tabelas normalizadas no ambiente relacional de origem e a utilização somente de fatos numéricos.

b) aumento do tempo de processamento para formação do cubo, a rapidez na consulta de dados agregados e o aumento da utilização da área de armazenamento dedicada ao cubo.

c) necessidade da utilização de star schema para carga dos dados, a rapidez no tempo de processamento para construção do cubo e a maior fluidez nas operações de slice and dice.

d) rapidez na consulta a partir das operações de pivot, a eliminação da necessidade de cálculos no ambiente relacional e o aumento do tempo nas consultas com operações de drill up e drill down.

e) diminuição da utilização da área de armazenamento do data warehouse, a eliminação de cálculo de dados no ambiente OLAP e o aumento da acuidade de valores totais em níveis hierárquicos mais altos das dimensões do cubo.

Parte inferior do formulário

11. A volatilidade dos dados é característica intrínseca a data warehouses.

**PORQUE**

Sistemas ROLAP possuem um conjunto de interfaces e aplicações que dão ao Sistema Gerenciador de Banco de Dados Relacionais características multidimensionais.
Analisando as afirmações acima, conclui-se que

Parte superior do formulário

a) as duas afirmações são verdadeiras e a segunda justifica a primeira.

b) as duas afirmações são verdadeiras e a segunda não justifica a primeira.

c) a primeira afirmação é verdadeira e a segunda é falsa.

d) a primeira afirmação é falsa e a segunda é verdadeira.

e) as duas afirmações são falsas.

12. O texto a seguir se refere à modelagem de Data Warehouse.

Se na modelagem do Data Warehouse for adotada uma abordagem _____, cada elemento de dados (por exemplo, a venda de um item) será representado em uma relação, chamada tabela de fatos, enquanto que as informações que ajudam a interpretar os valores ao longo de cada dimensão são armazenadas em uma tabela de dimensões, uma para cada dimensão. Esse tipo de esquema de banco de dados é chamado um esquema estrela, em que a tabela de fatos é o centro da estrela e as tabelas de dimensões são os pontos. Quando a abordagem _____ é escolhida, um operador específico que faz a agregação prévia da tabela de fatos ao longo de todos os subconjuntos de dimensões é utilizado e pode aumentar consideravelmente a velocidade com que muitas consultas _____ podem ser respondidas.

Considerando a ordem das lacunas, qual sequência de termos completa corretamente o texto acima?

Parte superior do formulário

a) MOLTP, ROLTP, OLTP. b) ROLTP, MOLTP, OLTP. c) ROLAP, MOLAP, OLAP.

d) ROLAP, MOLAP, OLTP. e) MOLAP, ROLAP, OLAP.

# Unidade 4 : Mineração De Dados

- Conceitos de mineração e descoberta de conhecimento
- Técnicas de mineração de dados e algoritmo
    - ✓ Análise de associação
    - ✓ Classificação e previsão de dados
    - ✓ Segmentação e analise de cluster
- Aplicação de mineração de dados
- Ferramentas de mineração de dados

Sistemas e Linguagens para Mineração de Dados.

## Palavras Chave/ Termos

**Algoritmo:** Um conjunto de passos ordenados para solução de um problema, como uma fórmula matemática ou instruções num programa. É o embrião de qualquer programa para computador, sendo o responsável por realizar todas as transformações necessárias a fim de se atingir um determinado objetivo.

**Data Mining – Mineração de dados:** É o meio de descobrir tendências, padrões e relacionamentos entre os dados. Pode-se empregar técnicas de Estatística ou Inteligência Artificial para encontrar informações que podem não estar imediatamente visíveis, ou seja, contra-intuitivas. Pode descobrir associações (correlações entre eventos), seqüências (eventos que levam a outros), classificações (padrões para estabelecer perfis), clustering (encontrar e visualizar novos grupos) e previsões (descobrir padrões que levam a prever o futuro).

**Crosstab** ou **Pivot Table**: Objeto que possibilita análises avançadas de dados. É composto por três partes: linhas, colunas e dados. Oferece recursos como drag e drop (arrastar e soltar) e slice and dice (fatiar e combinar).

**índice de armazenamento em cluster:** Um índice cuja seqüência de valores-chave corresponde estritamente à seqüência de linhas armazenadas em uma tabela. O grau de correspondência é medido através de estatísticas que são utilizadas pelo otimizador.

**Árvore de decisão :**Uma forma de "estruturar e representar conhecimento". Especificamente no contexto de mineração de dados, a expressão se refere a um tipo de algoritmo de aprendizado de máquina

**Associação:** Um tipo de análise que pode ser feita utilizando mineração de dados. Através da exploração das transações já realizadas pelos clientes, descobrir produtos que normalmente são vendidos na mesma transação ou para o mesmo cliente em um determinado tempo.

**Modelo de mineração:** A saída de uma função de mineração de dados que descreve padrões e relacionamentos que são descobertos em dados históricos. Um modelo de mineração de dados pode ser aplicado em novos dados para prever os prováveis novos resultados.

**função de associações:** Um algoritmo de mineração que localiza associações, padrões ou regras escondidos em dados.

**Confiança:**Intervalo previsto de uma variável de saída, dado um conjunto de valores de variáveis de entrada.

## Data Mining

Segundo Eduardo Corrêa Gonçalves (data???) Definição simples para mineração de dados (data mining):

- Processo realizado através de estratégias automatizadas que tem por objetivo a descoberta de conhecimento valioso em grandes bases de dados.

Esquema conceitual: um "pequeno diamante de informação" é extraído a partir de uma verdadeira "montanha de dados"!

A mineração de dados baseia-se na utilização de algoritmos capazes de vasculhar grandes bases de dados de modo eficiente e revelar padrões interessantes, escondidos dentro da "montanha de dados".

A mineração de dados permite a descoberta de conhecimento, a partir dos dados armazenados utilizando os conceitos e as técnicas de forma a analizar as tendências e padroes associados ás informações bruto para facilitar os decisores na tomada de decisão.

Diferentes autores procuram definir mineração em função das áreas de negócios, mas tudo leva a um caminho comum, que é a procura de conhecimentos.

Por ser considerada multidisciplinar, as definições acerca do termo Mineração de Dados variam com o campo de atuação dos autores. Destacamos neste trabalho três áreas que são consideradas como de maior expressão dentro da Mineração de Dados: Estatística, Aprendizado de Máquina e Banco de Dados. Em Zhou (data??), é feita uma análise comparativa sobre as três perspectivas citadas.

- Em Hand et al (data???), a definição é dada de uma perspectiva estatística: "Mineração de Dados é a análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam tanto úteis quanto compreensíveis ao dono dos dados".
- Em Cabena et al.(data???) a definição é dada de uma perspectiva de banco de dados: "Mineração de Dados é um campo interdisciplinar que junta técnicas de máquinas de conhecimentos, reconhecimento de padrões, estatísticas, banco de dados e visualização, para conseguir extrair informações de grandes bases de dados".

