UnB Departamento de Ciência da Computação

Business Intelligence

Data science moves from the specialist to the everyman. Familiarity with data analysis becomes part of the skill set of ordinary business users, not experts with “analyst” in their titles. Organizations that use data to make decisions are more successful, and those that don’t use data begin to fall behind.

The Economist

Edison Ishikawa, D. Sc.

Introdução

- Objetivo
  - Fornecer uma visão geral de Data Warehouse, OLAP e Data Mining

Sumário

- Introdução
- Desenvolvimento
  - Data Warehouse
  - Online Analytical Processing
  - Data Mining
- Considerações finais
- Referências

Introdução

- Business Intelligence – termo da moda
- Ferramentas de BD encontradas sob o título BI
  - Data warehouse
  - On-line analytical processing (OLAP)
  - Data mining
- Funcionalidades dessas ferramentas são complementares e relacionadas

Data warehouse

- Leva em consideração o armazenamento, a manutenção e a recuperação eficiente de dados históricos

OLAP

- É um serviço que fornece respostas rápidas a consultas ocasionais realizadas no data warehouse

Data mining

- Algoritmos que encontram padrões dos dados e informam modelos aos usuários

Business Inteligence

- Todas as três ferramentas estão relacionadas ao modo como os dados em um data warehouse são organizados logicamente, e o desempenho é altamente sensível às técnicas de projeto de BD utilizadas
- Objetivo
  - Fornecer informações úteis de apoio à decisão

Data warehouse

- Grande repositório de dados históricos que podem ser integrados para apoiar decisões
- Surgiu da tecnologia de Sistemas de Apoio a Decisão (DSS) e dos Sistemas de Apoio Executivo (ESS)
- Uso é nitidamente diferente dos sistemas operacionais (OLTP)
  - Contém dados exigidos nas operações do dia-a-dia de uma empresa
  - Dados costumam mudar rápido e constantemente
  - Tamanho das tabelas são mantidos relativamente pequenos , eliminando-se dados antigos de tempos em tempos

Data warehouse

- Diferenças com OLTP
  - Recebe periodicamente dados históricos em lotes e cresce com o tempo
  - Tamanho pode chegar de centenas de gigabytes a alguns terabytes
  - Faz consultas a grandes quantidades de dados e tem que retornar resultados rapidamente

Data warehouse x OLTP

- Os aspectos contrastantes entre os data warehouse e os sistemas operacionais resultam em uma abordagem de projeto diferenciada para o data warehouse

Data warehouse x OLTP

OLTP

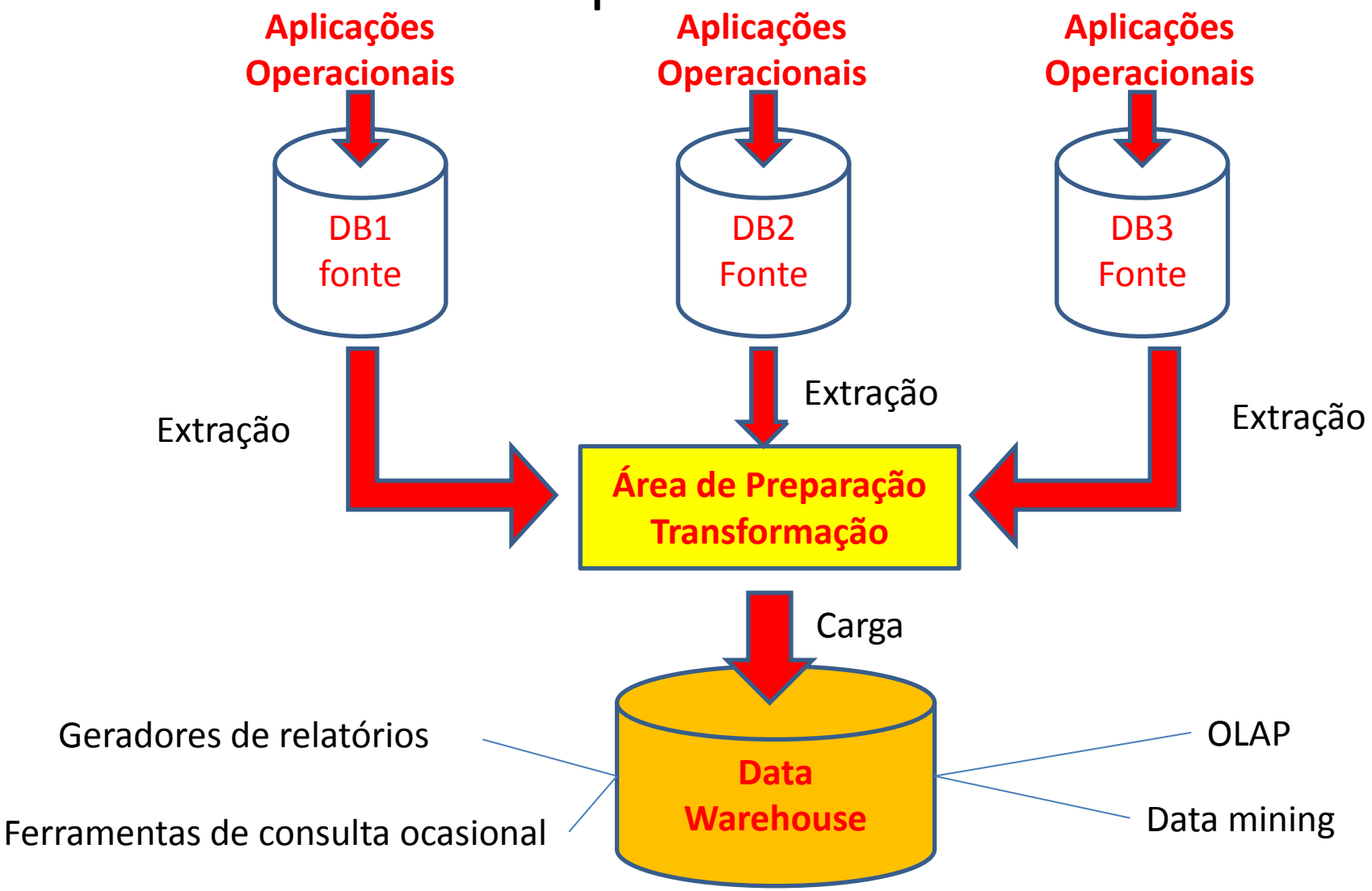
- Orientado a transação
- Milhares de usuários
- Geralmente pequeno (MB ate vários GB)
- Dados atuais
- Dados normalizados (muitas tabelas, poucas colunas por tabelas)
- Atualizações contínuas
- Consultas de simples a complexas

