Data Warehouse

1º Semestre

Mestrado em Desenvolvimento de Software e Sistemas Interativos

Instituto Politécnico de C.Branco

11ª Edição – 2021/22

Slides #4 – Modelação Multidimensional

Eurico Lopes

V.09-10.10.19

1

Índice

2

LR – Levantamento de Requisitos

- **Sumário**
  - Neste capítulo apresenta-se o método – *Dimensional Modeling* – utilizado para modelar Data Warehouses dimensionais.
  - A compreensão deste método é indispensável para o desenho de um Data Warehouse.

3

ER - Entidades e Relacionamentos

- Método com objetivos de normalização e eliminação de Redundância
- Adequado para sistemas transacionais, porque simplifica, nomeadamente, as transações de atualização e "lockup"
- Qualquer sistema suportado por um modelo relacional tem, no mínimo, dezenas de tabelas relacionadas
- Modelos pouco adequados à pesquisa:
  - A complexidade e extensão dos modelos não permitem a sua rápida interpretação
  - Penalizam drasticamente a performance de "queries",pelo número usual de "joins" necessários

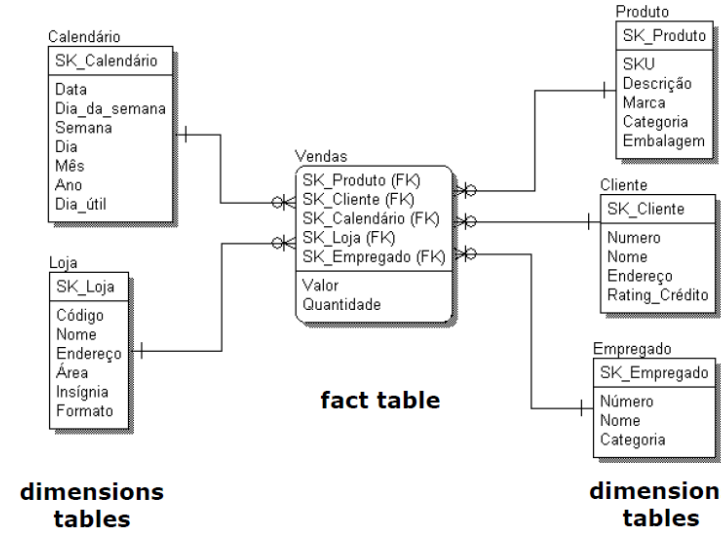
4

MM - Modelagem Multidimensional

- Técnica de modelagem cujo objetivo é organizar os dados segundo modelos "standard" (modelos em estrela/"**star schemas**"), intuitivos e que otimizam a performance de acesso.
- Cada modelo é composto por uma tabela com chaves compostas (tabela de factos/ "**fact table**") e um conjunto de pequenas tabelas (dimensão/ "dimension") chaves.
- Cada dimensão tem uma chave simples correspondente a cada um dos componentes da chave composta da tabela de factos.
- As tabelas de factos traduzem, na terminologia E&R, relacionamentos *n-para-n* (muitos-para-muitos).
- Os atributos próprios da tabela de factos representam sobretudo valores numéricos passíveis de serem somados (aditivos).

5

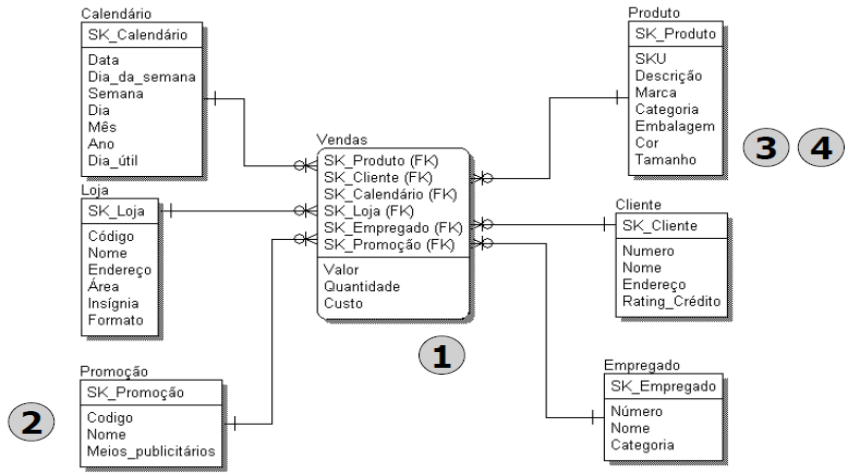
MM - Modelagem Multidimensional



6

MM – Pontos fortes

- **Os modelos facilmente acomodam alterações de desenho:**



7

MM – Pontos fortes

1. Adição de novos atributos à tabela de factos desde que consistentes com a granularidade atual
2. Adição de novas dimensões desde que os registos atuais da tabela de factos assumam um único valor desta dimensão
3. Adição de novos atributos a uma dimensão
4. Baixar a granularidade (mais granular) de uma dimensão a partir de um ponto no tempo