Mineração de Dados é um passo no processo de Descoberta de Conhecimento que consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta que, sob certas limitações computacionais, produzem um conjunto de padrões de certos dados." Apesar das definições sobre a Mineração de Dados levar a crer que o processo de extração de conhecimento se dá de uma forma totalmente automática, sabe-se hoje que de fato isso não é verdade [39]. Apesar de encontrarmos diversas ferramentas que nos auxiliam na execução dos algoritmos de mineração, os resultados ainda precisam de uma análise humana. Porém, ainda assim, a mineração contribui de forma significativa no processo de descoberta de conhecimento, permitindo aos especialistas concentrarem esforços apenas em partes mais significativa dos dados.

O conhecimento descoberto através de processos de mineração de dados é considerado interessante quando apresenta certas propriedades:

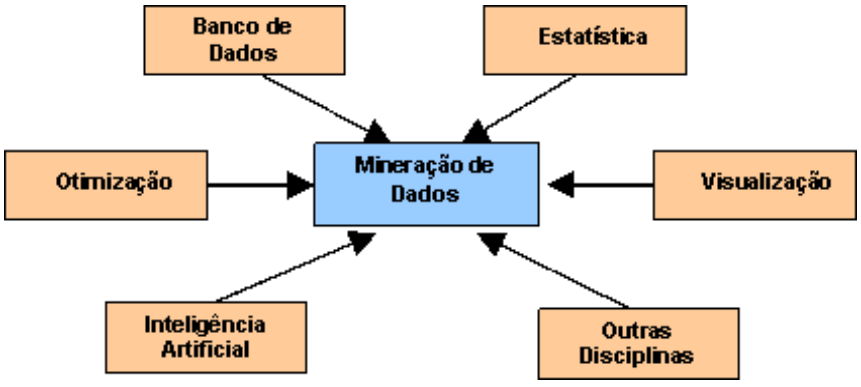



Figura 1:Mineração de Dados

Fonte: http://www.devmedia.com.br/artigo-sql-magazine-10-introducao-ao-data-mining/7820 acesso dezembro de 2015

## Mineração de Dados: Introdução e Aplicações

Os bancos relacionais, quando bem projetados, permitem a extração de diversas informações usando SQL. O mecanismo é simples: elabora-se um problema, é realizado um mapeamento para a linguagem de consulta, e esta consulta é submetida ao SGBD. Observe que esse processo resolve questões que necessariamente devem ser definidas, ou seja, as informações extraídas são respostas a uma consulta previamente estruturada. No entanto, dados armazenados podem esconder diversos tipos de padrões e comportamentos relevantes que a princípio não podem ser descobertos utilizando-se SQL. Além disso, por mais que o analista seja criativo, ele irá apenas conseguir elaborar diversas questões de forma que se tenham resultados práticos no final.

Atualmente, encontram-se comercialmente disponíveis diversas ferramentas de mineração de dados que auxiliam cientistas a classificar e segmentar dados, formular hipóteses, realizar diagnósticos. Ajudam analistas a entender e prever necessidades e interesses dos clientes, descobrirem perfis de comportamento, detecção de fraudes, aprovação de crédito e de apólice.

Assim, mineração de dados pode ser definida como o processo automatizado de descoberta de novas informações a partir de grandes massas de dados. A mineração de dados é uma área extensa, interdisciplinar e envolve o estudo de diversas técnicas (ver figura 13).

A mineração de dados não ocorre somente em bancos de dados relacionais. Hoje em dia pode-se trabalhar com diversas fontes tais como textos, arquivos de log, data warehouses, entre outras.

Data Mining Não são poucas as sínteses oferecidas por diversos autores para conceituar o significado e as características de Data Mining. Em tradução livre, Data Mining aparece como Miniração de Dados. Grosso modo define-se como uma ferramenta de inteligência capaz de extrair informações pertencentes em um DW ou DM. Um recurso que é eficaz para filtrar o conhecimento que se mostra relevante para a empresa de uma forma mais fácil, apresentando uma síntese mais clara para o utilizador. Encontra-se na literatura uma série de definições e conceitos sobre Data Mining, a seguir algumas delas: O Data Mining, ou mineração de dados, é um meio de extrair informações, anteriormente desconhecidas, da base de dados acessíveis nos DW. As ferramentas de DM usam algoritmos automatizados e sofisticados para descobrir padrões ocultos, correlações e relações entre os dados organizacionais. Essas ferramentas são usadas para projetar tendências e comportamentos futuros, permitindo que as empresas tomem decisões proativas e abalizadas. Data Mining, ou Mineração de Dados, pode ser entendido como o processo de extração de informações, sem conhecimento prévio, de um grande banco de dados e seu uso para tomada de decisões.

Segundo Silva, "Data Mining é uma técnica para determinar padrões de comportamento, em grandes bases de dados, auxiliando na tomada de decisão".

No cotidiano, percebe-se que a utilização do Data Mining pode ser útil na busca por dados que antes não eram facilmente encontradas. Por permitir essa descoberta, conseqüentemente, propicia a metamorfose destas informações em ações e resultados. O objetivo do data mining é descobrir, de forma automática ou semiautomática, o conhecimento que está —escondido‖ nas grandes quantidades de informações armazenadas nos bancos de dados da organização, permitindo agilidade na tomada de decisão. Uma organização que emprega o data mining é capaz de: Para Primak , o Data Mining está mais relacionado com processos de análise de inferência do que com os de análise dimensional de dados, representando assim uma forma de busca de informação baseada em algoritmos que objetivam o reconhecimento de padrões escondidos nos dados e não necessariamente revelados pelas outras abordagens analíticas, como o OLAP (cubo).

