Modelação Multidimensional

Exemplo: Distribuição Retalhista

Sumário

- O Processo de análise
- Apresentação do caso
- Análise do caso
- Atributos das tabelas de dimensões
- Estender o modelo
- Notas sobre as dimensões
- Resumo / ideias a reter

O Processo de Análise

Passos do Processo de Análise. Factores

- Processo de análise em 4 passos
  - Seleccionar o processo de negócio a modelar
  - Declarar qual a granularidade do processo
  - Escolher as dimensões
  - Identificar os factos numéricos das tabelas de factos
- Elementos para a análise
  - Requisitos do negócio
  - Dados realmente disponíveis

O Processo de Análise

O processo de negócio e a sua granularidade

- Processo de negócio
  - actividade de negócio desenvolvida pela organização e suportada por sistemas de informação (Ex: compra de matérias primas, encomendas, vendas, etc)
  - Orientar a análise/desenho ao processo e não à organização
- Declarar a granularidade
  - Indicar o significado preciso de cada registo das tabelas de factos
  - Não esquecer que é sempre possível agregar mas não o inverso
  - Não esquecer qual a granularidade disponível nas fontes
  - Exemplos: linha de ticket das compras, snapshot diário de níveis de cada produto num sistema de inventário

O Processo de Análise

Identificar os processos a modelar e granularidade

- Identificar os processos a modelar
  - Combinar a **percepção do negócio** com os **dados disponíveis**
    - Ex: Base de dados com os movimentos diários por produto. Que produtos são vendidos em que lojas com que preços e em que dias?
- Determinar a granularidade da tabela de factos em cada processo do negócio
  - Determina a dimensionalidade da base de dados e tem um grande impacto na tamanho da base de dados.
    - Ex: Código de produto por loja por promoção por dia
  - O granularidade deve ser tão baixa quanto possível, pois para responder às interrogações a base de dados precisa de ser "cortada" de forma "precisa"

O Processo de Análise

Granularidade e dimensões

- A granularidade determina a dimensionalidade primária da tabela de factos.
  - Ex: tempo, produto e loja são as dimensões primárias
  - Dimensões adicionais podem ser adicionadas se compatíveis com a granularidade definida.
    - Ex:
      - Promoção em que o produto foi vendido
        Vendedor que forneceu o produto na loja
      - Gestor encarregado da loja nesse dia
  - Se for necessário adicionar uma dimensão não compatível com a granularidade definida, então é necessário rever a granularidade

O Processo de Análise

As dimensões e as tabelas de factos

- Dimensões
  - Como são em geral descritos os dados do "domínio"?
  - Processo + granularidade => dimensões
  - Dimensão tempo
  - Para cada dimensão:
    - Listar todos os atributos descritivos
- Tabelas de Factos
  - O que se pretende medir
  - Factos pertencentes a diferentes granularidades devem estar em tabelas de factos diferentes
  - As medidas são em geral aditivas

O Processo de Análise

Medidas da tabela de factos

- É necessário escolher que medidas básicas serão consideradas tendo em conta a sua disponibilidade bem como o processo necessário para a sua recolha.
  - Ex: No final de cada dia é necessário recolher o sumário das vendas diárias de cada loja:
    - Para cada produto:
      - Valor total das vendas
      - Número de unidades vendidas
      - Custo total dessas unidades vendidas
      - Número de clientes que compraram esse produto
- Estimar a dimensão da tabela de factos
    - Ex: 2 anos (2*365); 30 000 Produtos; 3000 produtos vendidos diariamente; 20 lojas; 47 milhões de registos.

Onde estamos?

- O Processo de análise
- Apresentação do caso
- Análise do caso
- Atributos das tabelas de dimensões
- Estender o modelo
- Notas sobre as dimensões
- Resumo / ideias a reter

Apresentação do caso

Uma empresa grossista

- 100 grandes superfícies de vendas (supermercado), espalhadas geograficamente por três estados.
- Todos os *departamentos* em cada superfície de vendas:
  - Mercearias; Comida congelada; Carne; Artigos limpeza e higiene; Padaria; Florista; Equip. eléctricos e electrónicos; Vinhos; ….
- Aproximadamente 60 000 produtos individuais nas prateleiras (unidades de stock armazenáveis - USA)
  - 55 000 USA provenientes de produtores externos (códigos de barras - Código Universal de Produto - CUP). 1 CUP => 1 USA
  - Diferentes formas de empacotamento de um mesmo produto correspondem a diferentes CUPs (e portanto USAs)
  - 5 000 USA produtos internos (carne, padaria, etc) sem CUP.