8

MM - Dimensões Coerentes *(conformed dimensions)*

- Dimensão coerente é uma dimensão que tem o mesmo significado qualquer que seja tabela de factos com a qual possa ser ligada:
  - Uma dimensão coerente é partilhada pelos diversos data marts que a referenciam. Exemplos:
    - cliente, fornecedor, produto, tempo (calendário)
  - Do ponto de vista da consistência é uma das vantagens dos modelos ER aplicadas na modelagem multidimensional;
  - Tornam possível a mesma interpretação do conceito e respetivos atributos ao longo dos diferentes data marts;
  - Potenciam o cruzamento de informação de diferentes data marts;
  - Representa 80% do esforço de modelagem.

9

MM - Factos Coerentes *(conformed facts)*

- Definição coerente de factos:
  - Indispensável quando se pretende usar a mesma terminologia e cruzar informação de diferentes data marts;
  - Estes factos devem estar definidos para o mesmo contexto de dimensões entre os diferentes data marts;
    - Exemplos: tempo (calendário) e unidades de medida
  - Nos casos em que não seja possível desenhar os factos coerentes deve ter-se o cuidado de os designar com nomes diferentes;
  - Uma forma fácil de garantir esta coerência é exprimir os factos segundo dimensões atómicas.

10

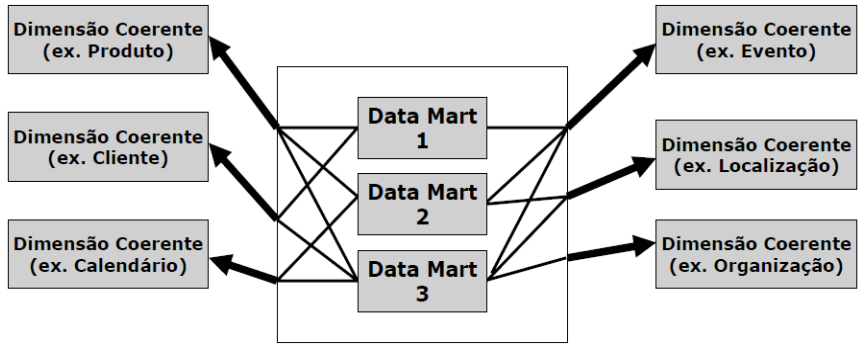
MM - Factos Coerentes *(conformed facts)*

- Granularidade dos factos:
  - A definição granular de factos, quando possível, potencia a extração de informação que é feita usualmente nos sistemas operacionais
  - Prepara o DW para futura exploração com técnicas de data mining.

11

MM - Dimensões e Factos Coerentes

- Um DW com dimensões e factos coerentes são como um "bus interface" que permite a adição sucessiva de novos data marts



12

SK - Chaves anónimas *(Surrogate Keys)*

- As chaves anónimas nas dimensões coerentes:
  - Tornam o DW mais independente dos sistemas operacionais:
    - reutilização de chaves
    - Mudanças na estruturas das chaves
    - Possibilitam que no DW se mantenham diferentes versões do mesmo registo dos sistemas operacionais (diferentes chaves anónimas para a mesma chave operacional).

13

MM - Chaves anónimas

- **A minha ficha de cliente no meu Banco:**

Número Cliente	Nome Cliente	Estado Civil	Endereço	Localidade	CP	País	...	Data Vigor
123	Eurico Lopes	Casado	Rua Y	C.Branco	6000	Portugal		04-Jan

- **A minha informação no DW do meu Banco:**

SK Cliente	Número Cliente	Nome Cliente	Estado Civil	Endereço	Localidade	CP	País	...	DataVigor
100001	123	Eurico Lopes	Casado	Rua x	C-Branco	3700	Portugal		01-Jan
100002	123	Eurico Lopes	Casado	Rua y	Lisboa	4700	Portugal		01-Out
100003	123	Eurico Lopes	Casado	Claypit	Leeds	LS2 8WG	Inglaterra		01-Mai
100004	123	Eurico Lopes	Casado	Rua k	Porto	4100	Portugal		03-Mai
100005	123	Eurico Lopes	Casado	Rua k	Porto	4100	Portugal		15-Mar