Data warehouse

- Orientado ao processo de negócios
- Poucos usuários (normalmente abaixo de 100)
- Grandes (de milhares de GB a vários TB)
- Dados históricos
- Dados não normalizados (poucas tabelas, muitas colunas por tabelas)
- Atualizações em lote
- Normalmente, consultas muito complexas

Data warehouse

arquitetura básica



Data warehouse

Características

1. Organizados de acordo com as áreas de interesse (semelhante a áreas funcionais – vendas, RH, etc..)
2. Dados são considerados não voláteis e são carregados em massa
3. Dados tendem a existir em vários níveis de granularidade
   - varia de acordo com as dimensões – dia, semana, mês, ano
   - Varia de acordo com o tempo – como dados são de natureza histórica, por exemplo, a moeda pode variar (cruzeiro, cruzado, real)

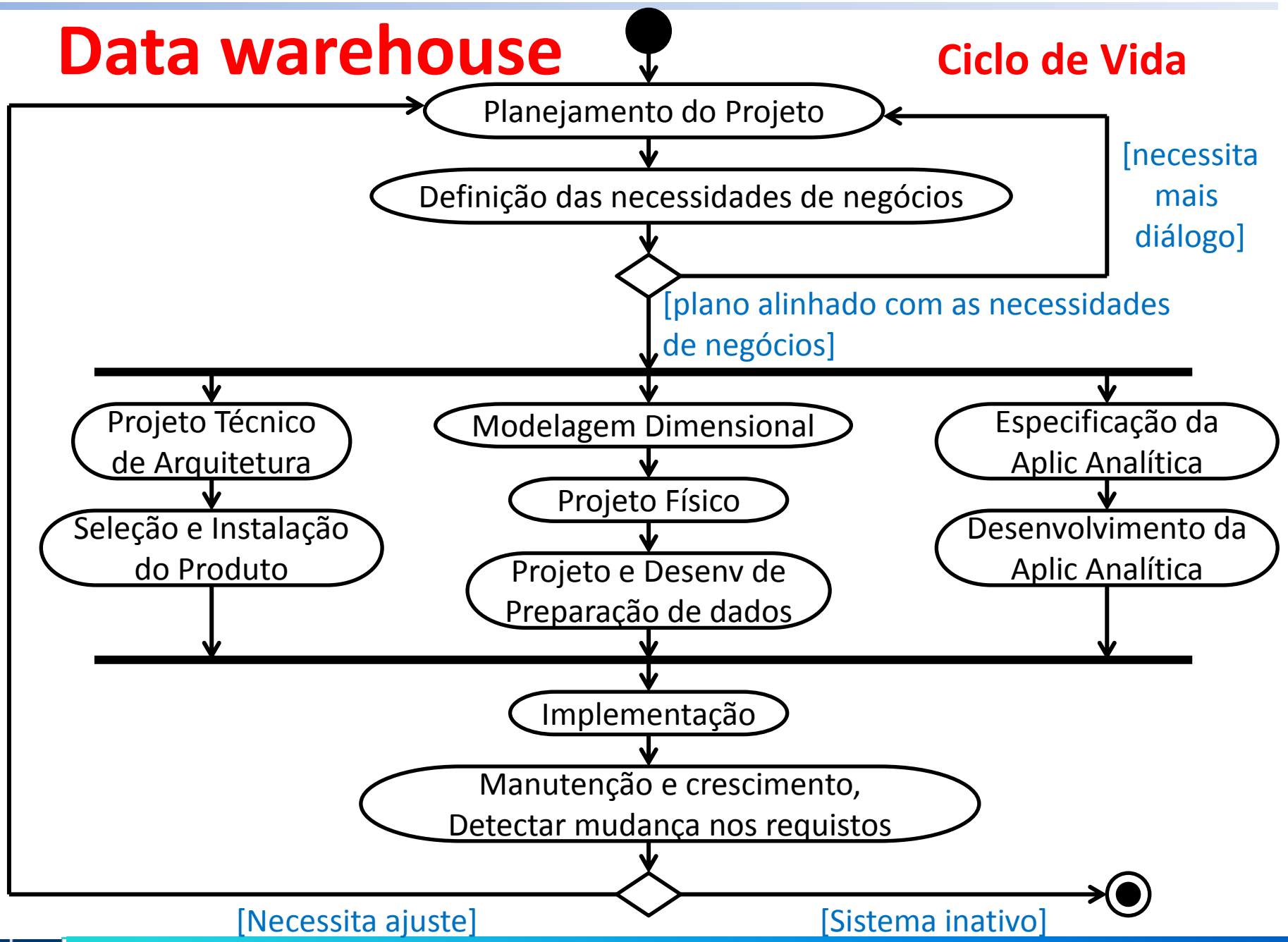
Data warehouse

Características

4. Deve ser flexível para atender rapidamente a necessidade por constantes mudanças
   - Definições de dados (esquemas) precisam ser suficientemente amplas para antecipar acréscimos de novos tipos de dados
5. Deve ter a capacidade de reescrever a história
   - Permitir análises hipotéticas (“o que acontece se”)
   - Deve permitir que se altere temporariamente os dados históricos com o objetivo de realizar análises hipotéticas

Data warehouse

Ciclo de Vida



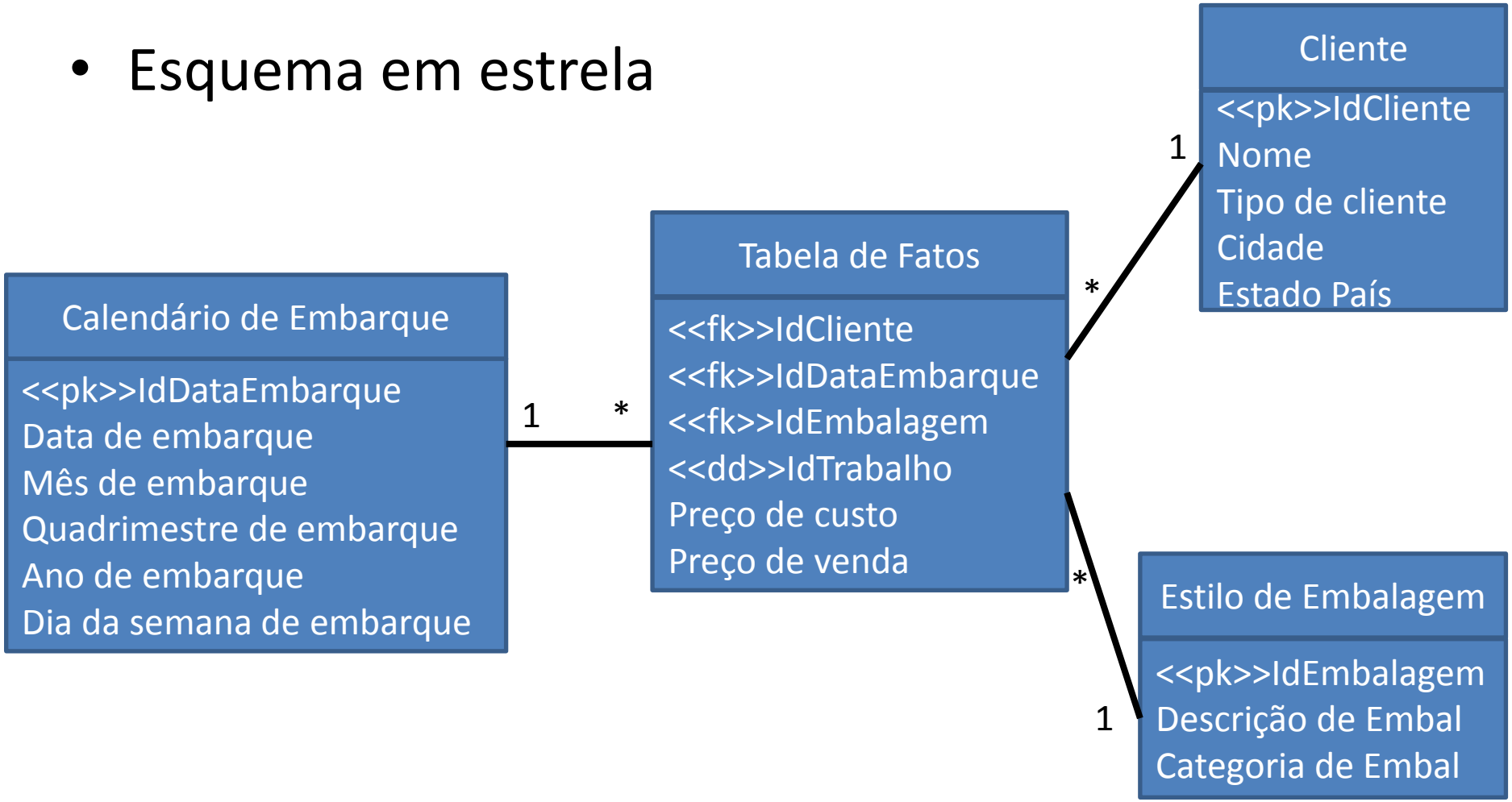
Data warehouse
Projeto Lógico

- Diferente da abordagem de normalização
  - Utilizada em sistemas operacionais - OLTP
- Definido pela abordagem de modelagem de dados **dimensional**
  - Esquema em estrela
  - Esquema em floco de neve

