



Modelando um Data Warehouse

GRIMALDO OLIVEIRA

Sobre Grimaldo



- Grimaldo Oliveira
 - grimaldo_lopes@hotmail.com
- Formação
 - Mestre em Tecnologias Aplicadas a Educação pela Universidade do Estado da Bahia.
 - Especialização em Análise de Sistemas pela Faculdade Visconde de Cairu.
 - Estatístico pela Universidade Federal da Bahia.
- Atividades
 - Mais de 15 anos atuando como Consultor de Business Intelligence.
 - Projetos Governos Maranhão, Mato Grosso e Bahia.
 - Coordenador da Área de Informações Estratégicas - PRODEB
 - Idealizador do Blog : BI com Vatapá – bicomvatapa.blogspot.com.
 - Livro: BI Como Deve Ser – bicomodeveser.com.br
 - Professor EAD - Aprenda Virtual - aprendavirtual.com

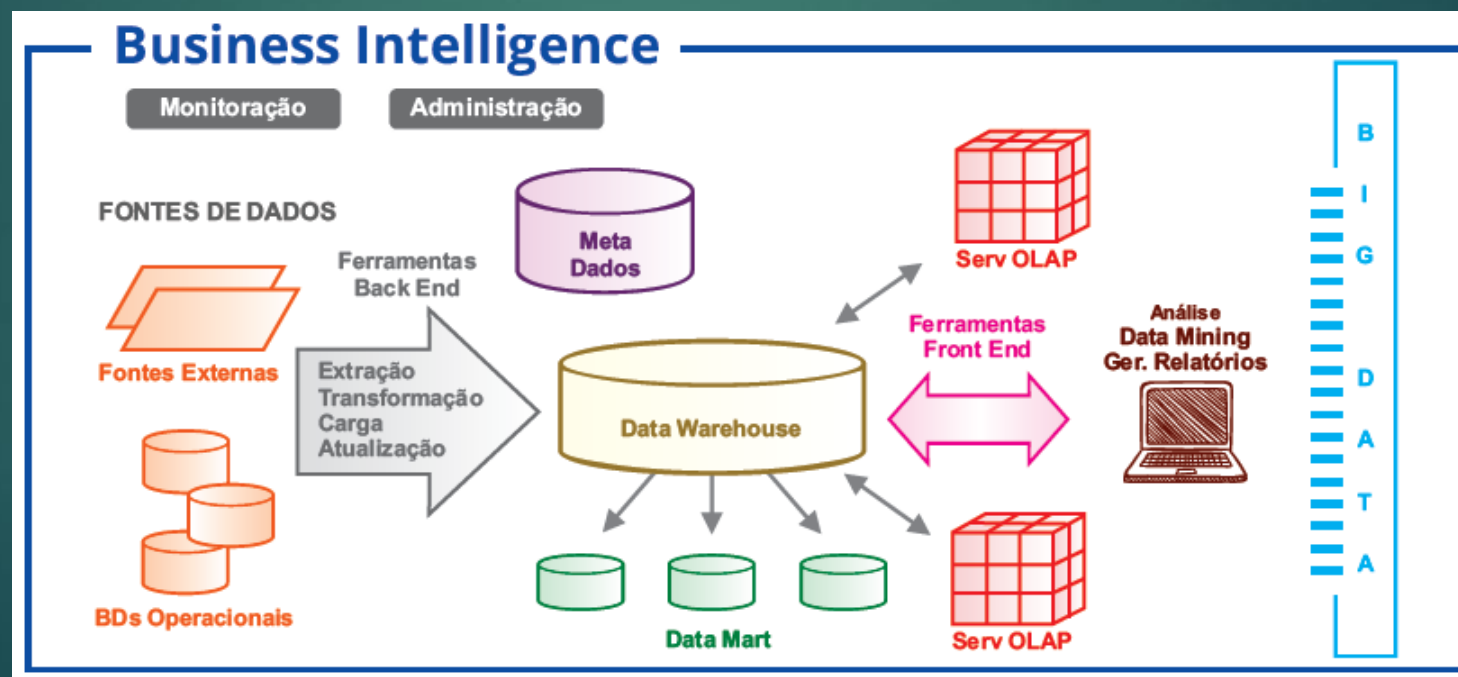
Agenda

- Objetivo
 - Apresentar os principais assuntos na construção de um projeto de Business Intelligence
 - O que é Business Intelligence
 - O que é Data Warehouse
 - Staging Area
 - Dimensão
 - Fato
 - Modelo Multidimensional
 - Etapas do processo de construção do Data Warehouse
 - Levantamento de dados com gestores: Matriz de Necessidades
 - Levantamento de dados nas bases transacionais: Fonte de Dados
 - Construção do modelo multidimensional

Business Intelligence

➤ O QUE É BUSINESS INTELLIGENCE

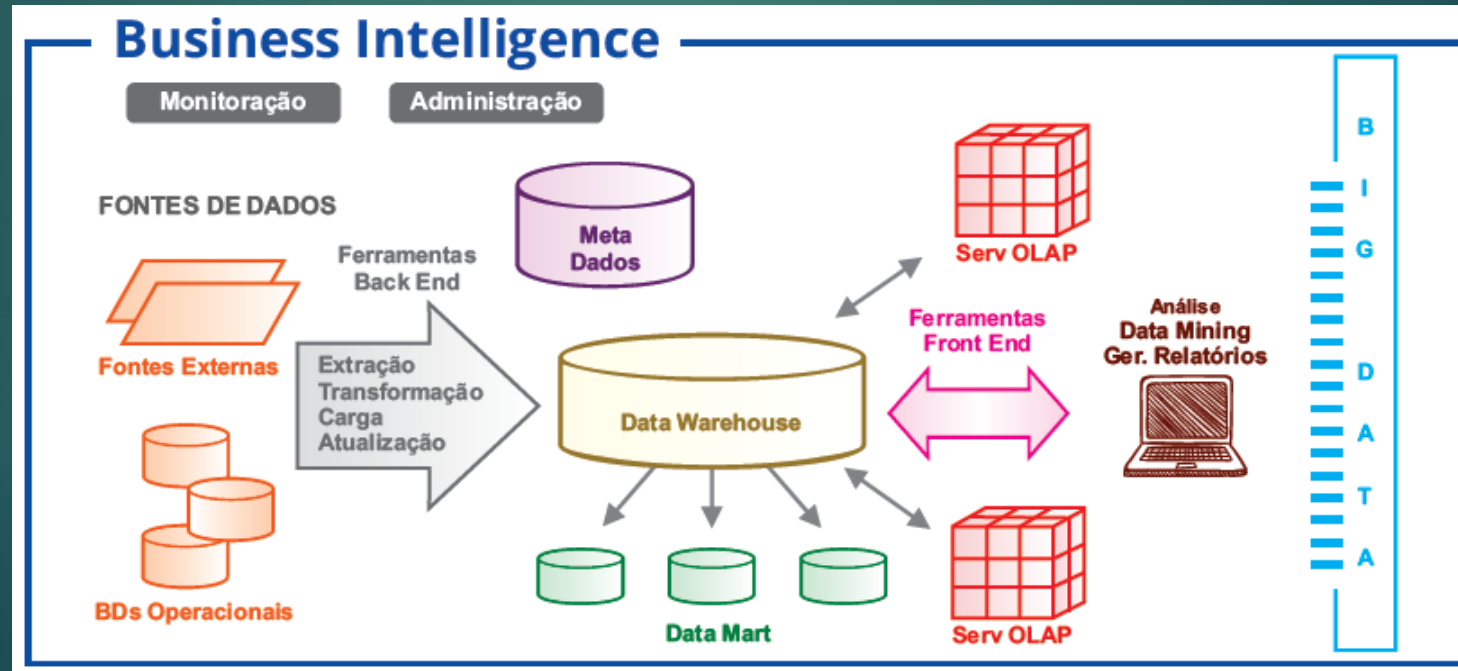
- Na tradução livre, significa Inteligência de negócios ou simplesmente “BI”. É um conjunto de técnicas e ferramentas que, reunidas, possibilitam ao gestor a tomada de decisão com base em métricas ou valores que são extraídos de seus diversos sistemas e consolidados.