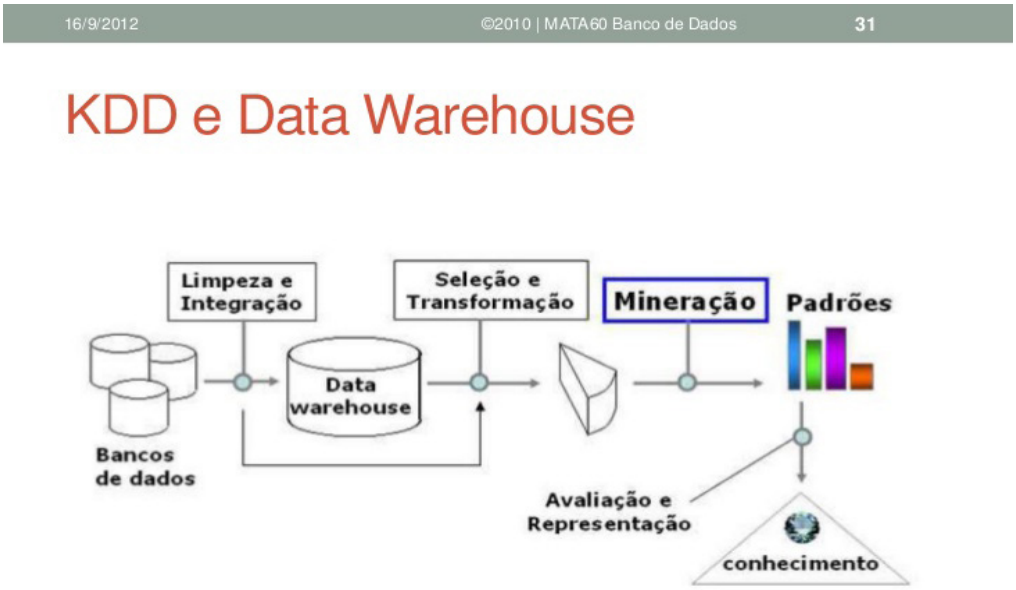



Figura 2:Processo KDD

Descoberta de Conhecimento em Banco de Dados. Na verdade, KDD é um processo mais amplo consistindo das seguintes etapas, como ilustrado na Figura 20:

1. **Limpeza dos dados:** etapa onde são eliminados ruídos e dados inconsistentes.
2. **Integração dos dados:** etapa onde diferentes fontes de dados podem ser combinadas produzindo um único repositório de dados.
3. **Seleção**: etapa onde são selecionados os atributos que interessam ao utilizador. Por exemplo, o utilizador pode decidir que informações como endereço e telefone não são de relevantes para decidir se um cliente é um bom comprador ou não.
4. Transformação dos dados: etapa onde os dados são transformados num formato apropriado para aplicação de algoritmos de mineração (por exemplo, através de operações de agregação).

5. Mineração: etapa essencial do processo consistindo na aplicação de técnicas inteligentes a fim de se extrair os padrões de interesse.
6. Avaliação ou Pós-processamento: etapa onde são identificados os padrões interessantes de acordo com algum critério do utilizador.
7. Visualização dos Resultados: etapa onde são utilizadas técnicas de representação de conhecimento a fim de apresentar ao utilizador o conhecimento minerado.

## Tarefas e Técnicas de Mineração de Dados

As tarefas e a técnicas de mineração de dados constituem um pilar fundamental na descoberta do conhecimento sem o  qual seria impossível inteirar  das tendências ocultas da informação.

É importante distinguir o que é uma tarefa e o que é uma técnica de mineração. A tarefa consiste na especificação do que estamos querendo buscar nos dados, que tipo de regularidades ou categoria de padrões temos interesse em encontrar, ou que tipo de padrões poderiam nos surpreender (por exemplo, um gasto exagerado de um cliente de cartão de crédito, fora dos padrões usuais de seus gastos). A técnica de mineração consiste na especificação de métodos que nos garantam como descobrir os padrões que nos interessam.

Os dois objetivos de mais alto nível que se tem interesse com a aplicação de técnicas de mineração de dados são a predição e a descrição [FAY 96, MEN 98]. Os padrões do tipo preditivo são os que procuram predizer valores futuros de um ou mais atributos variáveis do banco de dados com base em valores conhecidos de outros atributos. Os do tipo descritivo, por sua vez, procuram descrever padrões encontrados nos dados em um formato que seja interpretável pelo homem.

Principais tarefas de mineração:

- Associação,
- sequência,
- Classificação,
- Cluster,
- outlier.

Para que isso possa acontecer tem de ser utilizadas as tarefas da mineração (classificação, estimativa ou regressão, associação, clusterização) e as técnicas existentes (regras de associação, regras de classificação, árvores de decisão, agrupamento).

Como mencionado, várias são as tarefas que podem ser utilizadas na implementação destas estratégias, onde as principais são:

| Estratégia | Algoritmos |
|---|---|
| Classificação | árvores de decisão e redes neurais |
| Agregação | métodos estatísticos e redes neurais |
| Associação | métodos estatísticos e teoria de conjuntos |
| Regressão | métodos de regressão e redes neurais |
| Predição | métodos estatísticos e redes neurais |

**Com base na tabela pode se concluir** que redes neurais é a que apresenta maior abrangência, podendo ser aplicada em praticamente todas as estratégias.

## Análise de Regras de Associação.

O objetivo da técnica é avaliar que valores aparecem muito juntos nas mesmas transações ou eventos (por exemplo, carrinhos de compras), mas também pode ser utilizada para identificar relações entre atributos dentro de uma mesma entidade (ex.: clientes do sexo feminino costumam morar mais no bairro X). Como a própria palavra sugere associa a ocorrência de itens que aparecem associados

Esta regra tem como o inputUma regra de associação é um padrão da forma $X \rightarrow Y$ , onde X e Y são conjuntos de valores (artigos comprados por um cliente, sintomas apresentados por um paciente, etc). Uma regra de associação é uma expressão da forma $X \rightarrow Y$, onde X e Y são itemsets. Por exemplo, {Banana, leite} → {café} é uma regra de associação. A idéia por trás desta regra é que pessoas que compram Banan e leite têm a tendência de também comprar café, isto é, se alguém compra Banan e leite então também compra café. Repare que esta regra é diferente da regra {café} → {Banana,leite}. A toda regra de associação $A \rightarrow B$ associamos um grau de confiança, denotado por conf($X \rightarrow Y$).

Dado um conjunto de transações, 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 freqüência, são confiáveis e fazem predições interessantes. Regras que refletem casos que raramente ocorrem tendem a ser desprezadas. Por este motivo, deve-se estipular um valor mínimo para o suporte, geralmente referido como suporte_mínimo. O suporte de uma regra $X \rightarrow Y$, dado por suporte(XY), é a probabilidade de um transação satisfaça tanto X como Y. Caso o suporte de um conjunto de itens for maior ou igual ao mínimo estabelecido (suporte(X) ≥ suporte_mínimo), diz-se que este conjunto de itens é freqüente. Por outro lado, as regras também precisam refletir a realidade com um certo grau de certeza de estarem certas. Regras que raramente estão corretas, ou seja, que a implicação raramente se confirma, também não são regras úteis. Portanto, deve-se estipular um valor mínimo para a confiança da regra, geralmente referido como confiança_mínima. A confiança é dada por suporte(XY) / suporte(X), e representa a probabilidade de que uma transação satisfaça Y, dado que ela satisfaz X.

Este grau de confiança é simplesmente a porcentagem das transações que suportam Y dentre todas as transações que suportam X, isto é:

$$Conf(X \rightarrow Y) = \frac{\text{Numero de Transações que suportam ( X } \cup \text{ Y)}}{\text{Numero de tranasações que suportam X}}$$

As regras de associação representam padrões onde a ocorrência de eventos em um conjunto é alta.