Data warehouse

Modelagem de Dados Dimensional

- Esquema em estrela



Data warehouse

Modelagem de Dados Dimensional

- Esquema em estrela
  - Organizado em uma grande tabela de fatos central
    - Possui dois tipos de atributos
      - de dimensão – IdCliente, IdDataEmbarque, IdEmbalagem, *IdTrabalho (dimensão degenerada - <<dd>>)*
        - em geral possui relacionamentos de chave estrangeira <<fk>> / chave primária <<pk>> com as tabelas de dimensão
      - de medidas – Preço de custo e Preço de venda
        - Valores a serem agregados quando as consultas agrupam linhas
  - Muitas tabelas de dimensão menores
    - Cliente, Calendário de Embarque e Estilo de Embalagem

Data warehouse

Modelagem de Dados Dimensional

- Esquema em estrela
  - Consultas normalmente utilizam atributos nas tabelas de dimensão para selecionar as linhas pertinentes das tabelas de fatos
    - Exemplo: consultar o preço de custo e de venda de todas as tarefas em que mês de embarque seja janeiro de 2012
  - Os atributos da tabela de dimensão também são usados para agrupar as linhas de maneiras úteis para se explorar informações de resumo
    - Exemplo: ver o custo total e o preço de venda de cada Ano de Embarque de 2012

Data warehouse

Modelagem de Dados Dimensional

- Esquema em estrela
  - Tabelas de dimensão permitem diferentes *níveis* de detalhe para o usuário examinar
    - Exemplo: Esquema do slide 18 permite que as linhas da tabela de fatos sejam agrupadas por:
      - Data de embarque
      - Mês
      - Quadrimestre
      - Ano
    - Esse níveis formam uma hierarquia
    - Existe uma segunda hierarquia na dimensão Calendário de Embarque
      - Dias da semana

Data warehouse
Modelagem de Dados Dimensional

- Esquema em estrela
  - Tabelas de dimensão permitem diferentes *níveis* de detalhe para o usuário examinar (Hierarquia)
  - Usuário pode subir ou descer em uma hierarquia
    - Drill-down – descer para examinar dados mais detalhados
    - Roll-up – subir em uma hierarquia para resumir detalhes

Data warehouse

Modelagem de Dados Dimensional

- Esquema em estrela
  - Os atributos de dimensão da tabela de fatos é uma candidata a chave
  - O nível de detalhe definido pelo atributo de dimensão é a *granularidade* da tabela de fatos
  - A granularidade deve ser o nível mais detalhado disponível que o usuário desejaria examinar
    - Às vezes isto significa que uma dimensão degenerada (IdTrabalho) precisa ser criada
      - Finalidade de IdTrabalho é distinguir as linhas no nível correto de granularidade
      - Sem IdTrabalho a tabela de fatos agruparia tarefas semelhantes, não permitindo ao usuário examinar os preços de custo e de venda de tarefas individuais

Data warehouse
Modelagem de Dados Dimensional

- Esquema em estrela (resumo)
  - Facilita a resposta rápida a consultas envolvendo um grande conjunto de dados históricos
    - Por isso não usa tabelas normalizadas
  - Principais dados detalhados são centralizados na tabela de fatos
  - Informações dimensionais e hierarquias mantidas nas tabelas de dimensão