Data Warehouse

➤ O QUE É DATA WAREHOUSE

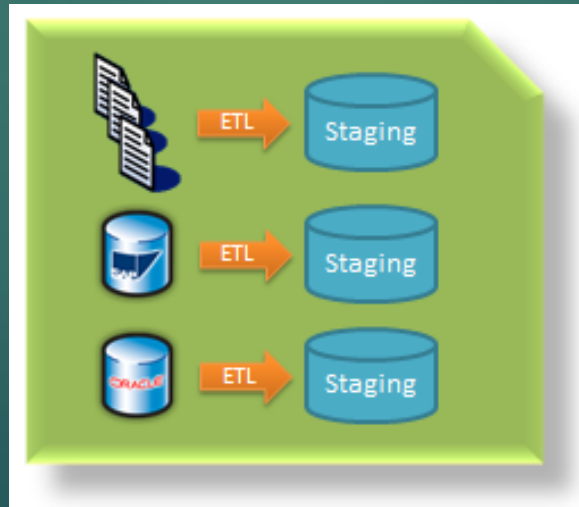
- Também conhecido como “armazém de dados”, é um banco de dados que integra e consolida as diversas fontes de dados (arquivos de texto, banco de dados, mensagens, dentre outras) de uma organização para apoio na tomada de decisão.



Staging Area

➤ STAGING AREA

- Staging Area ou área auxiliar serve como ponto único para a carga efetiva no Data Warehouse.
- A cada carga seu conteúdo é limpo.
- Evita acesso à produção em caso de recarga durante o dia.



Dimensão e Fato

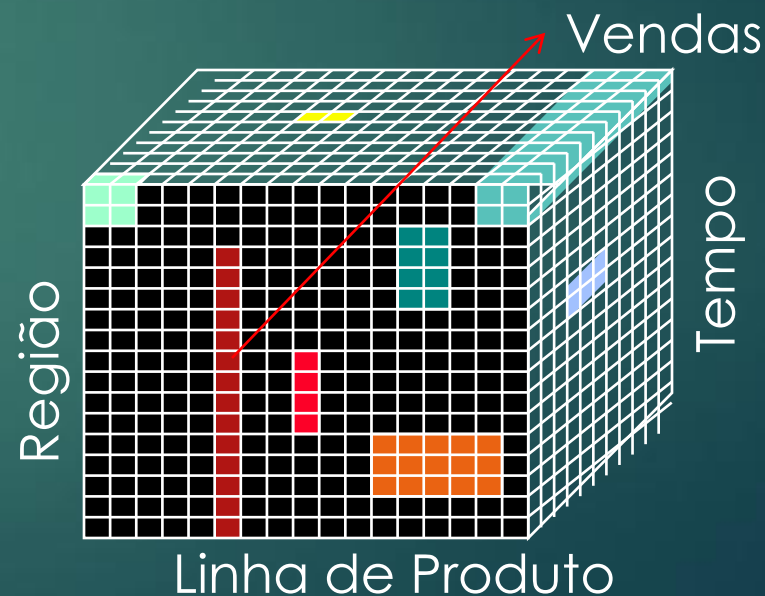
➤ DIMENSÃO

- Contém os descritores textuais do negócio.
- Exemplo : Tempo , Cliente, Produto, Tipo de Embalagem, Situação, etc.

➤ FATO

- Termo utilizado para a medição do negócio.
- Exemplo: quantidade de produtos vendidos, preço de compra, preço de venda, lucro, etc.

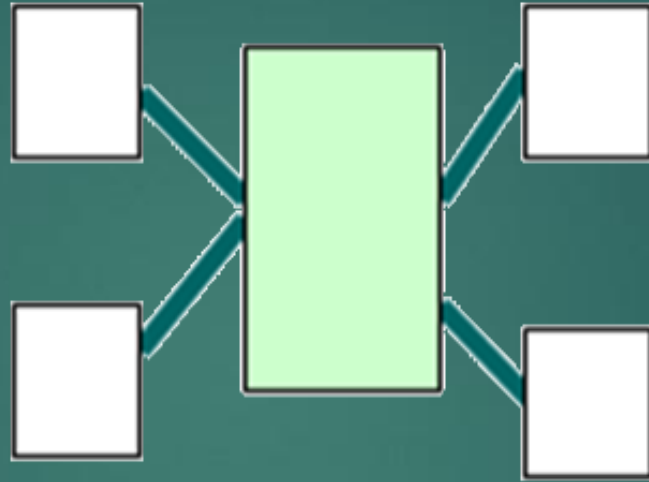
- Dimensões
 - Região
 - Linhas de Produto
 - Tempo (obrigatória)
- Fato
 - Vendas