Para melhorar o processamento dos algoritmos de associação, na fase de pré processamento, são associados números a cada artigo da loja como ilustrado na Tabela  para facilitar a representação dos mesmos e atender as especificações dos algoritmos.

Título e número desta tabela

| IDT (Identificador da Transação) | Itens comprados |
|---|---|
| 11 | {1, 3, 5} |
| 12 | {2, 1, 3, 7, 5} |
| 13 | {4, 9, 2, 1} |
| 14 | {5, 2, 1, 3, 9} |
| 15 | {1, 8, 6, 4, 3, 5} |
| 16 | {9, 2, 8} |

Na tabela acima as transacções são identificadas pelo código IDT e não pelo código do cliente. Assim, um mesmo cliente é contado várias vezes a cada vez que realiza uma compra na loja. É importante lembrar que o que interessa é a transação (a compra), e não o cliente. Na descoberta de associação, o cliente é irrelevante, pois o que interessa na identificação dos padrões são as ocorrências dos produtos comprados e não quem os comprou. É de extrema importância identificar os itens que normalmente são comprados em conjunto,isto é denominado de itemset cada conjunto de itens comprados pelo cliente em uma única transação. Um itemset com k elementos é chamado de k-itemset. Suponha que o gerente decida que um itemset que apareça em pelo menos 50% de todas as compras registadas seja considerado frequente. Por exemplo, se a tabela em cima for considerada um banco de dados, então o itemset {1, 3} é considerado frequente, pois aparece em mais de 60% das transações. Porém, se você for muito exigente e decidir que o mínimo para ser considerado frequente é aparecer em pelo menos 70% das transações, então o itemset {1, 3} não é considerado freqüente.

Conjunto de itens frequentes:

| Suporte | Conjunto de itemset |
|---|---|
| 66,6% | {1,3} |
| 33,3% | {2,3} |
| 16,6% | {1,2.7} |
| 5% | {2,9} |

Número e título da tabela???

Regras de associações encontradas

| Regra | Grau de Confiança |
|---|---|
| 1→3 | 100% |
| 2→3 | 100% |

"Formalmente, uma regra de associação é uma implicação da forma X → Y, onde X e Y são conjuntos de itens tais que X Y = . Convém destacar que a interseção vazia entre antecedente e conseqüente da regras assegura que não sejam extraídas regras óbvias que indiquem que um item está associado a ele próprio." (GOLDSCHMIDT, 2005, p. 61).

## Sequência

O algoritmo GSP é projetado para a tarefa de mineração de padrões sequencias ou sequências. Uma sequência ou padrão sequencial de tamanho k (ou k-sequência) é uma coleção ordenada de itemsets < I1, I2, ..., In >. Por exemplo, s = é um padrão sequencial. Uma k-sequência é uma sequência com k itens. Um item que aparece em itemsets distintos é contado uma vez para cada itemset onde ele aparece. Por exemplo, < {1, 2} >, < {1}, {2} >, < {1}, {1} > são 2-sequências. Denição 5.2.2 Sejam s e t duas sequências, s = < i1i2...ik > e t = < j1j2...jm >. Dizemos que s está contida em t se existe uma subsequência de itemsets em t, I1, ..., Ik tal que i1 ⊆ I1, ... , ik ⊆ Ik. Por exemplo, sejam t =< {1, 3, 4}, {2, 4, 5}, {1, 7, 8} > e s =< {3}, {1, 8} >. Então, é claro que s está contida em t, pois {3} está contido no primeiro itemset de t e {1,8} está contido no terceiro itemset de t. Por outro lado, a sequência s 0 =< {8}, {7} > não está contida em t. De agora em diante, vamos utilizar a seguinte nomenclatura para diferenciar as sequências: sequências que fazem parte do banco de dados de sequências (correspondem a alguma sequência de itemsets comprados por algum cliente) são chamadas de sequência do cliente. Sequências que são possíveis padrões que podem aparecer nos dados são chamadas de padrão sequencial

Definimos suporte de um padrão sequencial s com relação a um banco de dados de sequências de clientes D como sendo a porcentagem de sequências de clientes que suportam s, isto é: sup(s) = número de sequências de clientes que suportam s número total de sequências de clientes Um padrão sequencial s é dito frequente com relação a um banco de dados de sequências de clientes D e um nível mínimo de suporte α, se sup(s) ≥ α. Assim como acontecia com os itemsets, a propriedade de ser frequente também é antimonotônica, no que diz respeito a padrões sequenciais. Isto é, se s é frequente então todos os seus subpadrões (ou subsequências) são frequentes. O problema da mineração de sequências consiste em, dados um banco de dados de sequências D e um nível mínimo de suporte α, encontrar todos os padrões sequenciais com suporte maior ou igual a α com relação a D. Seguindo a mesma idéia dos algoritmos da família Apriori, o algoritmo GSP gera as k-sequências frequentes (sequência com k itens) na iteração k. Cada iteração é composta pelas fases de geração, de poda e de validação (cálculo do suporte).

Fórmula para cálculo de Suporte:

$$\text{sup}\,(x) = \frac{\text{numero de sequencias de clientes que suportam x}}{\text{Numero total sequencias de Clientes}}$$

## Técnicas para Classificação e Análise de Clusters

Trata-se das tarefas de classificação e análise de clusters. Para cada uma destas tarefas, veremos algumas técnicas comunmente utilizadas para realizá-las. Começaremos pela Classificação. Suponha que você é gerente de uma grande loja e disponha de um banco de dados de clientes, contendo informações tais como nome, idade, renda mensal, profissão e se comprou ou não produtos eletrônicos na loja. Você está querendo enviar um material de propaganda pelo correio a seus clientes, descrevendo novos produtos eletrônicos e preços promocionais de alguns destes produtos. Para não fazer despesas inúteis você gostaria de enviar este material publicitário apenas a clientes que sejam potenciais compradores de material eletrônico. Outro ponto importante : você gostaria de, a partir do banco de dados de clientes de que dispõe no momento, desenvolver um método que lhe permita saber que tipo de atributos de um cliente o tornam um potencial comprador de produtos eletrônicos e aplicar este método no futuro, para os novos clientes que entrarão no banco de dados. Isto é, a partir do banco de dados que você tem hoje, você quer descobrir regras que classificam os clientes em duas classes : os que compram produtos eletrónicos e os que não compram. Que tipos de atributos de clientes (idade, renda mensal, profissão) influenciam na colocação de um cliente numa ou noutra classe ? Uma vez tendo estas regras de classificação de clientes, você gostaria de utilizá-las no futuro para classificar novos clientes de sua loja. Por exemplo, regras que você poderia descobrir seriam:

Se idade está entre 30 e 40 e a renda mensal é `Alta' então ClasseProdEletr = `Sim'. Se idade está entre 60 e 70 então ClasseProdEletr = `Não'. Quando um novo cliente João, com idade de 25 anos e renda mensal `Alta' e que tenha comprado discos, é catalogado no banco de dados, o seu classificador lhe diz que este cliente é um potencial comprador de aparelhos eletrônicos. Este cliente é colocado na classe ClasseProdEletr = `Sim', mesmo que ele ainda não tenha comprado nenhum produto eletrônico. O que é um classificador ?. Classificação é um processo que é realizado em três etapas :

1. Etapa da criação do modelo de classificação. Este modelo é constituído de regras que permitem classificar as tuplas do banco de dados dentro de um número de classes pré-determinado. Este modelo é criado a partir de um banco de dados de treinamento. cujos elementos são chamados de amostras ou exemplos.