Data warehouse

Modelagem de Dados Dimensional

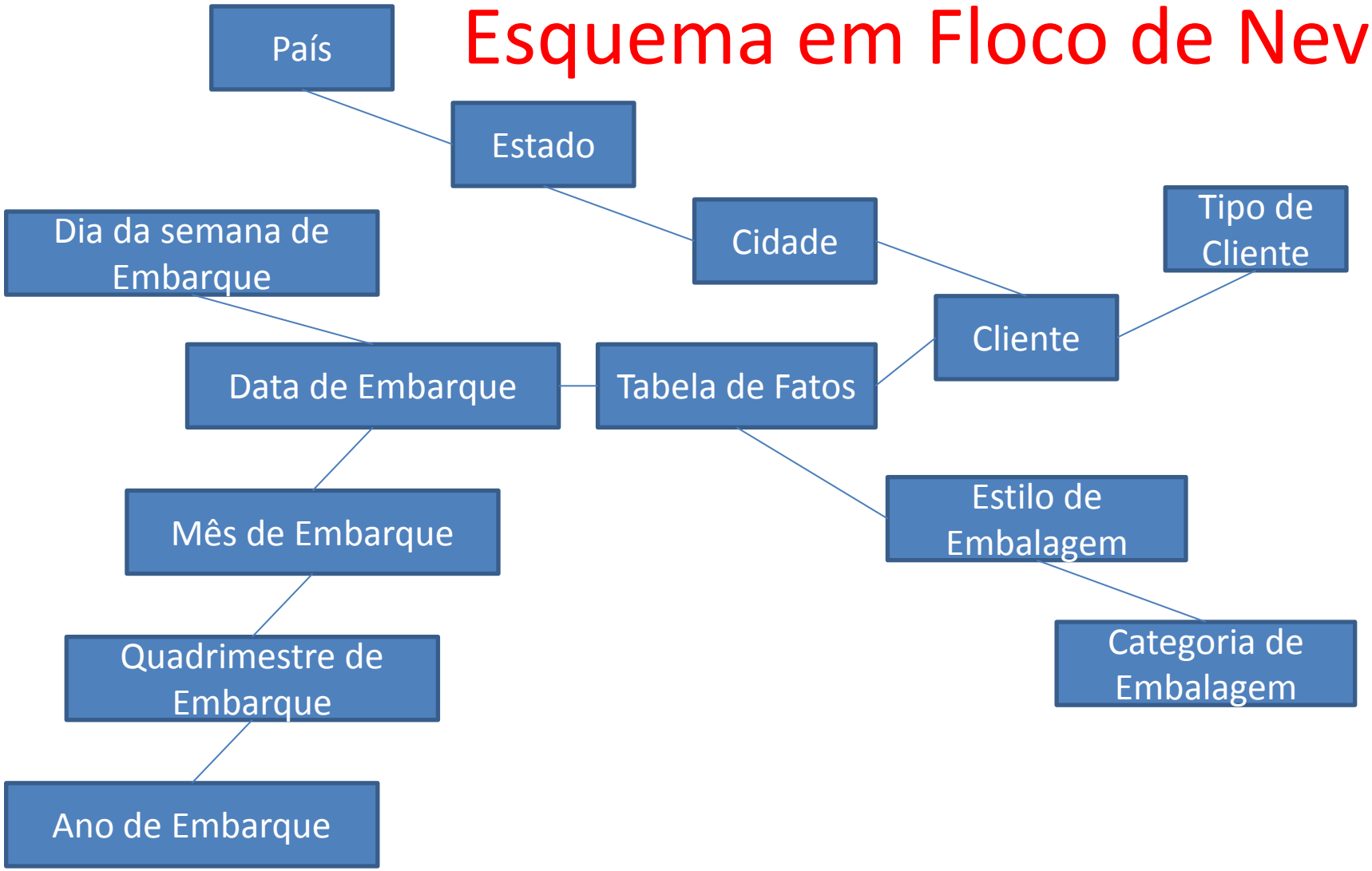
- Esquema em estrela (resumo)
  - Níveis hierárquicos não estão na 3ª FN, por que
    - Processo de normalização desmembraria cada tabela de dimensão do esquema no slide 18 em várias tabelas
    - Esquema normalizado resultante exige maior processamento de junção em muitas consultas
    - Tabelas de dimensão são pequenas em comparação a tabela de fatos
    - Tabelas de dimensão geralmente mudam lentamente
    - Núcleo das operações no data warehouse são operações de leitura

Data warehouse
Esquema em Floco de Neve

- Variação do esquema em estrela
  - Tabelas de dimensão são normalizadas
    - Cada nível hierárquico se torna sua própria tabela
- Está caindo em desuso
  - Esquema em estrela é mais rápido e simples
  - É mais fácil para o usuário entender quando cria consultas

Data warehouse

Esquema em Floco de Neve



On-Line Analytical Processing
OLAP

- Projeto e implementação de tabelas de fatos estratégicas são uma boa técnica quando se tem um pequeno conjunto de consultas frequentes
- Mas como realizar consultas ocasionais, que não foram previstas?
- Exemplo:
  - Procurar tarefas que não tem sido lucrativas
    - Necessita consulta capaz de subir e descer várias dimensões de dados
- Solução: **OLAP**

On-Line Analytical Processing
OLAP

- Sobrepõe o data warehouse
- Seleciona automaticamente um conjunto estratégico de ***visões-resumo*** e salva as ***tabelas-resumo automáticas (AST)*** em disco com visões normalizadas
- Mantém essas visões, deixando-as alinhadas com as tabelas de fatos à medida que novos dados chegam