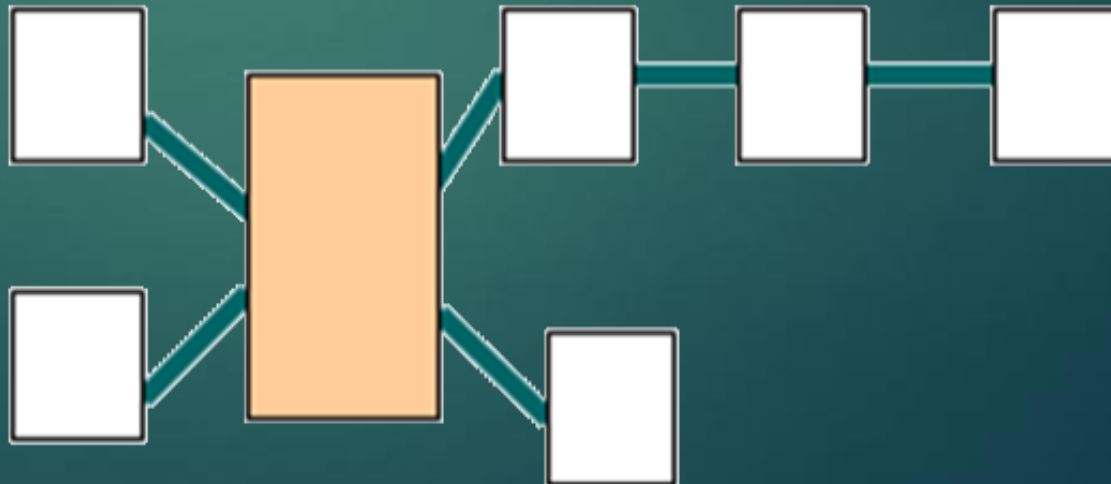
Modelagem Multidimensional

➤ MODELAGEM MULTIDIMENSIONAL

STAR SCHEMA

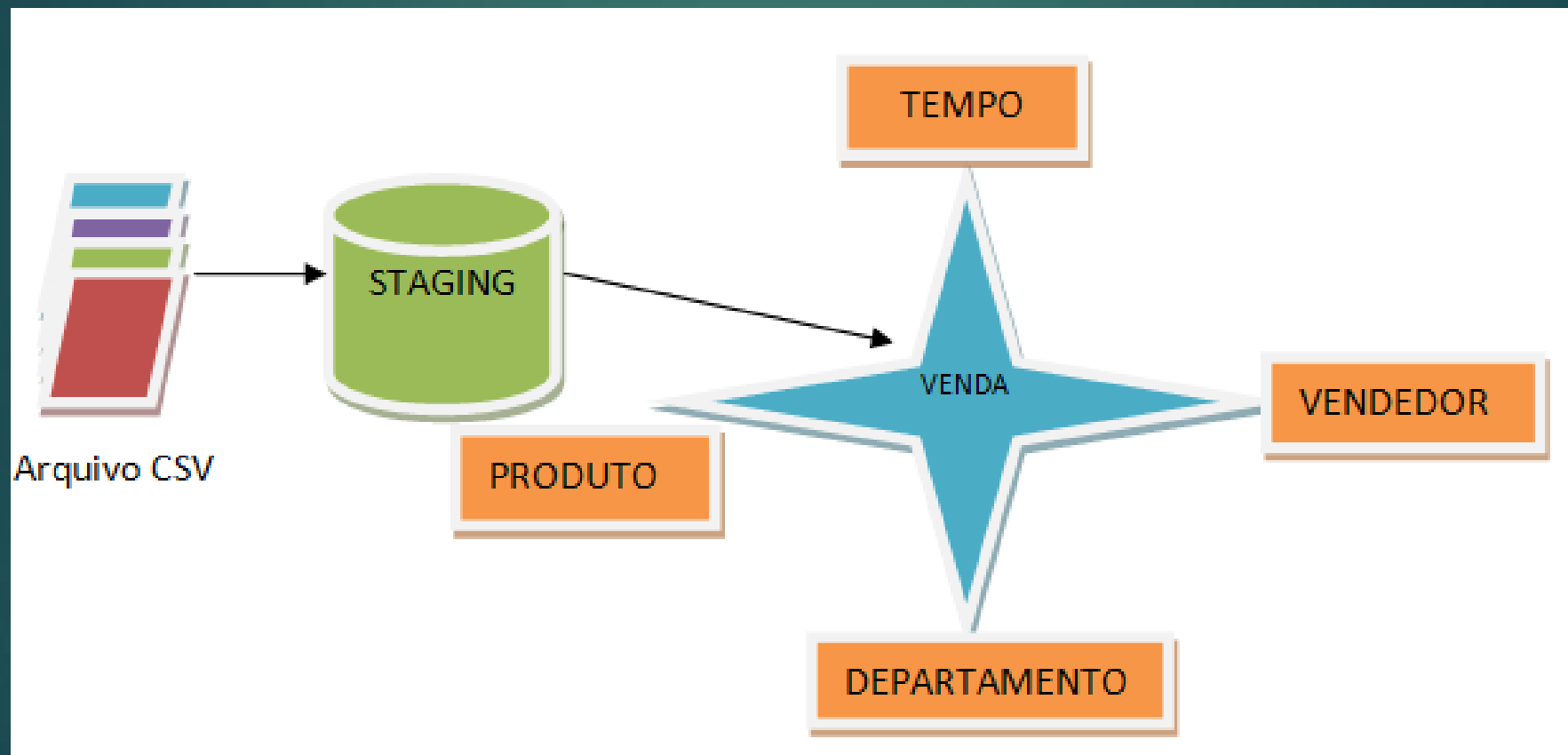


SNOW FLAKE



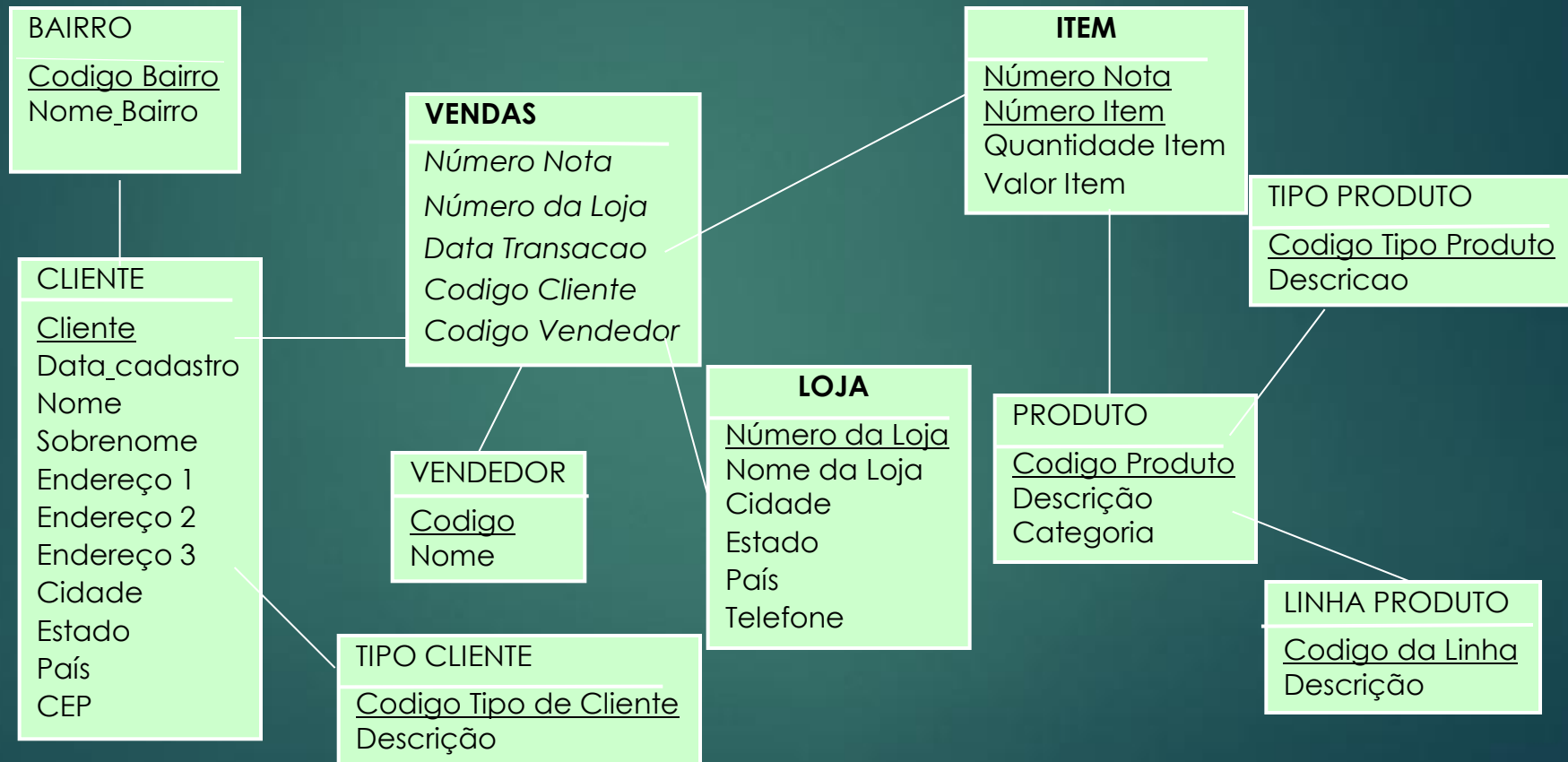
Exemplo de um Modelo Multidimensional

➤ DATA WAREHOUSE DE VENDAS



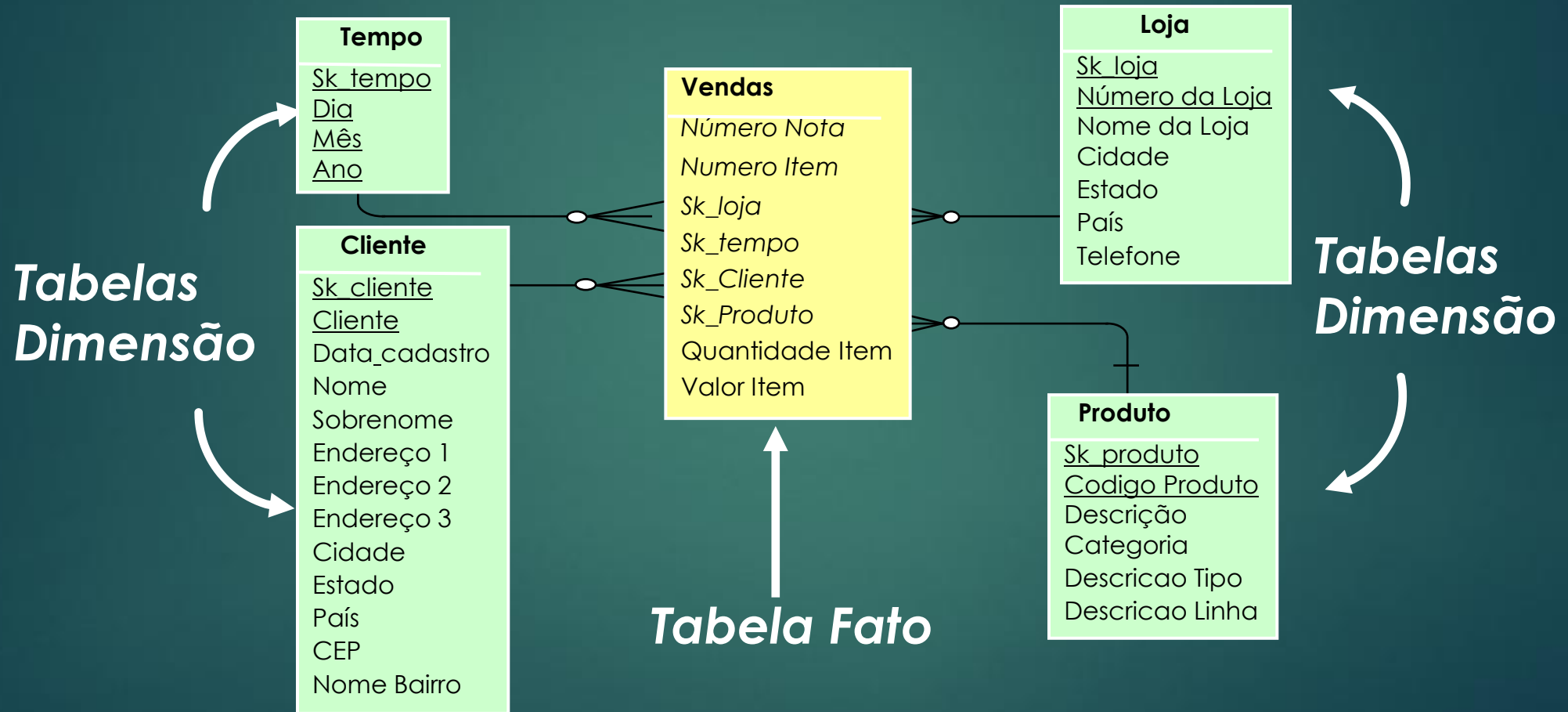
Modelo de Dados Operacional

➤ TABELAS OPERACIONAIS



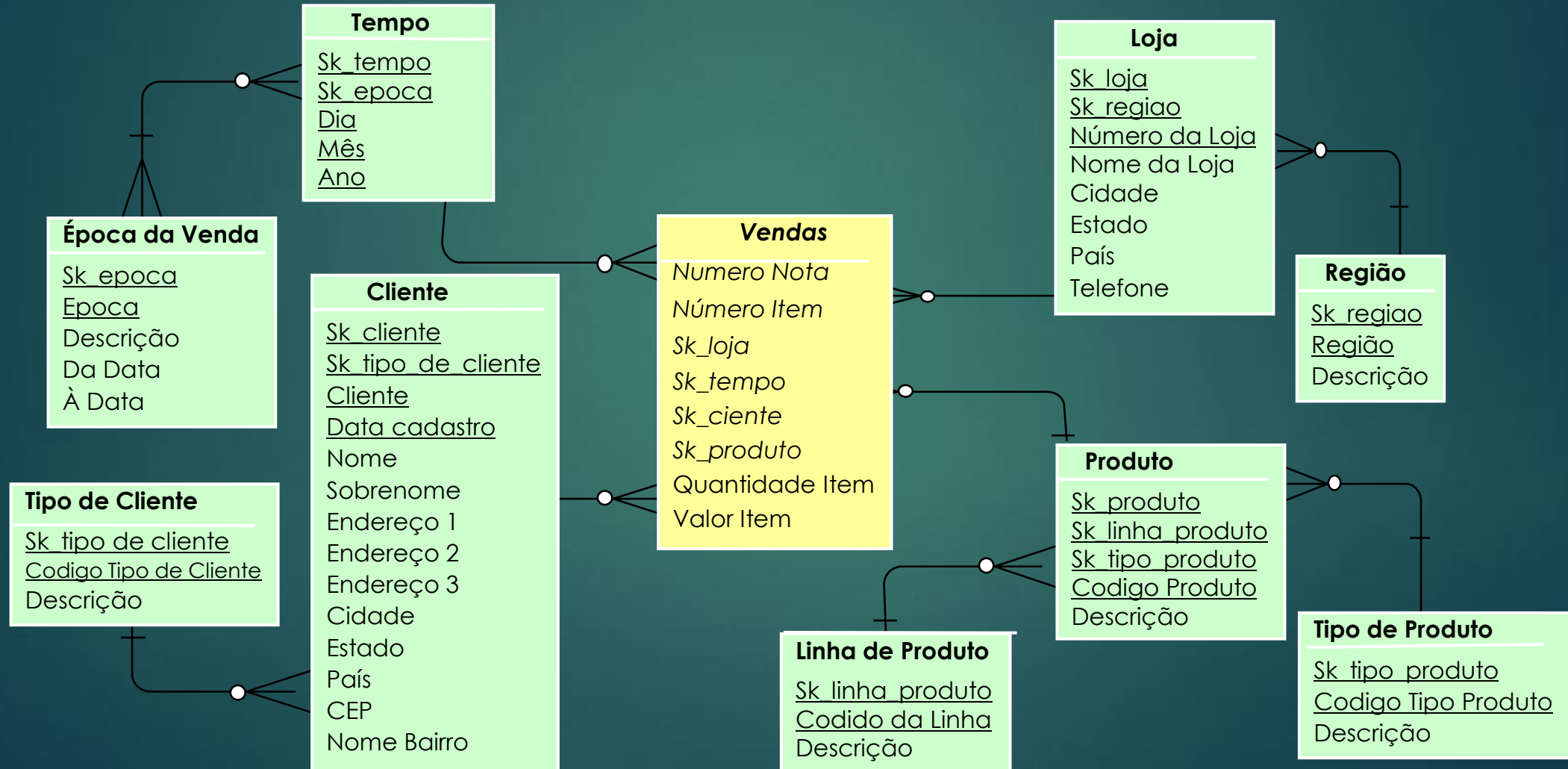
Modelo Multidimensional – Star Schema

➤ MODELO MULTIDIMENSIONAL – STAR SCHEMA