2. Etapa da verificação do modelo ou Etapa de Classificação : as regras são testadas sobre um outro banco de dados, completamente independente do banco de dados de treinamento, chamado de banco de dados de testes. A qualidade do modelo é medida em termos da percentagem de tuplas do banco de dados de testes que as regras do modelo conseguem classificar de forma satisfatória. É claro que se as regras forem testadas no próprio banco de dados de treinamento, elas terão alta probabilidade de estarem corretas, uma vez que este banco foi usado para extraí-las.

Por isso a necessidade de um banco de dados completamente novo. Diversas técnicas são empregadas na construção de classificadores. Neste curso vamos ver duas técnicas: árvores de decisão e redes neurais.

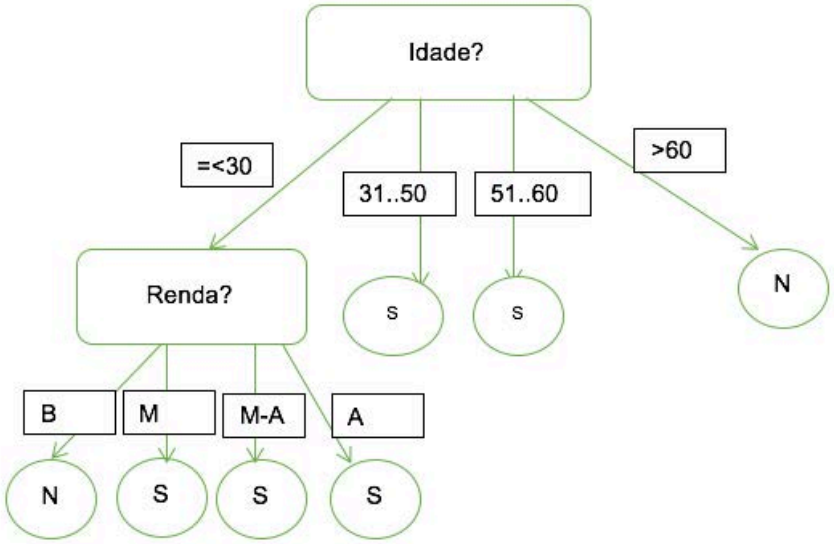
Árvore de Decisão

Uma árvore de decisão é uma estrutura de árvore onde:

- Cada nó interno é um atributo do banco de dados de amostras, diferente do atributo-classe,
- As folhas são valores do atributo-classe,
- Cada ramo ligando um nó-filho a um nó-pai é etiquetado com um valor do atributo contido no nó-pai. Existem tantos ramos quantos valores possíveis para este atributo.
- Um atributo que aparece num nó não pode aparecer em seus nós descendentes.

| Nome | Idade | Renda | Profissão | ClasseProdEletr |
|---|---|---|---|---|
| Daniel | =< 30 | Média | Estudante | Sim |
| João | 31..50 | Média-Alta | Professor | Sim |
| Carlos | 31..50 | Média-Alta | Engenheiro | Sim |
| Maria | 31..50 | Baixa | Vendedora | Não |
| Paulo | =< 30 | Baixa | Porteiro | Não |
| Otávio | > 60 | Média-Alta | Aposentado | Não |

A figura 21mostra a construção possível de uma árvore de decisão relativamente a base de dados acima.




**Equação 21:Arvore de Decisão**

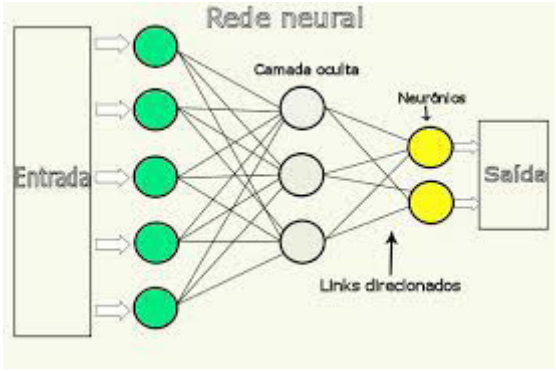
No banco de dados de amostras  na tabela acima onde foram utilizados os valores do atributo foram categorizados, numa lista de atributos candidatos de acordo com critério.

Output: Método para construção da árvore de decisão

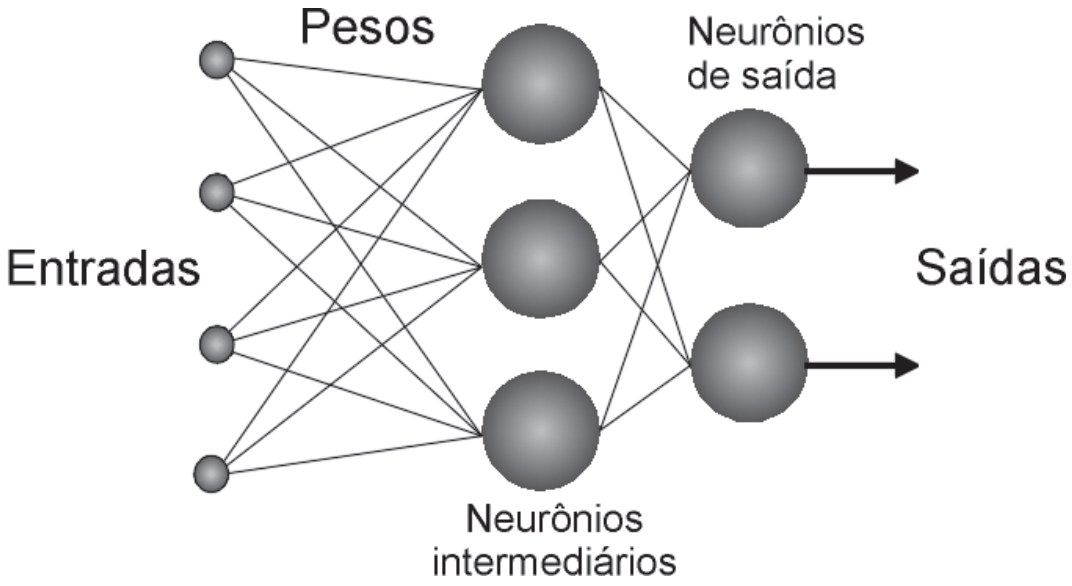
1. Crie um nó N (Idade?); Associe este nó o banco de dados.
2. Se todas as tuplas da base de dados pertencem à mesma classe C então transforme o nó N num Ramo da árvore (n=<30)  que poderá ter outros ramos ou folhas etiquetada por SlN.
3. Caso contrário: Se Cand-List = ∅ (>=60) então transforme N numa folha etiquetada com o valor do atributo-Classe que mais ocorre na base dados.
4. Caso contrário: calcule (Cand-List). Esta função retorna o atributo com o maior ganho de informação.