Apresentação do caso

Pontos de entrada de informação no S. Operacional

- Caixas (POS - Point of Sale)
  - Através dos códigos de barra nos produtos CUP, e nalguns USA não CUP.
  - Por entrada manual para alguns USA
- Pontos de entregas fornecedores
  - Apenas uma fracção dos armazéns utilizam a tecnologia de scanner para registar as entregas em tempo real.
- Departamento de Fornecedores e Contas a pagar
  - O completo conhecimento do material que entrou no supermercado só é possível, em muitos casos, por via dos pagamentos efectuados e por inspecção directa

Apresentação do caso

Principais preocupações / Objectivos

- A logística de encomendas, armazenamento nas prateleiras e venda dos produtos.
- Maximizar o lucro em cada supermercado.
  - Cobrar o máximo possível em cada produto,
  - Baixar os custos de aquisição dos produtos e os custos fixos
  - Atrair o máximo número de clientes
- Decisões mais significativas a tomar
  - Preços
  - Promoções (reduções temporárias de preços, anúncios, etc)
    - Baixas de preço servem para atrair clientes mas a venda é feita com prejuízo e a promoção pode baixar as vendas de outros produtos similares

Onde estamos?

- O Processo de análise
- Apresentação do caso
- Análise do caso
- Atributos das tabelas de dimensões
- Estender o modelo
- Notas sobre as dimensões
- Resumo / ideias a reter

Análise do caso

O processo de negócio a modelar

- O primeiro modelo de processo de negócio a construir deve ser aquele que maior impacto tiver nas expectativas dos utilizadores.
  - Deve responder às questões de negócio mais importantes e deve ser disponibilizado desde cedo aos utilizadores
- Analisar as compras dos clientes com base na informação recolhida nas caixas registadoras.
  - Deve ser possível analisar que produtos são vendidos, em que lojas, em que dias e qual o efeito das promoções

Análise do caso

Declarar a granularidade dos dados

- Neste caso temos dois níveis possíveis
  - Ao nível da linha de factura, isto é, quantas unidades são vendidas e a que preço em cada venda.
  - Ao nível das vendas realizadas para cada produto diariamente em cada loja
- Como pretendemos analisar o efeito das promoções e efectuar análises de associação de produtos comprados, é necessário considerar a granularidade mais baixa: **Linha de factura**
  - Não esquecer que é sempre possível agregar a partir de uma granularidade mais baixa, mas não o inverso

Análise do caso

Escolha das dimensões

- Granularidade determina as dimensões primárias
  - Um linha de factura corresponde a uma venda de um **produto** realizada numa **data**, numa **loja**.
- Encontrar outras dimensões que podem ser associadas
  - Muitas vezes o produto é vendido ao abrigo de uma promoção.
  - Várias linhas de factura estão associadas a um acto de venda (número de factura)

- Dimensões de base:
  - Data: data e não data + hora
  - Produto
  - Loja
  - Promoção: nem todas as vendas são feitas ao abrigo de uma promoção
  - Factura: número da factura

Análise do caso

Escolha das dimensões: StarSchema inicial

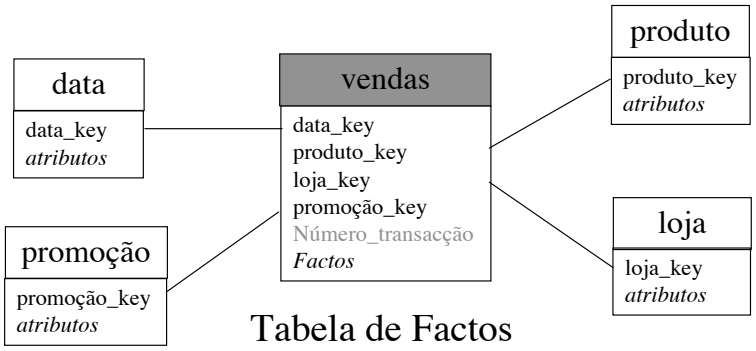


Tabela de Factos

Análise do caso

Identificar os factos

- Granularidade escolhida é chave para determinar os factos disponíveis. Numa linha de factura temos:
  - Quantidade: quantidade vendida em termos de número de unidades
  - Valor unitário
  - Valor total: = Valor unitário x Quantidade
  - Custo dos produtos vendidos: Em alguns sistemas de caixas é possível saber qual foi o preço a que a loja comprou o produto e portanto qual o custo (interno) dos produtos vendidos na transacção
- Factos
  - Unidades_vendidas
  - Valor_vendas
  - Custo
  - Lucro = Valor_vendas - Custo