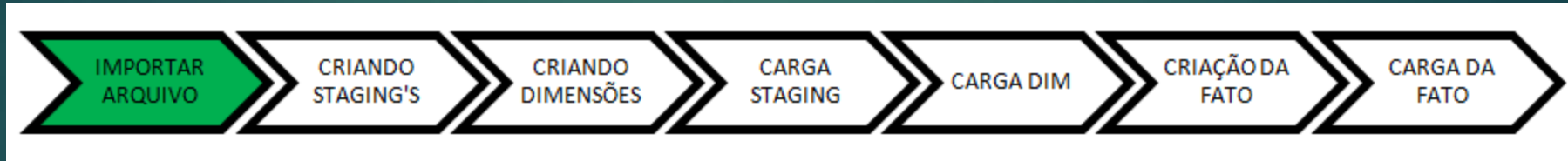
Modelo Multidimensional – Snow Flake

➤ MODELO MULTIDIMENSIONAL – SNOW FLAKE

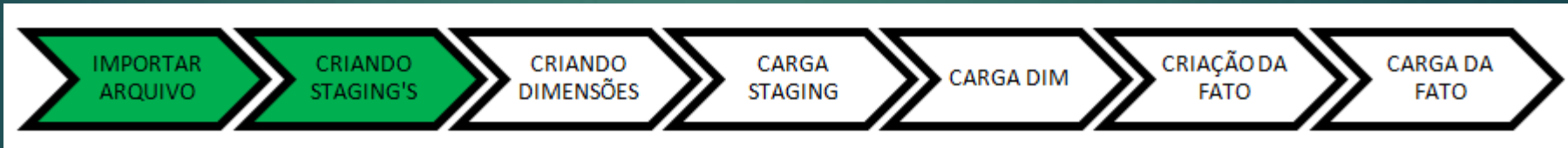


Construção do Data Warehouse

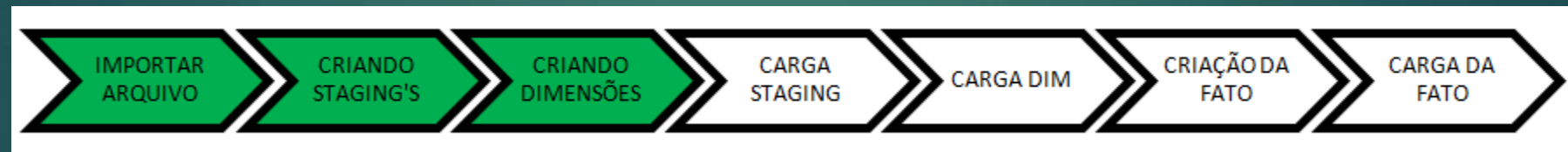
➤ Importando Arquivos



➤ Criando Staging



➤ Criando Dimensões



Construção do Data Warehouse

➤ Carga Staging



➤ Carga Dimensões



➤ Criação da Fato



➤ Carga da Fato



Documentos necessários

Finalidade: Coleta de dados com os gestores para a construção do BI.