## Rede neurais

Uma rede neural. O algoritmo de classificação terá como entrada um banco de dados de treinamento e retornará como output uma rede neural. Esta rede também poderá ser transformada num conjunto de regras de classificação, como foi feito com as árvores de decisão. A única diferença é que esta transformação não é tão evidente como no caso das árvores de decisão. As redes neurais foram originalmente projetadas por psicólogos e neurobiologistas que procuravam desenvolver um conceito de neurônio artificial análogo ao neurônio natural. Intuitivamente, uma rede neural é um conjunto de unidades do tipo:



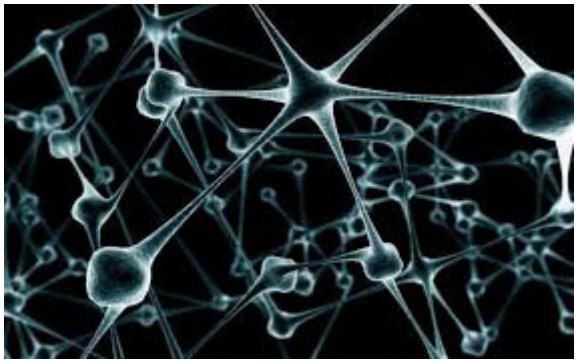

**Equação 22:Rede Neural**




**Equação 23:Rede Neural com peso**

Tais unidades são conectadas umas às outras e cada conexão tem um peso associado. Cada unidade representa um neurônio. Os pesos associados a cada conexão entre os diversos neurônios é um número entre -1 e 1 e mede de certa forma qual a intensidade da conexão entre os dois neurônios. O processo de aprendizado de um certo conceito pela rede neural corresponde à associação de pesos adequados às diferentes conexões entre os neurônios. Por esta razão, o aprendizado utilizando Redes Neurais também é chamado de Aprendizado Conexionista.

Se formos comparar diagrama de rede neural e a imagem pode se perfeitamente entender que o conceito neurónio vem exactamente da biologia em que vários neurónios estão interligados.



**Equação 24:Rede de neuronio**

A rede é composta de diversas camadas (verticais):

1. Camada de Input: consiste dos nós da primeira coluna na Figura 23. Estes nós correspondem aos atributos (distintos do atributo classe) do banco de dados .
2. Camada de Output: consiste dos nós da última coluna na Figura 23. Estes nós são em número igual ao número de classes. Eles correspondem, de fato, aos possíveis valores do Atributo-Classe.
3. Camadas escondidos ou intermediários: consiste dos nós das colunas intermediárias (indo da segunda até a penúltima). Numa rede neural existe pelo menos uma camada intermediária. Além disto, a seguinte propriedade se verifica: Cada nó de uma camada i deve estar conectado a todo nó da camada seguinte i + 1.

Uma rede neural é dita de n camadas se ela possui n − 1 camadas intermediárias. Na verdade, conta-se apenas as camadas intermediárias e a de output. Assim, se a rede tem uma camada intermediária, ela será chamada de rede de duas camadas, pois possui uma camada intermediário e uma de output. A camada de input não é contado, embora sempre exista.

## Clusterização ou Agrupamento

A técnica de Clustering recebe um grupo de elementos e daí identifica as classes. Ou seja, diferentemente da técnica de classificação, as classes não existem ainda ou não são conhecidas. O princípio básico da técnica é colocar no mesmo grupo os elementos mais similares e em grupos diferentes os elementos pouco similares. Este agrupamento é feito por algoritmos automáticos como o k Single-link, Strings e Cliques.

Agrupamento ou Clustering É um padrão que visa a agrupar objetos com base em alguma similaridade entre eles, com o objetivo de identificar aglomerações que descrevam dados. Pode-se citar entre as aplicações práticas deste método a descoberta de subgrupos de clientes que pode ajudar uma empresa a focar melhor seus produtos

## Avaliação da Unidade

Verifique a sua compreensão!

Uma aplicação específica está "varrendo" diversas bases de dados de uma organização: bancos de dados transacionais, planilhas eletrônicas, arquivos de texto, documentos eletrônicos, todos os tipos de "objetos" que possam conter informações estão sendo analisados em busca de padrões, tendências, transformações etc, que possam apoiar a Diretoria no processo de tomada de decisões.

O texto acima descreve:

a) Driew Down.

b) Pivot.

c) Data mining.

d) Datawarehouse.

e) ETL.

ETL e Pivot são conceitos associados a:

a) processo unificado de software.

b) desempenho de software.

c) notação de modelagem de processos.

d) inteligância de negócios

e) linguagem de execução de processos.

Data Mining é parte de um processo maior denominado:

a) Data Mart.

b) Database Marketing.

c) Knowledge Discovery in Database.

d) Business Intelligence.

e) Data Warehouse.

As principais técnicas utilizadas na mineração são:

a) a classificação,

b) a agregação,

c) a associação ou k-média,

d) a regressão e a predição.

Rede neuronais

As conexões entre os neurónios representam:

a) Pesos que armazenam o conhecimento

b) Representado no modelo e ponderam as entradas recebidas por cada neurônio da rede

c) A saída de um neurónio pode representar entrada para vários outros

d) Representa a multiplicação do pesos do neurónios

Data miming está a associado a resolução principalmente de dois tipos de problemas:

a) Predição

b) Descoberta de novo Conhecimento

c) Descoberta de conhecimento

d) Análise do mercado

Com base na tabela seguinte calcule grau de confiança e suporte para os itemset

| **IDT (Identificador da Transação)** | **Itens comprados** |
|---|---|
| 1001 | {1, 2, 5} |
| 1300 | {2, 1, 8, 7, 5} |
| 13001 | {4, 3, 2, 1} |
| 1411 | {5, 2, 1, 3, 4,9} |
| 1512 | {1, 6, 4, 3, 5} |
| 1602 | {9, 2,3, 8} |

## Leituras e Outros Recursos

As leituras e outros recursos desta unidade encontram-se na lista de "Leituras e Outros Recursos do curso".