On-Line Analytical Processing OLAP

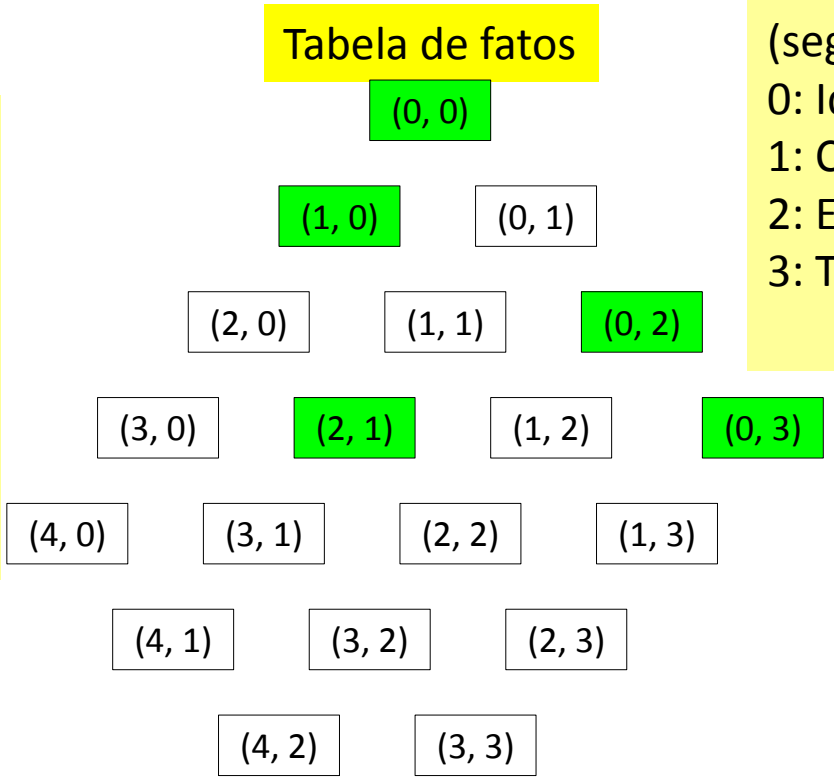
- Quando usuário solicita dados de resumo, o sistema OLAP descobre qual AST pode ser usada para dar resposta rápida à consulta indicada
- É uma boa solução quando existe a necessidade de exploração ocasional das informações resumo com base em grandes quantidades de dados residindo no data warehouse

OLAP

Como ele funciona?

- O sistema OLAP não pode não pode criar e manter todas as visões possíveis à medida que a dimensionalidade aumenta

Dimensão Data de Embarque
(primeira dimensão)
0: IdData
1: Mês
2: Quadrimestre
3: Ano
4: Tudo



Dimensão Cliente
(segunda dimensão)
0: IdCliente
1: Cidade
2: Estado
3: Tudo

Explosão Exponencial

$$\text{Visões possíveis} = \prod_{i=1}^{d} h_i$$

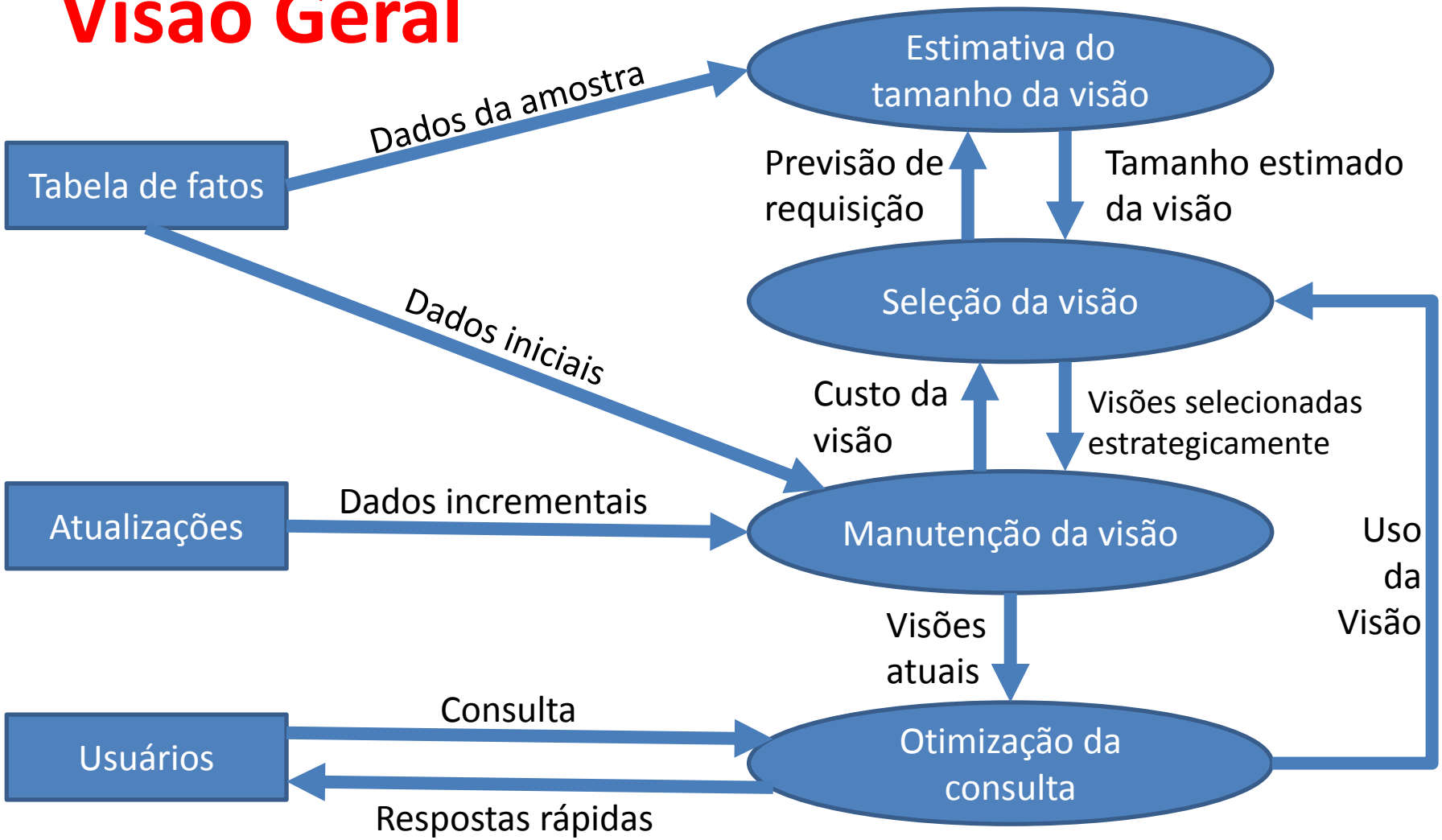
d = dimensões em um data warehouse

h_i = número de níveis hierárquicos ao longo de cada dimensão

OLAP
Como ele funciona?