Fatos		Diária
Dimensões		
Hóspede		✓
Tipo Quarto		✓
Código Tipo Quarto		
Tipo Quarto	HISTÓRICO	
Classe Quarto		✓
Tempo (Data Registro Primeira Diária)		✓

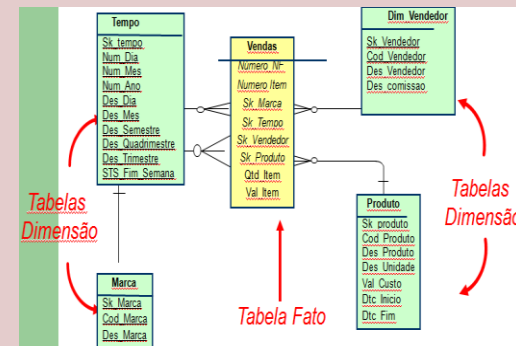
Matriz de Necessidades

Finalidade: Levantamento dos relacionamentos e objetos que armazenam os dados da empresa.

DIMENSÕES	ORIGEM	
	TABELA/VISÃO	CAMPO
Hóspede		
Nome Hóspede	HOSPEDE	NOM_HOSPEDE
Cidade Hóspede	CIDADE_ORIGEM	NOM_CIDADE
País Hóspede	PAIS_ORIGEM	NOM_PAIS
Aeroporto Hóspede	AEROPORTO_SAIDA	DES_AEROPORTO
Local Aeroporto Saída	AEROPORTO_SAIDA	NOM_LOCALIDADE
Código Hóspede	HOSPEDE	COD_HOSPEDE

Fonte de Dados

Finalidade: Modelo adequado para realizar as consultas nas bases que servirão ao BI



Modelagem Multidimensional

Levantamento dos Dados



Matriz de Necessidades

- Diagrama usado para documentar o Data Warehouse.
- Nas linhas coloca-se os descritores do negócio e nas colunas as métricas.
- Assinala-se um “X” quando a métrica for referente ao descritor.
- Útil para avaliar o impacto de alterações do seu DW.

Estudo de Caso

➤ Hotel Dallas

➤ Necessidades dos Gestores: Analisar Faturamento e Diárias do Hotel

DIÁRIAS

FATURAMENTO

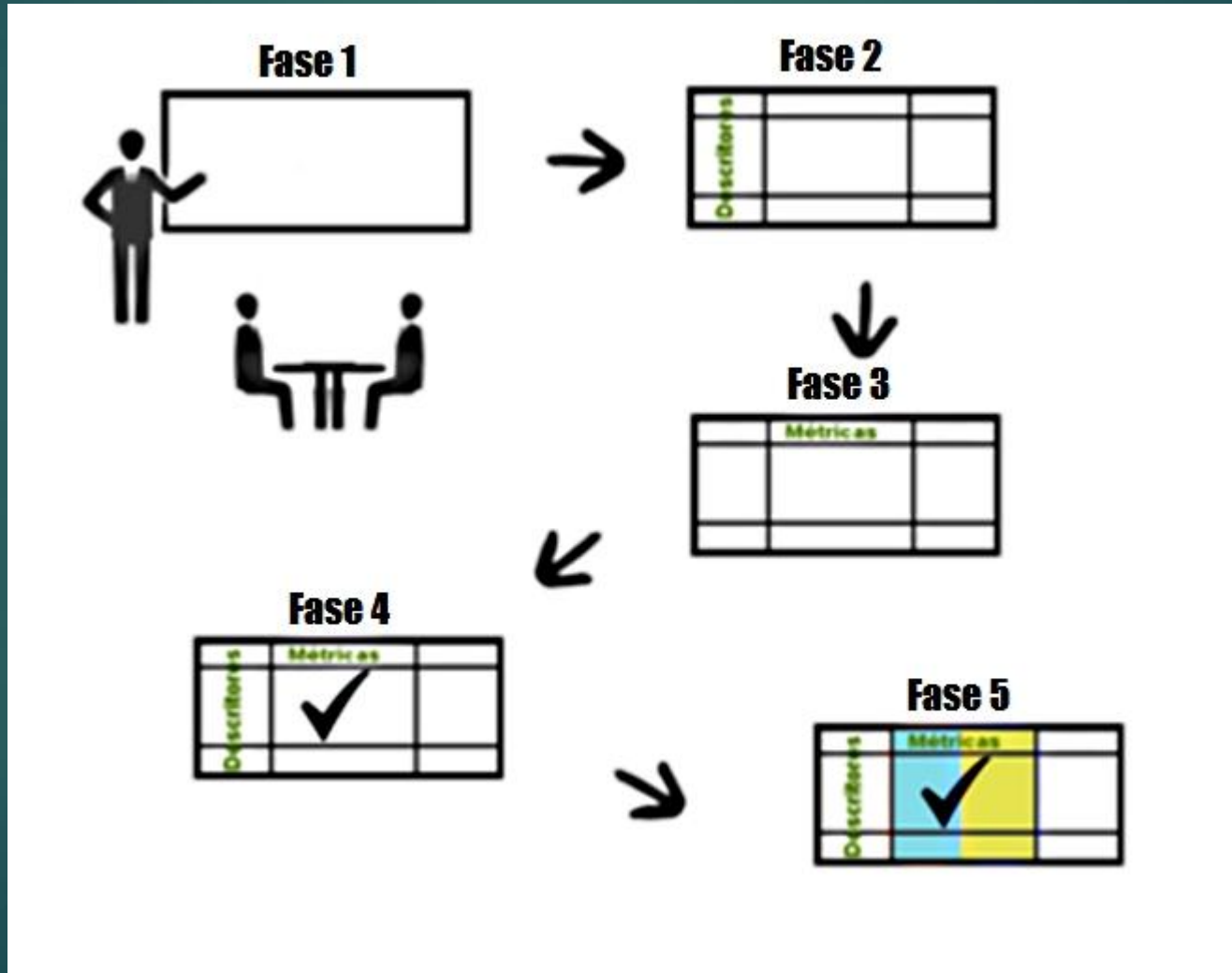


Levantamento dos dados com os Gestores

➤ Passo a Passo

- **Fase 1-** Convoque uma reunião com todos os gestores que utilizarão o BI.
- **Fase 2-** Com um projetor e uma planilha eletrônica faça uma abertura explicando como será preenchida a Matriz de Necessidades.
- **Fase 3-** Solicite que os gestores explanem sobre as principais necessidades na análise de dados do Hotel, indagando: quais métricas e descritores (valores e cruzamentos) vocês desejam analisar?
- **Fase 4-** Comece a desenvolver a Matriz de Necessidades, completando com os dados informados pelos gestores.
- **Fase 5-** Ao completar a Matriz de Necessidades com todas as informações discutidas na reunião, verifique se as perguntas dos gestores podem ser respondidas pelo documento criado.