  Factos aditivos por todas as dimensões
  - Margem de Lucro = Lucro / Valor Venda
  - Valor unitário: também não é aditivo; além disso não é relevante.

Análise do caso

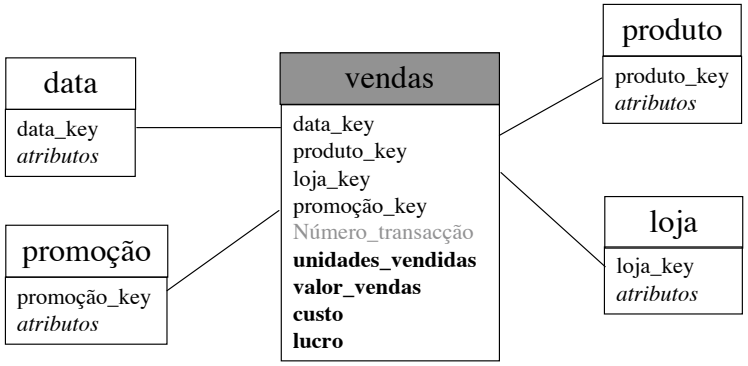
Identificar os factos: Discussão

- Factos de base aditivos
  - Unidades_vendidas
  - Valor_vendas
  - Custo
- Factos calculados e aditivos. Guardam-se ou calculam-se?
  - Lucro = Valor_vendas - Custo
  - Uniformidade nos valores independentemente do utilizador/relatório
- Factos calculados não aditivos. Calculam-se no fim
  - Margem de Lucro = Lucro / Valor Venda
  - Aggreg(Margem de Lucro) = Soma(Lucro) / Soma(Valor Venda)

- Estimar a dimensão da tabela de factos
  - Neste exemplo podemos considerar 2 biliões de linhas por ano

Análise do caso

Tabela de Factos



Onde estamos?

- O Processo de análise
- Apresentação do caso
- Análise do caso
- Atributos das tabelas de dimensões
- Estender o modelo
- Notas sobre as dimensões
- Resumo / ideias a reter