14

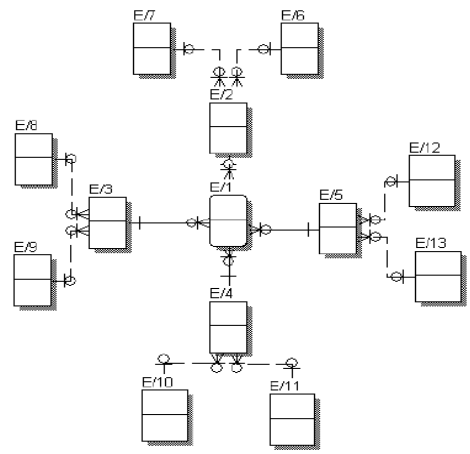
MM - Do E&R ao MM

- **Um modelo E&R converte-se num modelo multidimensional:**
  1. Isolando partes do modelo por processo de negócio;
  2. Promovendo as entidades que representam relacionamentos *n-para-n* com atributos numéricos e aditivos a tabelas de factos;
  3. Desnormalizar as restantes tabelas em tabelas com chaves simples relacionadas diretamente com as tabelas de factos.
- **Um DW "grande" contará com 10 a 25 modelos em estrela, cada um deles com 5 a 15 dimensões, muitas delas partilhadas por vários modelos**

15

MM - Floco de Neve (*snowflaking*)

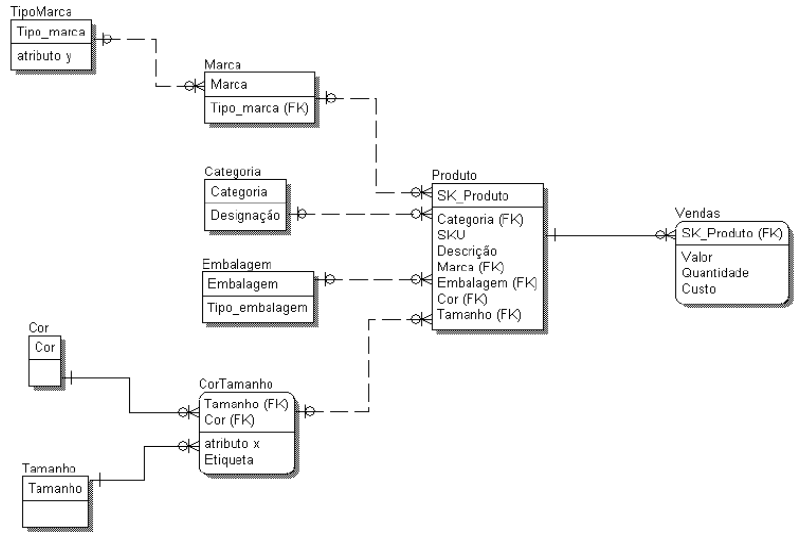
- **Estamos em presença de um "floco de neve" quando alguns campos, usualmente de cardinalidade baixa, são retirados de uma dimensão para uma tabela separada e ligados por um relacionamento**



- Aplicação da normalização aos modelos em estrela

16

MM - Floco de Neve (*snowflaking*)



17

MM - Floco de Neve (*snowflaking*)

- Solução a evitar porque:
  - Tornam os modelos mais complexos
  - Prejudicam a navegação pelo modelo das ferramentas de query
  - Inadequado com bitmap indexes
    - Os bitmap indexes tiram partido dos atributos com baixa cardinalidade que são retirados pela *snowflake*