Dinâmica de construção da Matriz de Necessidades



Matriz de Necessidades

- Perguntas dos gestores:
 - 1) Gostaria de analisar o **Total do Valor do Bar**, gasto pelo **hóspede** (nome e cidade de origem).
 - 2) Quero analisar o **Total de Diárias**, gasto pelo **hóspede** da **classe econômica**.
 - 3) Identifique os **hóspedes** do hotel Dallas, que vieram do **país 'Brasil'**.
 - 4) Solicito analisar a **Média do valor da hospedagem** de todos os **hóspedes** que são oriundos da **cidade de 'Salvador'** e que saíram do **aeroporto de 'Porto Seguro'**.
 - 5) Gostaria de realizar uma agregação de todos os valores passados ao governo da região com o **valor turismo**, separados por **país de origem** do hóspede e **cidade de origem** do hóspede.
 - 6) Gostaria que fosse exibido uma análise dos **dias da semana** que as pessoas mais se hospedam (**diária**) no Hotel Dallas, separados por **mês, bimestre e semestre**.
 - 7) Quero criar uma fórmula **Receita = valor da hospedagem + valor do bar – valor do turismo** por nome de **hóspede**.

Matriz de Necessidades

Fatos		Diária	Faturamento		
Dimensões		Valor Diária	Valor Hospedagem	Valor Bar	Valor Turismo
Hóspede		✓	✓	✓	✓
Código Hóspede					
Nome Hóspede	HISTÓRICO				
Cidade Hóspede	HISTÓRICO				
País Hóspede	HISTÓRICO				
Aeroporto Hóspede	HISTÓRICO				
Local Aeroporto Saída	HISTÓRICO				
Tipo Quarto		✓			
Código Tipo Quarto					
Tipo Quarto	HISTÓRICO				
Classe Quarto		✓			
Código Classe Quarto					
Classe Quarto	HISTÓRICO				
Tempo (Data Registro Primeira Diária, Data Entrada, Data Saída)		✓	✓	✓	✓
Ano					
Mês					
Dia					
Dia da Semana					
Bimestre					
Trimestre					
Semestre					

Fonte de Dados



Fonte de Dados

- Diagrama usado para documentar os campos levantados na Matriz de Necessidades oriundas das bases transacionais da organização.
- Nas linhas coloca-se os descritores do negócio e os relacionamentos que existem no modelo, as métricas são identificadas e destacadas.
- Você deve ficar atento as chaves primárias (PK) para mapear os dados corretamente.
- Útil para avaliar quantos descritores e métricas você terá no seu DW.

Modelo de Dados

Hospede

Cod_hospede
Nom_hospede
Des_endereco
Cod_cidade
Cod_aeroporto
Dtc_cadastro

Cidade Origem

Cod_cidade
Nom_cidade
Cod_pais

Aeroporto Saída

Cod_aeroporto
Des_aeroporto
Nom_Localidade

Tipo classe quarto

Cod_classe_quarto
Des_classe_quarto

Diaria

Cod_diaria
Cod_hospede
Cod_tipo_quarto
Cod_classe_quarto
Val_diaria
Qtd_diaria
Qtd_pessoas
Dtc_registro_primeira_diaria

Pais Origem

Cod_pais
Nom_pais

Tipo quarto

Cod_Tipo_quarto
Des_Tipo_quarto

Faturamento

Cod_faturamento
Cod_hospede
Dtc_entrada
Dtc_saida
Val_hospedagem
Val_turismo
Val_bar

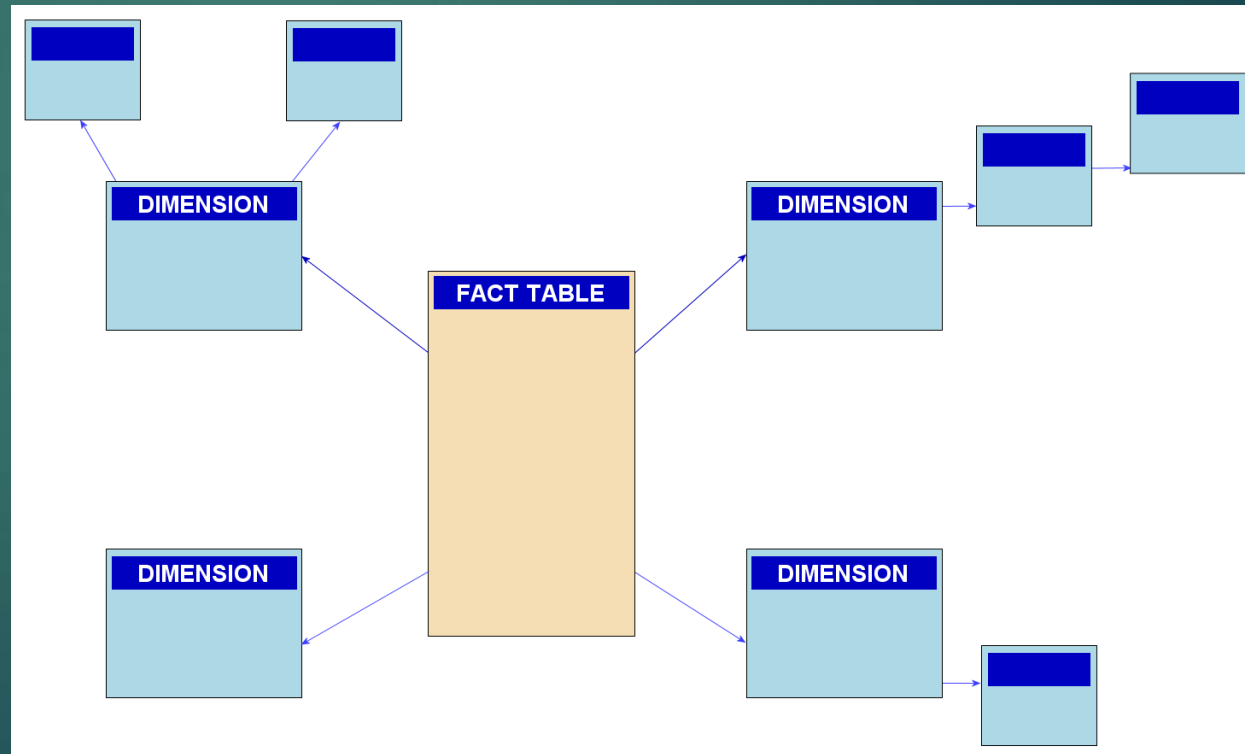
Fonte de Dados (Dimensões)

DIMENSÕES	ORIGEM			OBSERVAÇÃO
	TABELA/VISÃO	CAMPO	RELACIONAMENTO	
Hóspede				
Nome Hóspede	HOSPEDE	NOM_HOSPEDE		
Cidade Hóspede	CIDADE_ORIGEM	NOM_CIDADE	HOSPEDE.COD_CIDADE=CIDADE_ORIGEM.COD_CIDADE	
País Hóspede	PAIS_ORIGEM	NOM_PAIS	HOSPEDE.COD_CIDADE=CIDADE_ORIGEM.COD_CIDADE CIDADE_ORIGEM.COD_PAIS=PAIS_ORIGEM.COD_PAIS	
Aeroporto Hóspede	AEROPORTO_SAIDA	DES_AEROPORTO	HOSPEDE.COD_AEROPORTO=AEROPORTO_SAIDA.COD_AEROPORTO	
Local Aeroporto Saída	AEROPORTO_SAIDA	NOM_LOCALIDADE		
Código Hóspede	HOSPEDE	COD_HOSPEDE		
Tipo Quarto				
Tipo Quarto	TIPO_QUARTO	DES_TIPO_QUARTO		
Código Tipo Quarto	TIPO_QUARTO	COD_TIPO_QUARTO		
Classe Quarto				
Classe Quarto	TIPO_CLASSE_QUARTO	DES_CLASSE_QUARTO		
Código Classe Quarto	TIPO_CLASSE_QUARTO	COD_CLASSE_QUARTO		
Tempo (Data Registro Primeira Diária)				
Ano				A dimensão TEMPO não tem origem.
Mês				
Dia				
Data				

Fonte de Dados (Fatos)