**Bibliografia**


1. AMORIM, Sergio R. Leusin e CHERIAF, Malik, SISTEMA DE INDEXAÇÃO E RECUPERAÇÃO DE INFORMAÇÃO EM CONSTRUÇÃO BASEADO EM ONTOLOGIA, UFF - Universidade Federal Fluminense,
2. CASSIOPEIA: UM MODELO DE AGRUPAMENTO DE TEXTOS BASEADO EM SUMARIZAÇÃO. http://nlx.di.fc.ul.pt/~guelpeli/Arquivos/Tese.pdf acessado Novembro 2014
3. Christopher D. Mannin et al., Introduction to Information Retrieval ,Online edition (c),2009 Cambridge UP Online edition (c) Cambridge University Press Cambridge, England ,2009
4. Implicações da categorização e indexação na recuperação da informação tecnológica contida em documentos de patentes. http://nlx.di.fc.ul.pt/~guelpeli/Arquivos/Tese.pdf acessado Novembro 2014
5. De Medeiros Ericles Alves, TÉCNICA DE APRENDIZAGEM DE MÁQUINA PARA CATEGORIZAÇÃO DE TEXTOS. Trabalho de Conclusão de Curso- Engenharia da Computação Recife na Universidade politécnica Pernambuco, junho de 2004
6. Estratégia de busca na recuperação da informação: revisão da literatura, Ci. Inf., Brasília, v. 31, n. 2, p. 60-71, maio/ago. 2002
7. ERICK BONFIM MARCELLO , RECUPERAÇÃO DE DOCUMENTOS TEXTO USANDO UM MODELO PROBABILÍSTICO ESTENDIDO, tese de mestrado , SP 2006
8. Goodoy Viera-Angel Freddy e Dos Santos Garrdo Isora: Folksonomia como uma estratégia para Recuperação Colaborativa da Informação. DataGramaZero - Revista de Ciência da Informação - v.12 n.2 abr11 ARTIGO 02
9. Hagar Espanha Gomes e Maria Luiza de Almeida Campos,Tesauro e normalização terminológica: o termo como base para intercâmbio de informações ,DataGramaZero - Revista de Ciência da Informação - v.5 n.6 dez/04
10. Marcondes Carlos H. et al. Saberes e BIBLIOTECAS DIGITAIS SABERES E Práticas. Salvador/BrasíliaUFBA/IBICT 2005 http://livroaberto.ibict.br/bitstream/1/1013/1/Bibliotecas%20Digitais.pdf
11. R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. Addison-WesleyLongman,

12. Setzer Valdemar. Dado, Informação,Conhecimento e Competencia. DataGramaZero - Revista de Ciência da Informação - n. zero dez/99 .

13. http://exame.abril.com.br/tecnologia/glossario/ - ACESSADO Novembro 2014

14. http://marco.uminho.pt/~macedo/thesis/glossario.html ACESSADO NOV 2014

15. BARQUIM, 1997; CHAUDHURI, 1997; COREY, 2001, Uma Arquitetura de Gestão de Dados em Ambiente Data Warehouse

16. Raghu Ramakrishnan,Johannes Gehrke : Sistemas de gerenciamento de banco de dados - 3.ed.

17. Arquitectura de um SGDB http://www.devmedia.com.br/arquitetura-de-um-sgbd/25007 acessado Novembro 2014

18. Base de dados, http://blog.gaudencio.net.br/2012/08/banco-de-dados-conceituando-banco-de_5437.html acessado Novembro 2014

19. Ribeiro Rocha,Jaqueline et al.ESTUDO DE METODOLOGIAS PARA A CONSTRUÇÃO DE VOCABULÁRIOS CONTROLADOS NO ÂMBITO DA CIÊNCIA DA INFORMAÇÃO, Universidade Estadual de Londrina (UEL)

20. Robson Luís Gomes dos Santos, Usabilidade de interfaces para sistemas de recuperação de informação na web Estudo de caso de bibliotecas on-line de universidades federais brasileiras

21. OSCAR BUGMANN SANDRO, SISTEMA TUTOR PARA ENSINO DE SQL, Trabalho de Conclusão de Curso II do curso de Ciência da Computação — Bacharelado. BLUMENAU,2012

22. LEONARDO DAITX DE BITENCOURT, UM SISTEMA VOLTADO À INDEXAÇÃO E RECUPERAÇÃO DE INFORMAÇÃO INTEGRADO À ONTOLOGIA Grau de Bacharel em Tecnologias da Informação Araranguá 2013

23. Sistemas de Recuperação da Informação (Slides Adaptados de Cristopher D. Manning)

24. Recuperação de Informação em Bases de Texto, http://www.di.uevora.pt/~pq/ribt_aula6.pdf acessado em Novembro 2014

25. Prof. Ulrich, Tópicos Especiais em Ciência da Computação:Sistemas de Recuperação da Informação, http://www.dsc.ufcg.edu.br/~ulrich/disciplinas/SRI.html Acessado Novembro de 2014

26. OLINDA NOGUEIRA PAES CARDOSO, Recuperação de Informação

27. Zuchini Márcio Henrique Modelos Computacionais Adaptativos para Recuperação de Informação Adaptative Computing Models for Information Retrieval , USF, ESAMC

28. Prof. Dr. Leandro Balby Marinho Vocabulário de Termos e Listas de Postings http//www.dsc.ufcg.edu.br/~lbmarinho acessado Novembro 2014

29. D. Antonio Gabriel López Herrera, Modelos de Sistemas de Recuperacion de Informaci´on Documental Basados en Informaci´on Ling¨ u´ıstica Difusa Memoria de Tesis presentada para optar al grado de Doctor en Informática Granada 2006

30. de Araújo Júnior Rogério Henrique e Kira Tarapanoff Precisão no processo de busca e recuperação da informação: uso da mineração de textos, Ci. Inf., Brasília, v. 35, n. 3, p. 236-247, set./dez. 2006

31. Etapas do Processo de Mineração de Textos – uma abordagem aplicada a textos em Português do Brasil, 2006

32. E. A. M. Morais A. P. L. Ambrósio Technical Report - INF_005/07 - Relatório Técnico December - 2007 - Dezembro

33. JEROCIR BOTELHO MARQUES DE JESUS TESAURO: UM INSTRUMENTO DE REPRESENTAÇÃO DO CONHECIMENTO EM SISTEMAS DE RECUPERAÇÃO

34. Cássio de Albuquerque Melo, Cássio Centro de Informática Mecanismos de Ranqueamento na Web aplicados na Busca e Recuperação de Componentes de Software, Universidade Federal de Pernambuco,2006

35. Helder Joaquim Carvalheira Gomes ,Text Mining: Análise de Sentimentos na classificação de notícias dissertação de mestrado, Trabalho para obtenção do grau de Mestre em Estatística e Gestão de Informação,Universidade Nova de Lisboa 2012

36. PROBABILÍSTICOS EM RECUPERAÇÃO DE INFORMAÇÃO EM BASES TEXTUAIS, Dissertação submetida à Universidade Federal de Santa Catarina para a obtenção do grau de Mestre em Ciências da Computação, Florianópolis, outubro de 2001

37. C. J. van RIJSBERGEN,INFORMATION RETRIEVAL,Information Retrieval Group, University of Glasgow http://www.dcs.gla.ac.uk/Keith/Preface.htm acessado Novermbro 2014

38. MARCUS VINICIUS CARVALHO GUELPELI, CASSIOPEIA: UM MODELO DE AGRUPAMENTO DE TEXTOS BASEADO EM SUMARIZAÇÃO, Tese de Doutorado apresentada para obtenção do Grau de Doutor, Niterói 2012

39. LEANDRO KRUG WIVES TECNOLOGIAS DE DESCOBERTA DE CONHECIMENTO EM TEXTOS APLICADAS À INTELIGÊNCIA COMPETITIVA, PORTO ALEGRE, JANEIRO DE 2002.