- Criar ou manter todas as visões não dá!
  - Explosão exponencial
- Sistema OLAP precisa oferecer resposta rápida enquanto mantém o sistema dentro da limitação dos recursos
  - Seleciona um subconjunto estratégico de visões para materializar

OLAP
Visão Geral

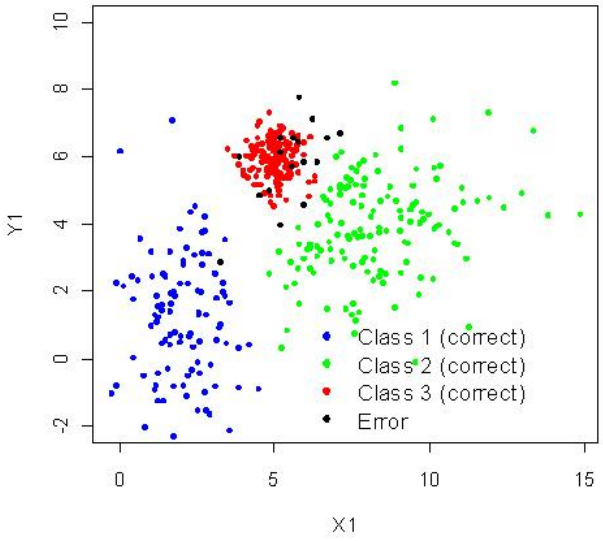


Data Mining

- Duas técnicas genéricas para extrair conhecimentos de um BD
  1. Usuário pode ter uma hipótese para confirmar ou refutar
     - Análise feita com consultas padrões e análise estatística
  2. Fazer com que o computador procure correlações nos dados e apresente hipóteses promissores para que o usuário leve em consideração
     - Data mining
       - Machine learning
       - Knowledge discovery

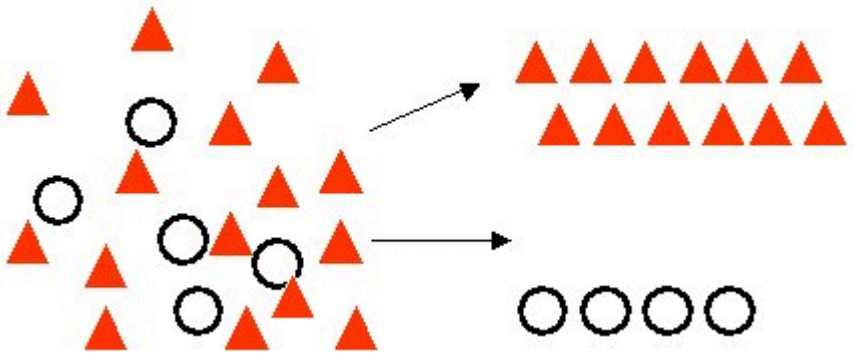
Data Mining

- Algoritmos tentam solucionar diversos problemas comuns
  - Problema de Categorização
    - Dado um conjunto de casos com valores conhecidos para alguns parâmetros, classificar os casos
    - Ex: dadas observações de pacientes, sugerir diagnóstico



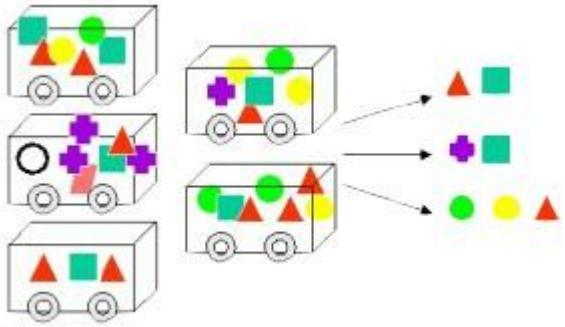
Data Mining

- Algoritmos tentam solucionar diversos problemas comuns
  - Problema de Agrupamento
    - Dado um conjunto de casos, encontrar agrupamentos naturais dos casos
    - Útil na identificação de segmentos de mercado