Atributos das Dimensões

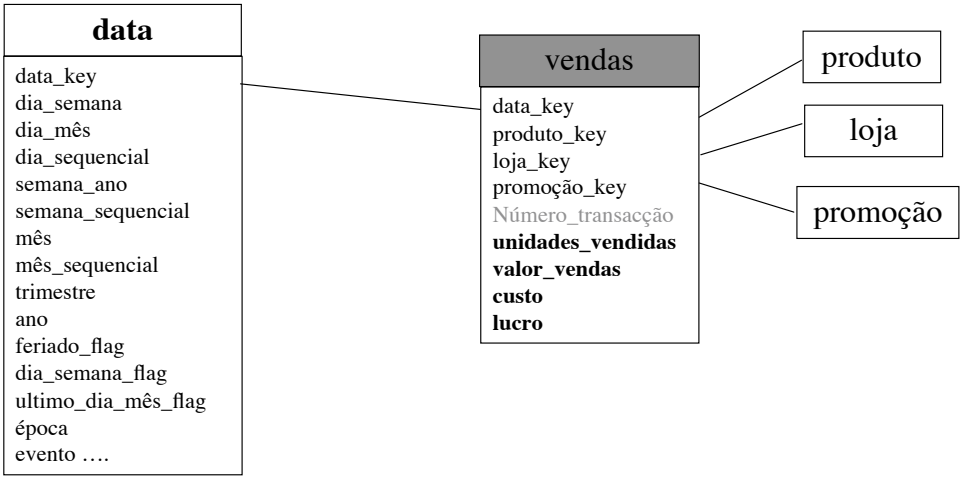
Discussão preliminar

- Porquê usar uma dimensão Data?
  - Porque não usar um atributo data na tabela de facto que seria directamente usado nas restrições?
    - Tamanho: 8 bytes para representação de data vs 4 bytes para inteiros
    - Evitar o join com a tabela Data (que é pequena)?
    - Atributos da dimensão tempo
- E o tempo?
  - Quando é necessário registar factos *ao longo do dia* usa-se uma dimensão Data e uma Dimensão tempo_do_dia.
  - O número de registos da dimensão **Data** é de 365 dias por ano. O número de registos da dimensão **tempo_do_dia** é de 24 **Horas** ou de 1440 minutos. Qualquer destas tabelas pode ser criada à priori.
  - A tamanho de uma tabela tempo seria de 8760 por ano (ao nível das horas) ou de 525 600 por ano (ao nível de minuto)

Atributos

- Data_key (inteiro)
- Data (tipo de dados data)
- Dia da semana (segunda, terça, …, domingo)
- Números relativos a uma data inicial.
  - Número do Dia gregoriano (consecutivos a começar numa dada data)
  - Número da Semana gregoriana (similar, mas a contar semanas)
  - Número do Mês gregoriano (similar, mas a contar semanas)
- Número do dia em relação à semana, mês, ano, ano fiscal, período fiscal
  - Dia do mês (1, …, 31), Dia do ano, Dia do ano Fiscal, …
- Número da semana em relação ao mês, ano
- ….
- Indicador de feriado
- Indicador de dia de semana (trabalho)
- Etc.

Dimensão de Data



- dia_semana e dia_mês - comparar as compras entre diferentes dias da semana ou do mês;
- feriado_flag, dia_semana_flag, ultimo_dia_mês_flag - comparação com dias especiais
- dia_sequencial, semana_sequencial, mês_sequencial - diferença entre datas
- época - (ex: Natal, Páscoa, etc)
- evento - (jogo Uefa, etc)

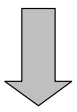
Dimensão Produto

- Hierarquia
  - USA (número e descrição)
  - Tamanho embalagem
  - Marca
  - Subcategoria
  - Categoria
  - Departamento
- Outros atributos
  - Tipo de embalagem
  - ...

- Manutenção actualizada da lista de USA
  => actualização da dimensão produto
- Não é necessário normalizar !
- Roll up / Roll down
  Agregar / Desagregar
- **É possível agregar e desagregar com outros atributos não pertencentes à hierarquia.**

Dimensão Produto: Roll up / Roll down

Dep.	Valor Vendido	Unidades Vendidas
D-1	780	263
D-2	1044	509
D-3	213	444
D-4	95	39



Desagregou departamento por marca

Dep.	Marca	Valor Vendido	Unidades Vendidas
D-1	M-1	300	160
D-1	M-2	480	103
D-2	M-5	...	…..
...	……	….	…..

Dimensão Produto: Roll up / Roll down (2)

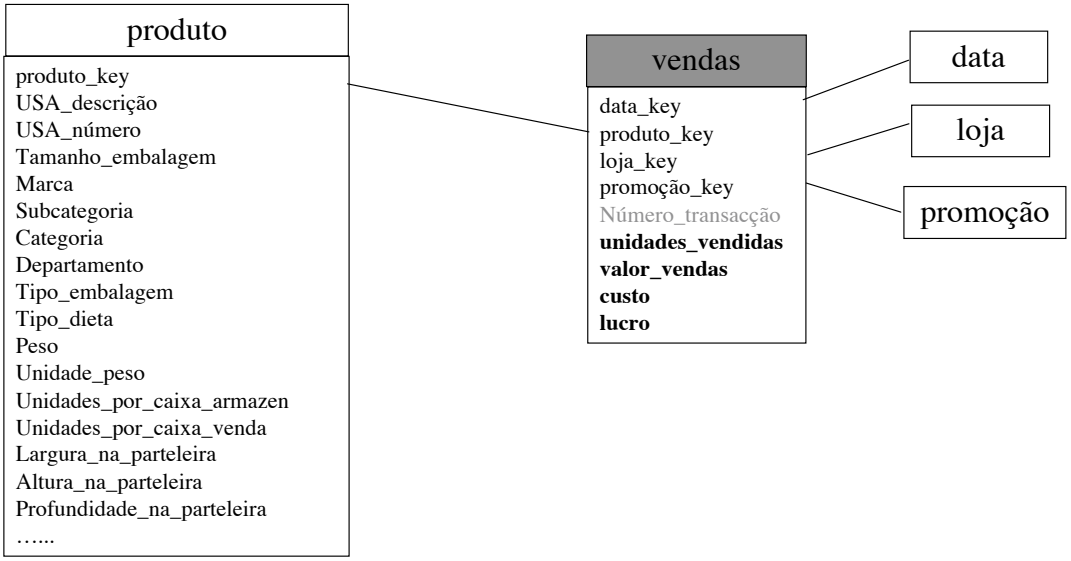
Dep.	Valor Vendido	Unidades Vendidas
D-1	780	263
D-2	1044	509
D-3	213	444
D-4	95	39