18

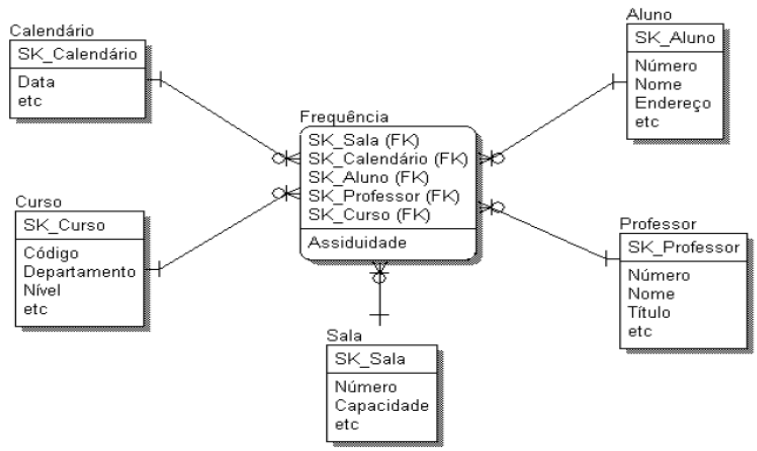
MM – Tipos de Factos

- Factos Aditivos (SQL SUM function)
  - O caso mais comum, em que os factos são medidas numéricas que podem ser somados
    - Ex.: vendas em quantidade, vendas em valor
- Factos semi-aditivos (SQL AVG function)
  - Factos que expressam medidas de intensidade, isto é, posições de uma determinada medida num determinado instante
    - Ex. Nível de stock, saldo contabilístico
- Factos não aditivos (Factless Facts)
  - Ex.: Eventos e ocorrências

19

MM - Factless Fact Tables

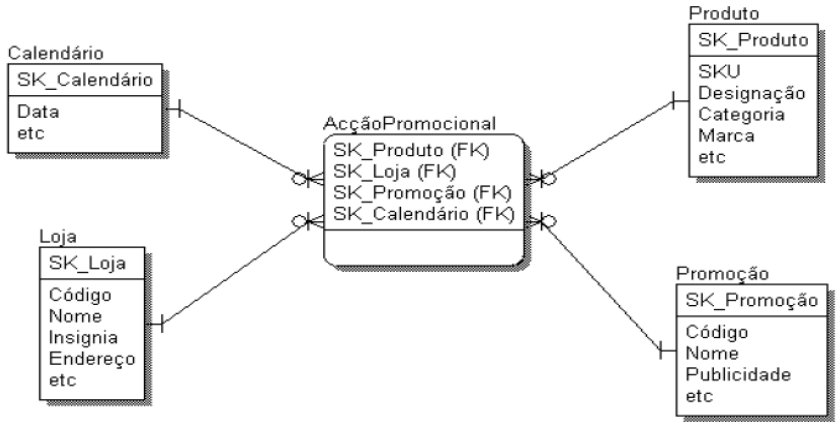
- Solução para modelar **eventos**



20

MM - Factless Fact Tables

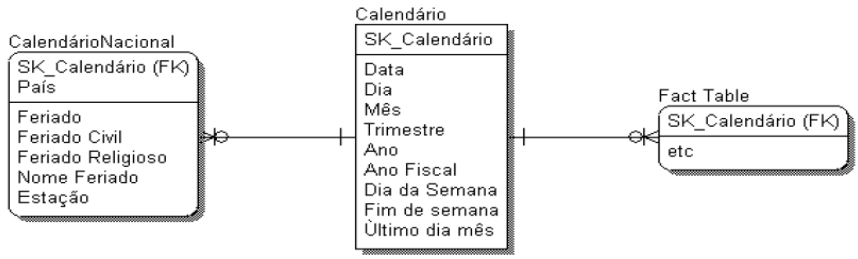
- Solução para modelar **ocorrências**



21

MM - Algumas Dimensões Genéricas

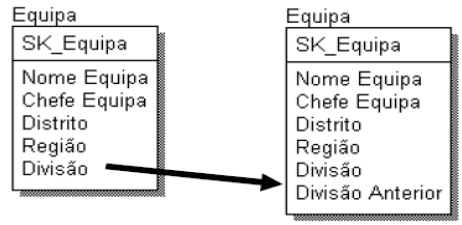
- Dimensão **Calendário**



22

MM - Gestão de Alterações nas Dimensões

- **Opções possíveis:**
  1. Ignorar a alteração
  2. Sobrepor a alteração -> **ignorar a "história"**
  3. Acrescentar um novo registo à dimensão -> **múltiplas versões históricas**
  4. Utilizar um campo "valor anterior" -> **2 versões históricas** (atual e anterior)
  5. Criação de snapshots como variante da solução 3



23

MM - Gestão de Alterações nas Dimensões

- Exemplos de aplicação:
  - Dimensões com poucas alterações (*slowly changing dimensions*):
    - Aplicação do método 3
  - Dimensões pouco extensas com alterações frequentes (*rapidly changing dimensions*):
    - Aplicação do método 3 pode ser a solução