40. http://www.leandro.wives.nom.br/pt-br/publicacoes/eq.pdf Acessado Novembro 2014

41. Uma Arquitetura de Gestão de Dados em Ambiente Data Warehouse Alcione Benacchio (UFPR) E¬mail: alcione@inf.ufpr.br Maria Salete Marcon Gomes Vaz (UEPG, UFPR) E¬mail: salete@uepg.br

42. http://hdl.handle.net/10183/49413 acessado em Dezembro de 2015

43. Amo Sandra Técnicas de Mineração de Dados Universidade Federal de Uberlândia Faculdade de Computação deamo@ufu.br - http://www.deamo.prof.ufu.br/arquivos/JAI-cap5.pdf ACESSADO DEZEMBRO 2015

44. Alcione Benacchio and Maria Salete Marcon Gomes Vaz Uma Arquitetura de Gestão de Dados em Ambiente Data Warehouse (UFPR) E-mail: alcione@inf.ufpr.br (UEPG, UFPR) E¬mail: salete@uepg.br

45. Caldeira Carlos Pampulim, Data Warehousing,2º Edição Lisboa ,2012

46. W.H. Inmon What is a Data Warehouse?, Prisma, Volume 1, Numero 1, 1995).

47. https://www.oficinadanet.com.br/artigo/business_intelligence/importancia-de-um-sistema-de-apoio-a-decisao acessado Janeiro 2016

48. https://www.oficinadanet.com.br/artigo/business_intelligence/importancia-de-um-sistema-de-apoio 2015-a-decisao acessado em dezembro

49. Sistema de Apoio a Decisão http://www.devmedia.com.br/sistema-de-apoio-a-decisao/6201#ixzz3zzYQSyF0 Acessado Dezembro 2015

50. Sistema de Apoio a Decisão http://www.devmedia.com.br/sistema-de-apoio-a-decisao/6201#ixzz3zzXfx7TB Acessado Dezembro 2015

51. http://blogs.ambientelivre.com.br/paulo/modelagem-dimensional-tipos-de-dimensoes/ (Com Ajuda do PAPA da modelagem Dimensional – Ralph Kimbal)

52. MACHADO, F.N.R. Projeto de Data Warehouse, São Paulo: Érica, 2004.)

53. Machado Felipe Nery Rodrigues Tecnologia e Projeto de Data Warehouse,São Paulo, 2007.

54. Artigo SQL Magazine 10 - Introdução ao Data Mining http://www.devmedia.com.br/artigo-sql-magazine-10-introducao-ao-data-mining/7820#ixzz40Bssc03y Acessado Dezembro 2015

55. http://www.devmedia.com.br/artigo-sql-magazine-10-introducao-ao-data-mining/7820 Acessado Dezembro 2015

56. Artigo SQL Magazine 10 - Introdução ao Data Mining http://www.devmedia.com.br/artigo-sql-magazine-10-introducao-ao-data-mining/7820#ixzz40BsABd6i Acessado em Dezembro 2015

57. http://www.nwk.edu.br/intro/wp-content/uploads/2014/05/revista_cientifica_bsi_2011.pdf acessado em Dezembro de 2015

58. Mineração de Dados: Tarefas e Técnicas http://www.devmedia.com.br/mineracao-de-dados-tarefas-e-tecnicas/30919#ixzz40G30IY00 Acessado Dezembro 2015

59. http://corporate.canaltech.com.br/materia/business-intelligence/a-granularidade-de-dados-no-data-warehouse-26310/

60. Glossário, Disponível em: <http://www.digitro.com/pt/index.php/sala-imprensa/glossario>. Acesso em: 05 fevereiro 2011.

61. http://corporate.canaltech.com.br/materia/business-intelligence/Tipos-de-metricas-existentes-no-Data-Warehouse/ acessado Fev 2016

62. http://www.devmedia.com.br/qualidade-na-modelagem-dos-dados-de-um-data-warehouse/6978 acessado Fevereiro 2016

63. Qualidade na Modelagem dos Dados de um Data Warehouse http://www.devmedia.com.br/qualidade-na-modelagem-dos-dados-de-um-data-warehouse/6978#ixzz3zr5yIUOv Acessado Fevereiro 2016

64. Artigo SISTEMA DE APOIO À DECISÃO (Akyria Bolonine et al) http://www.devmedia.com.br/sistema-de-apoio-a-decisao/6201#ixzz41FSeCTnE acessado Dezembro 2015

65. Data Warehouse Aplicado ao Negócio, Erick Skorupa Parolin. S.Paulo 2011

66. Tipos de métricas existentes no Data Warehouse

67. http://corporate.canaltech.com.br/materia/business-intelligence/Tipos-de-metricas-existentes-no-Data-Warehouse/ Acessado Fevereiro 2016

68. Tabelas Fato: http://litolima.com/2010/01/07/modelagem-dimensional-conceitos-basicos/ acessado Janeiro 2016

69. http://www.verga.com.br/downloads/glossario-de-bi/ acessado Dezembro 2015

70. http://www.coladaweb.com/informatica/olap acessado Dezembro 2015

71. http://www.itnerante.com.br/profiles/blogs/artigo-suporte-a-decis-o-02-sobre-as-opera-es-de-olap acessado Dezembro 2015

72. https://www.inf.ufsc.br/~frank/bd_univag/DataWarehouse-Peq.pdf Acessado Dezembro 2015

**Sede da Universidade Virtual africana**

The African Virtual University
Headquarters

Cape Office Park

Ring Road Kilimani

PO Box 25405-00603

Nairobi, Kenya

Tel: +254 20 25283333

contact@avu.org

oer@avu.org

**Escritório Regional da Universidade Virtual Africana em Dakar**

Université Virtuelle Africaine

Bureau Régional de l'Afrique de l'Ouest

Sicap Liberté VI Extension

Villa No.8 VDN

B.P. 50609 Dakar, Sénégal

Tel: +221 338670324

bureauregional@avu.org