Desagregou departamento por tipo de embalagem

Dep.	Tipo Embalagem	Valor Vendido	Unidades Vendidas
D-1	E-1	100	50
D-1	E-2	280	75
D-1	E-5	...	…..
...	……	….	…..

Atributos das dimensões Produto

Tabela Produto



Dimensão Loja

- Dimensão geográfica do negócio
  - Uma ou mais hierarquias geográficas
    - Distrito / Concelho / Freguesia / Código postal
    - Região de vendas
- Atributos para caracterizar a organização da loja
  - Tipo de plano da loja
  - Dimensão da loja
  - Modelo financeiro
  - Número de empregados
  - ...

Dimensão Promoção

- Descreve as condições sob as quais decorreu uma promoção de um produto:
  - Reduções temporárias de preço; "coupons" de desconto; campanhas publicitárias; paineís
- Os gestores estão interessados em saber:
  - Os produtos em promoção aumentaram as vendas durante a promoção? (*lift*)
  - Depois da promoção houve uma baixa nas vendas que anulou os ganhos? (*time shifting*)
  - Outros produtos sofreram uma correspondente quebra nas vendas? (canibalização)
  - Os produtos em promoção tiveram um aumento das vendas tendo em conta o período anterior e posterior à promoção? (*crescimento de mercado*)
  - A promoção foi rentável considerando os aspectos anteriores e os custos directos da promoção?

Uma ou várias dimensões?

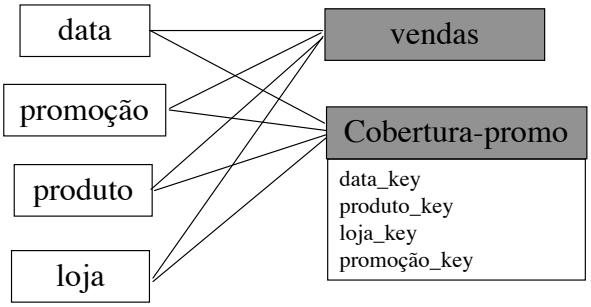
- As condições de uma promoção são os factores correlacionados de
  - Reduções temporárias de preço; “coupons” de desconto; campanhas publicitárias; paineis
- 4 Dimensões distintas?
  - A forte correlação não justifica separar em quatro dimensões
  - Uma única dimensão pode ser visitada de forma conveniente
  - As vantagens de passar para 4 dimensões poderiam ser:
    - Se os utilizadores pensarem em quatro mecanismos independentes (entrevistas!)
    - A administração da tabela única pode ser menos evidente pois necessita de uma chave artificial

E as vendas que não se realizam ao abrigo de promoções?

- Uma única dimensão promoções
  - Chave artificial
  - Cada registo refere-se a uma promoção combinada (de vários tipos de promoções)
  - Atributos classificativos e descritivos de cada tipo de promoções. Valores NULL quando os atributos não são aplicáveis.
  - Um registo especial singnificando que não há qulaquer promoção: “Sem promoção”
- Tabela de Factos
  - Na tabela de factos quando se regista uma venda numa data, numa loja de um producto que não está em promoção a chave estrangeira de promoção que se associa é aquela que corresponde a “Sem promoção”
- Regra geral
  - Evitar o uso de chaves nulas

Que produtos em promoção não foram vendidos?

- Uma promoção **A** sobre os produtos X e Y. Foram realizadas várias vendas de X e nehuma de Y
  - Na tabela de factos só existem registos (ligados à promoção **A**) das vendas de X.