24

MM - Gestão de Alterações nas Dimensões

- Exemplos de aplicação (cont.):
  - Dimensões extensas:
    - Compromisso entre a performance e o tratamento das alterações, conforme se trate de uma dimensão com alterações pouco ou muito frequentes
  - Dimensões extensas com alterações frequentes:
    - A solução de criação de snapshots pode ser uma boa opção como variante da solução pura de acrescentar novos registos

25

MM - Gestão de Alterações nas Dimensões

- A solução a adotar será sempre um compromisso entre os impactos técnicos e as necessidades do negócio, no tratamento da evolução temporal das dimensões:
  - que evoluções são importantes retratar
  - qual o impacto na arquitetura (disco!) e performance

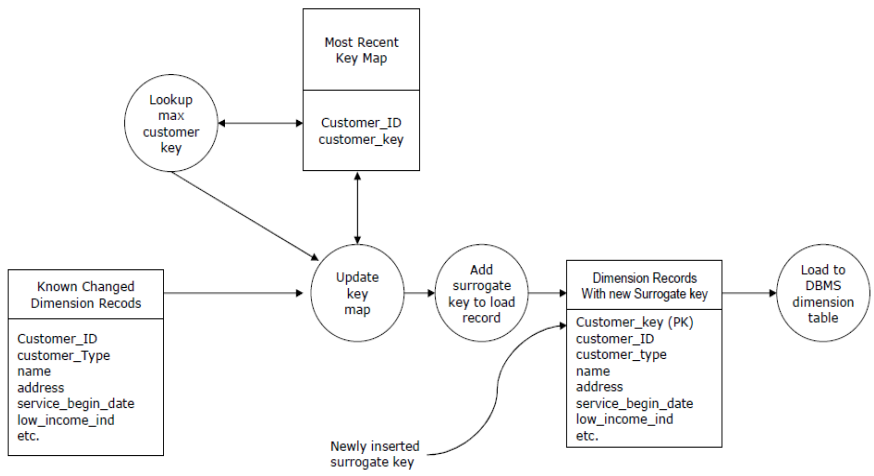
26

MM - Identificação de Alteração de Registos de Dimensões

- **Sistemas operacionais marcam os registos alterados**:
  - Status
  - Timestamp
- Comparação da "versão de hoje" com a "versão de ontem"
- Problema na identificação de registos apagados quando não tratados como registos inativos pelos sistemas operacionais

27

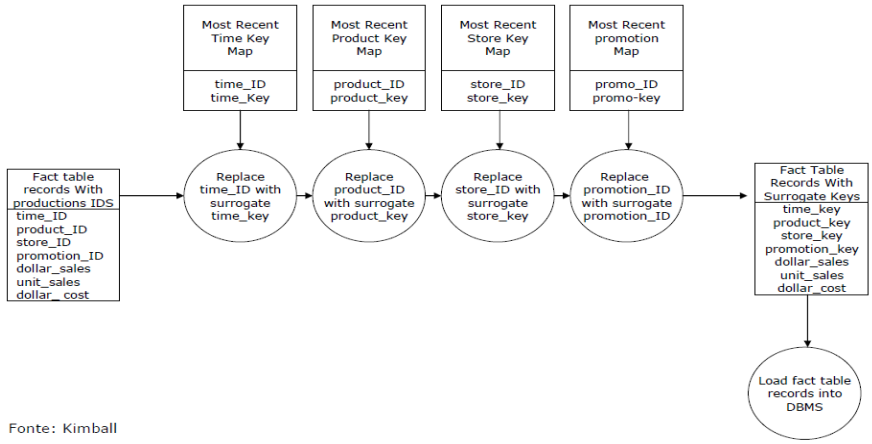
MM - Processamento de Alterações numa Dimensão



Fonte: Kimball

28

MM - Processamento de Tabelas de Factos



Fonte: Kimball

29

MM - Exemplos Práticos de Arquiteturas

- Data Warehouse integrado com ERP
  - **Sincronização de processos "batch"**
- Controlo centralizado do processo pelo Data Warehouse
  - **Mecanismos de sinalização**
- Extração a partir de áreas de interface partilhadas
  - **EAI**