FATOS				
Diária	ORIGEM			OBSERVAÇÃO
	TABELA/VISÃO	CAMPO	RELACIONAMENTO	
Métricas				
Valor Diária	DIARIA	VAL_DIARIA		Se o retorno das métricas for NULO,
Quantidade de Diária		QTD_DIARIA		
Quantidade de Pessoa		QTD_PESSOAS		
Dimensões Diária				
Hóspede	DIM_HOSPEDE	SK_HOSPEDE	DIARIA.COD_HOSPEDE=DIM_HOSPEDE.COD_HOSPEDE	
Tipo Quarto	DIM_TIPO_QUARTO	SK_TIPO_QUARTO	DIARIA.COD_TIPO_QUARTO=DIM_TIPO_QUARTO.COD_TIPO_QUARTO	
Classe Quarto	DIM_CLASSE_QUARTO	SK_CLASSE_QUARTO	DIARIA.COD_CLASSE_QUARTO=DIM_CLASSE_QUARTO.COD_CLASSE_QUARTO	
Tempo	DIM_TEMPO	SK_TEMPO	DIARIA.DTC_REGISTRO_PRIMEIRA_DIARIA=DIM_TEMPO.DATA	
Faturamento	ORIGEM			OBSERVAÇÃO
	TABELA/VISÃO	CAMPO	RELACIONAMENTO	
Métricas				
Valor da Hospedagem	FATURAMENTO	VAL_HOSPEDAGEM		Se o retorno das métricas for NULO,
Valor do Turismo		VAL_TURISMO		
Valor do Bar		VAL_BAR		
Dimensões Faturamento				
Hóspede	DIM_HOSPEDE	SK_HOSPEDE	FATURAMENTO.COD_HOSPEDE=DIM_HOSPEDE.COD_HOSPEDE	
Tempo	DIM_TEMPO	SK_TEMPO	FATURAMENTO.DTC_ENTRADA=DIM_TEMPO.DATA	
Tempo	DIM_TEMPO	SK_TEMPO	FATURAMENTO.DTC_SAIDA=DIM_TEMPO.DATA	

Modelagem Multidimensional



Agenda

- Objetivo
 - Apresentar os principais assuntos na construção de um projeto de Data Warehouse
 - ▶ Dimensões e Fatos
 - ▶ Modelagem Multidimensional
 - ▶ Fundamentos da Modelagem Multidimensional
 - ▶ Desnormalização: anti-forma normal.
 - ▶ Tipos de Dimensões
 - ▶ Dimensões: Chaves Artificiais e Histórico
 - ▶ Dimensões Hierárquicas
 - ▶ Tipos de Métricas
 - ▶ Tempo
 - ▶ Granularidade

Dimensão e Fato

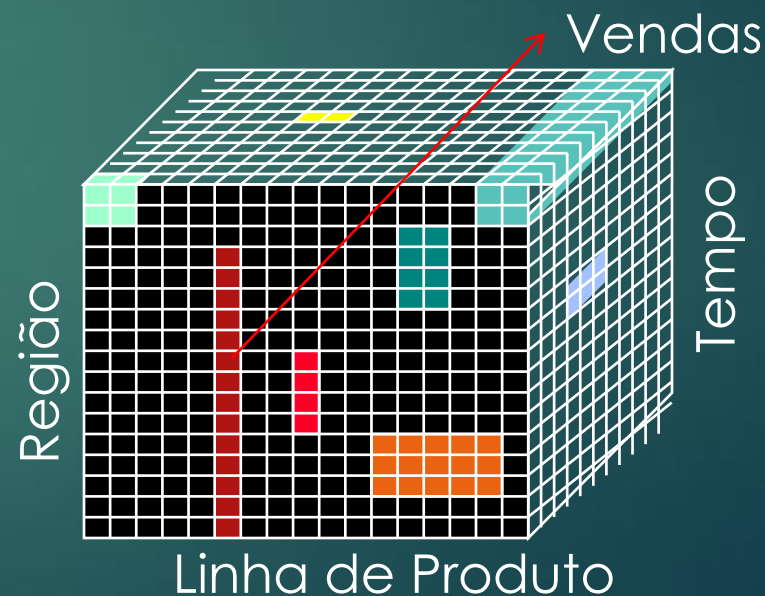
➤ DIMENSÃO

- Contém os descritores textuais do negócio.
- Exemplo : Tempo, Cliente, Produto, Tipo de Embalagem, Situação, etc.

➤ FATO

- Termo utilizado para a medição do negócio.
- Exemplo: quantidade de produtos vendidos, preço de compra, preço de venda, lucro, etc.

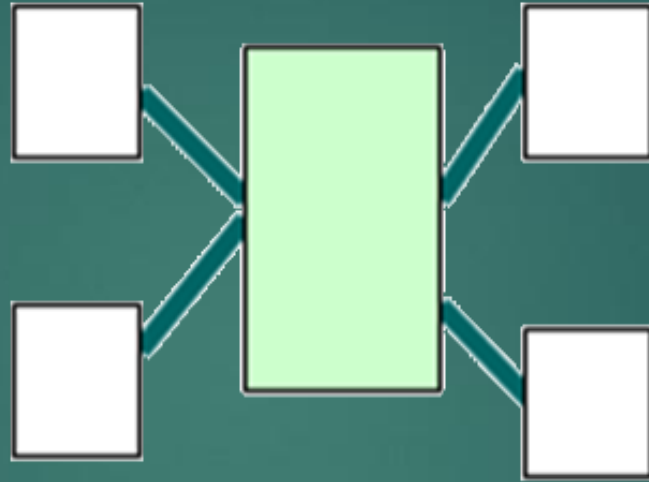
- Dimensões
 - Região
 - Linhas de Produto
 - Tempo (obrigatória)
- Fato
 - Vendas