Data ou Semana

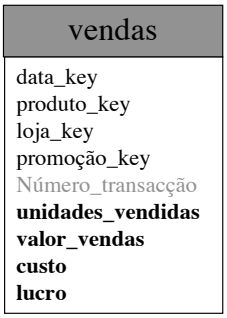
Factless Table

Um registo por cada produto numa promoção num dia numa loja

Granularidade diferente

Número da transacção

- Números de facturas, números de enconmendas, números de transacção constituem frequentemente chaves de dimensões degeneradas



- Dimensões degeneradas - dimensões vazias (sem atributos) e portanto sem tabela

Onde estamos?

- O Processo de análise
- Apresentação do caso
- Análise do caso
- Atributos das tabelas de dimensões
- Estender o modelo
- Notas sobre as dimensões
- Resumo / ideias a reter

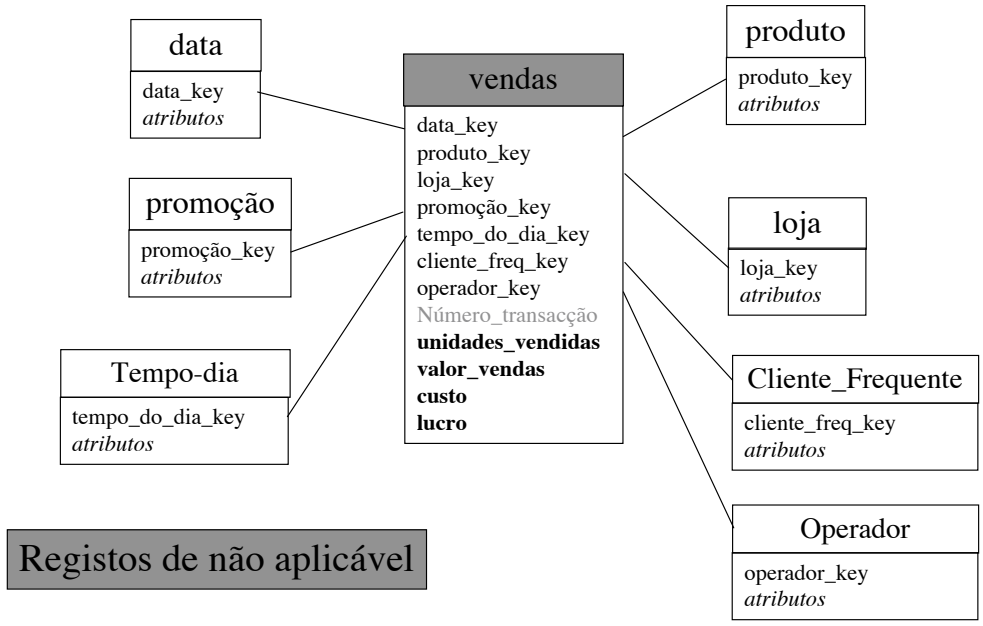
Estender o modelo

Novos requisitos

- Programa de cliente frequente
  - Criar uma tabela de dimensão de cliente frequente
    - Deve existir um registo cuja chave corresponderá a casos anteriores a este programa de cliente frequente. Esta chave será a que é colocada como chave estrangeira na tabela de factos para os factos históricos anteriores
    - Deve existir um registo cuja chave corresponderá a clientes que não aderiram ao programa de cliente frequente. Esta chave será a que é colocada em transações de clientes que não aderiram a este programa
  - E na tabela de factos
    - Juntar uma nova chave estrangeira na tabela de factos
- Controlo e análise dos operadores das caixas
  - Juntar uma dimensão **tempo do dia**
  - Juntar uma dimensão **operador de caixa**.
  - Juntar as novas chaves estrangeiras na tabela de factos (com os valores *correctos*).

Estender o modelo

Novos requisitos



Estender o modelo

Outras alterações

- Novos atributos de dimensões
  - Juntar os novos atributos
  - Se o valor dos atributos só faz sentido a partir de uma data determinda prever o valor "Não aplicável" ou "Não disponível" e colocar nos antigos registos da dimensão
- Novos factos
  - Pertencem ao mesmo evento e são da mesma granularidade: juntar nova coluna com os valores do novo facto (se não está disponível para registos históricos - NULL)
  - São de outra granularidade: criar nova tabela de factos
- Aumentar a granularidade de uma dimensão
  - É possível refinar uma dimensão. Construir uma nova dimensão que irá incluir os registos anteriormente existentes.
  - **Reconstruir** a tabela de factos para ligar à dimensão refinada. Aplicações anteriores podem continuar a funcionar.
- Outros casos
  - Novas tabelas de factos

Onde estamos?