30

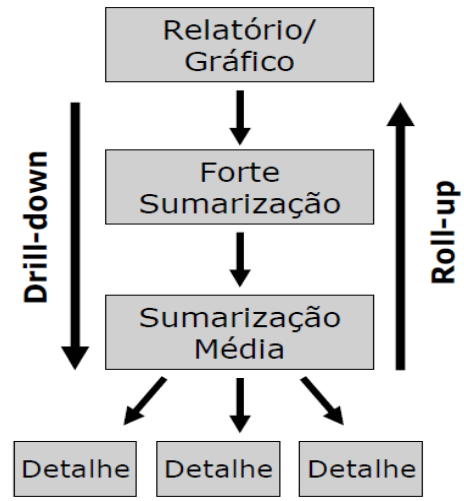
DW - Um Modelo em 4 Passos

- Seleção do data mart
  - Fontes da dados
- Granularidade da tabela de factos
- Seleção das dimensões relevantes
  - Tratamento das alterações nas dimensões
  - Hierarquias para possíveis agregações
- Seleção dos factos a retratar e respetivos atributos
  - Histórico a manter

31

MM - Porquê Agregar

- Tendência do método topdown para análise da informação
- Aproximar a representação dos dados à forma como são interrogados
- Performance, performance, performance, ...



32

MM - Porquê Não Agregar

- Necessário mais espaço de armazenamento das agregações e mais tempo para as executar;
- Introdução de maior complexidade nos modelos pela introdução de mais tabelas e consequente complexidade nos programas que lhes acedem;
- O hardcode das agregações e o automatismo da respetiva utilização limitam a flexibilidade de atuação do DBA;
- Dificuldade em perceber o que deve ser agregado.

33

MM – O Que Agregar

- Identificar no processo de levantamento de requisitos:
  - atributos usualmente agrupados segundo uma hierarquia;
  - combinações de agrupamentos usadas.
- Pela análise dos dados avaliar qual a redução de granularidade entre níveis de cada hierarquia:
- Recolher estatísticas de utilização:
  - Perceber a atual utilização de cada granularidade;
  - Perceber a necessidade de novas agregações;
  - Conhecer problemas de performance.

34

MM – Quando Agregar

- Agregação incremental:
  - Semanas incompletas / Meses incompletos;
  - Semanas/Meses completos com dias pontuais em falta;
  - Publicação de múltiplas versões da verdade;
  - Modalidade usada quando as agregações fazem parte do processo de carregamento.
- Agregar quando todos os dados atómicos estão disponíveis:
  - As partes comprometem o todo.

35

MM – Agregação: Não Esquecer

- Agregação = Redundância
  - Considerar overhead da introdução de agregações (pelo menos 100%);
  - As agregações devem reduzir pelo menos 10 vezes a respetiva tabela atómica;
  - Se o DBA “torcer o nariz” pague-lhe um jantar!
- Custos (tempo) de agregar ou reagregar
- A introdução de novas agregações pode e deve ser transparente para o utilizador.

36

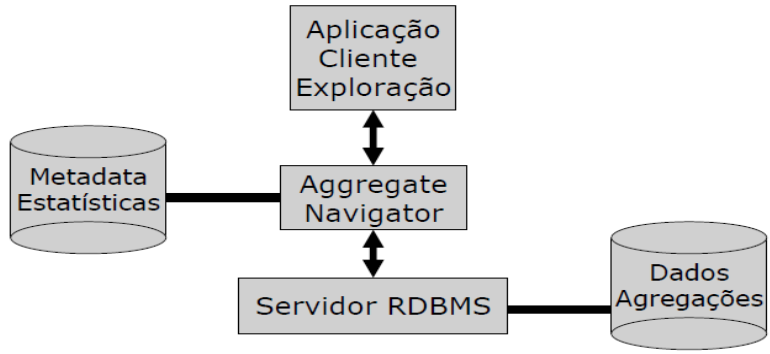
MM - Arquitetura de Agregação

- Realizar as agregações para tabelas separadas (lógica e fisicamente), das tabelas de factos atómicas:
  - Cada nível de agregação sua tabela.
- Desvantagens das agregações embebidas:
  - Atributos dos factos com campos maiores para suportarem valores maiores (SUM);
  - Introdução de novos atributos com valores nulos para a maior parte das ocorrências (MIN, MAX, AVG, …);
  - Tabelas maiores;
  - Dificuldades com as query tools.