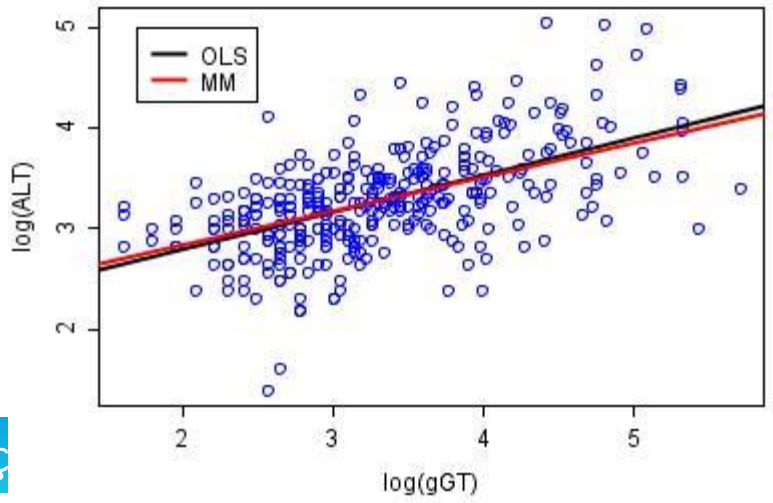
Data Mining

- Algoritmos tentam solucionar diversos problemas comuns
  - Problema de regras de associação, também conhecidas como análise de índices de preços ao consumidor
    - Empresas querem saber que itens frequentemente são comprados juntos
    - Conhecimento útil para tomar decisões sobre como dispor os produtos em um supermercado



Data Mining

- Algoritmos tentam solucionar diversos problemas comuns
  - Problema de previsão
    - A previsão é uma forma de data mining, usando dados conhecidos, e as tendências futuras são previstas com base no modelo
      - Modelo mais simples – mínimos quadrados



Gestão dos recursos de dados

- Projetar e configurar corretamente um Banco de Dados e o Data Warehouse é somente uma parte para se ter sistemas de informação confiáveis
- É preciso certificar que os dados para o negócio permaneçam precisos, confiáveis e prontamente disponíveis para quem deles necessita. É preciso estabelecer regras sobre como os dados serão organizados e armazenados, e quem terá permissão para vê-los ou alterá-los

Política de Informação

- Especifica regras para compartilhar, disseminar, adquirir, padronizar, classificar e inventariar informação.
- Elabora procedimentos e responsabilidades específicas, determinando quais usuários e unidades organizacionais compartilham a informação, para onde ela pode ser distribuída e quem é responsável por sua atualização e manutenção.

Administração de Dados

- Responsável pela políticas e procedimentos específicos pelos quais as informações podem ser gerenciadas como recurso organizacional
- Responsabilidades incluem desenvolvimento da política de informação, planejamento de dados, supervisão do projeto lógico do banco de dados e do desenvolvimento do dicionário de dados, e monitoração de como os especialistas em sistemas de informação e grupos de usuários finais utilizam estas informações.

Gestão de Banco de Dados

- Responsável por definir e organizar a estrutura a estrutura e o conteúdo do banco de dados, e também por sua manutenção.
- Em estreita colaboração com os usuários, o grupo de projeto determina o banco de dados físico, as relações lógicas entre elementos e as regras de acesso e procedimentos.

Assegurando a qualidade de dados

- Um banco de dados e uma política de informação bem projetados já é meio caminho andado para que a empresa tenha a informação de que precisa.
- Mas não é o suficiente...
  - Qual o impacto se um produto vendido estivesse com o preço errado no banco de dados?
  - O que aconteceria se o número do telefone ou o saldo bancário de um cliente estivesse incorreto?
- Informações incorretas, desatualizadas ou inconsistentes com outras fontes de informação criam sérios problemas operacionais e financeiros para as empresas. Levam a decisões incorretas, recall de produtos e até a prejuízos financeiros.

Assegurando a qualidade de dados

- Estes problemas com qualidade de dados podem ser causados por dados inconsistentes e redundantes produzidos por múltiplos sistemas
- Outra fonte de problemas é na entrada dos dados (grafia errada, números trocados, sem código, etc). Isto está aumentando à medida que os próprios cliente e fornecedores inserem seus dados via Web.
- Antes de implantar um novo banco de dados, as organizações precisam identificar e corrigir seus dados incorretos, além de estabelecer rotinas mais avançadas para editá-los quando o banco de dados estiver em operação

Auditoria de Qualidade de Dados

- Faz a análise da qualidade dos dados
  - É o levantamento estruturado da precisão e do nível de integridade dos dados em um sistema de informação
- Podem ser executadas com um levantamento completo dos arquivos de dados, de amostras desses arquivos ou da percepção dos usuários finais quanto à qualidade dos dados

Data Cleansing

- Limpeza e padronização
- Consiste em atividades para detectar e corrigir, dentro do banco de dados, informações incorretas, incompletas, formatadas inadequadamente ou redundantes
- Não corrige apenas os dados, mas também reforça a consistência entre diferentes conjunto de dados oriundos de sistemas de informação independentes.

Data Cleansing

- Existem software comerciais especializados que podem automaticamente fazer o Data Cleansing

Considerações Finais

- Problemas com qualidade de dados não são apenas empresariais. Elas também podem afetar seriamente indivíduos, causando problemas financeiros e até mesmo seu emprego e liberdade
- Quem nunca soube de algum caso de problema de crédito em que o cidadão estava erroneamente fichado no SEPROC?
- E a dificuldade para limpar a ficha?
- Que foi preso no lugar de outro por que no sistema estavam seus dados?

Referências

- Livros
  - Sistemas de informações gerenciais, Laudon & Laudon, 9ª Ed, Pearson, 2010
  - Projeto e modelagem de banco de dados, T. Teorey et all, 2ª Ed, Campus, 2014

Dúvidas