- O Processo de análise
- Apresentação do caso
- Análise do caso
- Atributos das tabelas de dimensões
- Estender o modelo
- Notas sobre as dimensões
- Resumo / ideias a reter

Notas sobre as dimensões

Normalização: *snowflaking*

- Argumentos a favor da normalização das dimensões?
  - Espaço ocupado pelas dimensões
    - Não é relevante pois é a tabela de facto que maior espaço ocupa (3 Mb / 10 Gb)
  - Manutenção das tabelas de dimensões
    - Mais fácil se normalizado. Tarefa realizada na área de *staging*.
  - Tende a facilitar a navegação através de hierarquias simples
- Argumentos em desfavor da normalização das dimensões?
  - Desenho mais complexo
    - Utilizadores
    - Optimizadores de queries
  - Tende a "limitar" a navegação nas dimensões
  - Indequado ao uso de indexes de bitmaps (aplicados a atrbitutos de baixa cardinalidade)

Notas sobre as dimensões

Des-normalizar a tabela de factos!

- A ideia de incluir chaves por cada um dos elementos frequentemente analizados:
  - Produto: Tipo de produto; Classe; Departamento, etc
  - Loja: Tipo de loja; Região; etc
  - Data: Semana; Mês; Trimestre; Ano
- => Produz sistemas
  - Gigantescos
  - Pouco simples

Notas sobre as dimensões

Chaves primárias das tabelas de factos

- Chaves das dimensões devem ser artificiais
  - em geral inteiros sem significado, quando muito ordenadas.
- Razões são de ordem dirversa:
  - Desacoplar as chaves do DW das do(s) OLTP
    - Suposições sobre as chaves *naturais* podem ser invalidadas no futuro. Por exemplo re-uso de chaves antigas
    - Integrar fontes diversas com sistemas inconsistentes de chaves naturais
    - É possível usar chaves artificiais que não teriam significado no OLTP, como por exemplo "Não aplicável", etc.
  - Dimensão data
    - As chaves artificias devem ser inteiros cuja sequência tem significado
    - Permite representar "Data desconhecida", "Ainda não aconteceu", etc
    - Permite o particionamento das tabelas de factos com todas as vantagens para indexação de novos dados

Notas sobre as dimensões

Chaves primárias das tabelas de factos

- Razões são de ordem dirversa (cont):
  - Desempenho e Espaço
    - Inteiros tão pequenos quanto for possível (sabendo quantas linhas são necessárias).
    - 4 bytes => 2^{32}

- É necessário manter na àrea de staging tabelas de referências cruzadas entre as chaves do DW e a das fontes, para um adequado carregamento

- Dimensões degeneradas
  - Podem ou não usar chaves artificias, dependendo se os números usados (neste caso número da transacão) são ou não únicos em diferentes locais (lojas) ou se são ou não reutilizados
  - Não esquecer que estas dimensões podem eventualmente deixar de ser degeneradas

Resumo / Ideias a reter

Ideias a reter

- Desagregação é apenas juntar mais cabeçalhos de linha das tabelas de dimensões. Criando mais uma coluna que é um atributo de uma tabela dimensão
- Agregação é apenas retirar cabeçalhos de linha.
- Não é necessária uma hierarquia explícita para suportar a desagregação.

Análise do cabaz de compras

Ideias a reter

- Chaves das tabelas de dimensões devem ser artificiais (números sequências) e não dependentes de qualquer significado existente no OLTP
- Evitar (absolutamente) chaves com NULL nas tabelas de factos. Nas tabelas de dimensões devem existir chaves correnspondentes a “Não aplicável”, “Não disponível”, etc.
- Tabelas de factos sem factos - factless tables - servem para contagens de eventos.
- Duas tabelas indexadas pelas mesmas dimensões podem representar granularidades diferentes
- É possível juntar novas dimensões a uma tabela de factos e as aplicações anteriores permanecerem inalteradas