37

MM - Navegação pelas Agregações

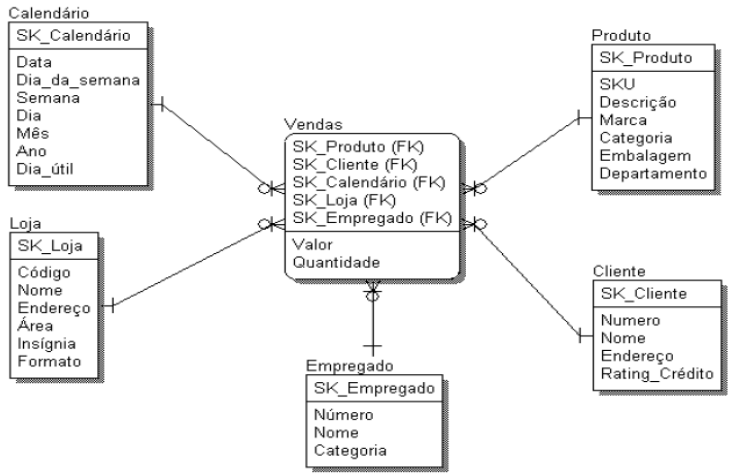
- O Aggregate Navigator intercepta pedidos de SQL atómicos e envia-os ao RDBMS como pedidos de SQL sobre agregações (aggregate-ware SQL).



38

MM - Exemplo de Agregação

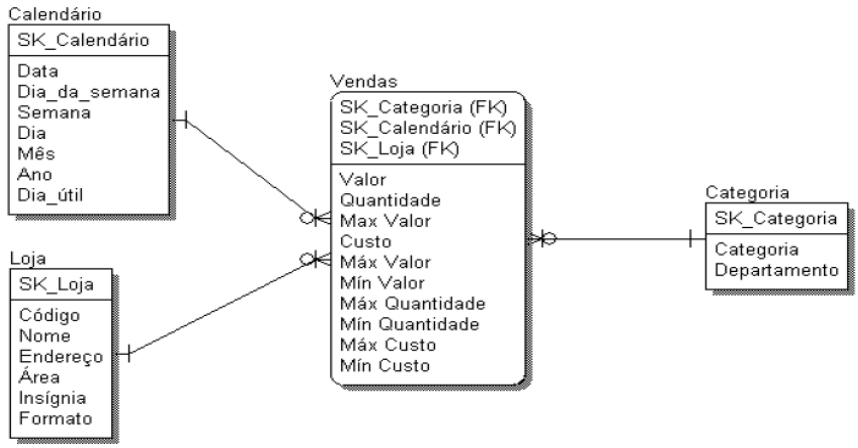
- Tabela Base:



39

MM - Exemplo de Agregação

- Agregação da Tabela Base:



40

MM – Exercício Prático (1)

- **Objetivo:**
  - Construir um modelo dimensional que suporte a exploração dos KPI's sobre o Service Desk;
- **Fontes:**
  - Lista de KPI's e respetiva definição
  - Modelo E&R do aplicativo Remedy Customer Support.

41

MM – Exercício Prático (2)

- **Objetivo:**
  - Identificar e desenhar as potenciais agregações ao modelo base do Service Desk;
- Fontes:
  - Modelo Dimensional do Service Desk previamente construído.

42

DW – Documentos complementares

- **Documentos disponibilizados na plataforma:**
  - Livro: Kimball et all, (2008) *The Data Warehouse Lifecycle Toolkit*, 2nd Ed. Wiley
    - **Ch06-A Graduate Course on Dimensional Modeling**
  - Database Management Systems, 2nd Edition. R. Ramakrishnan and J. Gehrke:
    - **12 – Ramakrishnan.zip**
      - Ch23a_DecSup-95.pdf (Data Warehousing and Decision Support – Slides part A)
      - Ch23b_Views-95.pdf (Data Warehousing and Decision Support – Slides part B
    - 14 - Exemplo DW Retalho

43

DW – Links

- **Links**
  - **Dimensional Modeling and E-R Modeling**
    - **http://www.dkms.com/papers/dmerdw.pdf**
  - **The Problem with Dimensional Modeling**
    - **http://www.information-management.com/issues/20000501/2184-1.html**
  - **DW Dimensional Modeling Techniques**
    - https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/
  - **Generating a Multi-Dimensional Model – ORACLE Tutorial**
    - https://www.oracle.com/webfolder/technetwork/tutorials/obe/db/sqldevdm/r40/datamodel4genmulti/datamodel4genmulti_otn.html

(Todos os links acedidos 10/out/2019)

44