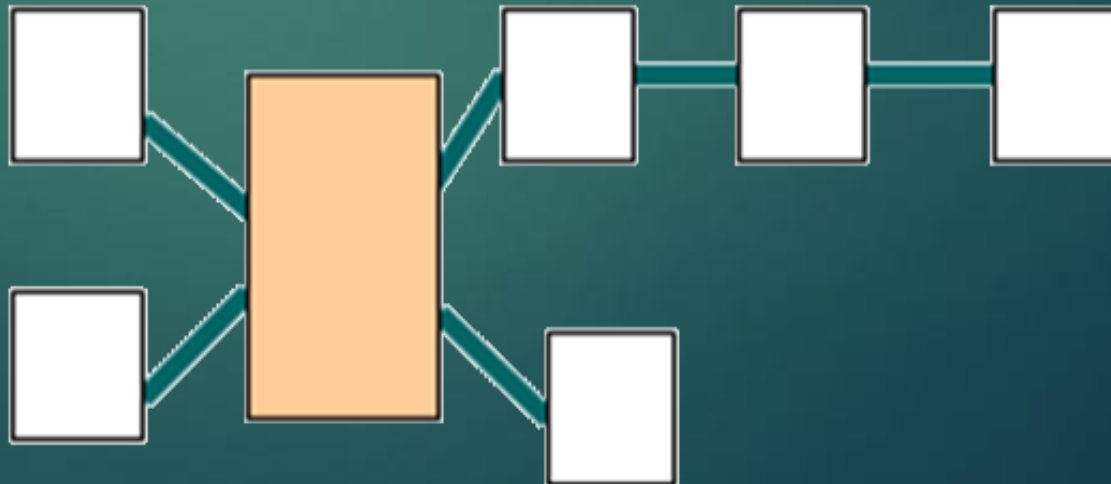
Modelagem Multidimensional

➤ MODELAGEM MULTIDIMENSIONAL

STAR SCHEMA

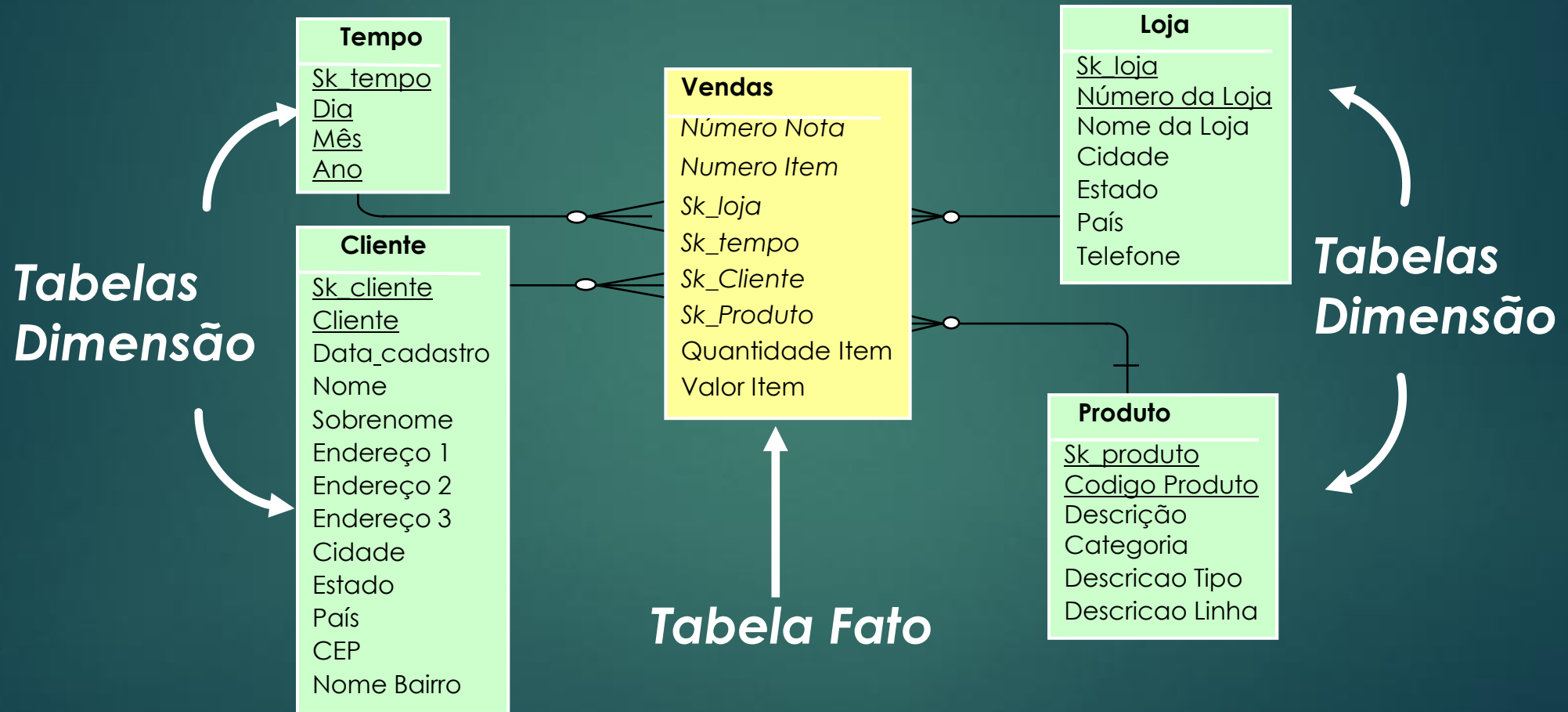


SNOW FLAKE



Modelo Multidimensional – Star Schema

➤ MODELO MULTIDIMENSIONAL – STAR SCHEMA



Desnormalização

- ▶ Só a 1ºFN deve ser respeitada. As demais FN obrigatoriamente devem ser ignoradas.
- ▶ Vantagens: Excelente tempo de 'query response'.
- ▶ Exemplo:

CLIENTE

Cod_cliente

@Cod_cliente

Nom_cliente

Nom_Cliente

Cod_tipo_cliente

Cod_tipo_cliente

TIPO CLIENTE

Des_tipo_cliente

Cod_tipo_cliente

Cod_condicao_tipo_cliente

Des_tipo_cliente

Des_condicao_tipo_cliente

Cod_condicao_tipo_cliente

Considerações sobre espaço

A desnormalização simplifica o modelo, otimizando as consultas geradas pela ferramentas OLAP. Porém, causa excesso de uso em disco.

Exemplo: Clientes do Bank Money

	<u>Normalizado</u>	<u>Desnormalizado</u>
Tamanho	400Mb	2Gb
Qtde Campos	~70	~180
Query Estatística	> 10hs	<15 min

Chaves Artificiais

- ▶ Permite que o controle de histórico dos dados seja facilmente implementado.
- ▶ Gera independência de relacionamento com outras tabelas.
- ▶ Relacionam as Dimensões e Fatos.
- ▶ Devem ser apenas números e não carregar em si nenhum significado.
 - ▶ Exemplo: 132453
 - ▶ Contra Exemplo: Jan/99 , IFZ01

Tipos de Dimensões

- ▶ Slow Changing Dimensions
- ▶ Fast Changing Monster Dimensions
- ▶ Dimensões Degeneradas

Slow Changing Dimensions

- ▶ **Tipo 1: Sobrescrever os Dados**

- ▶ O novo registro substitui o registro original. Só existe um registro no banco de dados - os dados atuais.
- ▶ Não mantém histórico.

Tipo1 : Sobrescrever os Dados

- Um exemplo seria de uma tabela de banco de dados que mantém as informações do fornecedor.

Supplier_key	Supplier_Name	Supplier_State
001	Phlogistical Sociedade de Abastecimento	CA

Supplier_key	Supplier_Name	Supplier_State
001	Phlogistical Sociedade de Abastecimento	IL

Slow Changing Dimensions

▶ **Tipo 2: Controle de Versão**

- ▶ Mantém histórico
- ▶ Um novo registro é adicionado na tabela de Dimensão. Dois registros existentes no banco de dados - os dados atuais e dados da história anterior.
- ▶ É recomendável para a maioria dos casos.

Tipo 2 : Controle de Versão

- ▶ Exemplo seria o de uma tabela de banco de dados que mantém as informações do fornecedor.

Supplier_key	Supplier_Code	Supplier_Name	Supplier_State	Data_inicial	Data_final
001	ABC	Phlogistical Sociedade de Abastecimento	CA	01 de janeiro-2000	21-Dez-2004
002	ABC	Phlogistical Sociedade de Abastecimento	IL	22-Dez-2004	

Exemplo : DIM_HOSPEDE

Seq_cliente

Cod_hospede

Nom_hospede

Des_endereco

Dtc_inicio

Dtc_fim

Sts_corrente

***Em branco: campos de controle
utilizados pela modelagem
multidimensional***

***Em verde: campos do sistema
fonte.***

Exemplo : DIM_HOSPEDE

Seq_cliente : 1
Cod_cliente : X1
Nom_cliente : Antônio dos Santos
Des_Endereco : Rua das Margaridas
Dtc_inicio : 01/03/2010
Dtc_fim : <null>
Sts_corrente : Ativo

E se o cliente
vier a trocar de
endereço?

Exemplo : DIM_HOSPEDE

Seq_cliente : 3435

Cod_cliente : X1

Nom_cliente : Antônio dos Santos

Des_Endereco : Rua das Rosas

Dtc_inicio : 01/03/2012

Dtc_fim : <null>

Sts_corrente : Ativo

Um **novo** registro é
inserido na base de
dados.

E o registro anterior?

Exemplo : DIM_HOSPEDE

Seq_cliente : 1

Cod_hospede : X1

Nom_hospede : Antonio dos Santos

Des_Endereco : Rua das Margaridas

Dtc_inicio : 01/03/2010

Dtc_fim : 01/03/2012

Sts_corrente : Inativo

Ambos os registros
estão no banco de
dados.

Considerações

- ▶ O uso de chaves artificiais permite o controle do histórico e facilita o relacionamento com as tabelas Fato.
- ▶ Os campos de controle de data (Dtc_inicio e Dtc_fim) permitem saber quando a informação refletia o operacional.
- ▶ O campo “*Sts_corrente*” permite a fácil identificação do registro que contém os dados mais recentes.

Fast Changing Dimensions

- ▶ Algumas dimensões que possuem grande volume de registros e muitos campos, crescem rapidamente, explodindo o espaço físico de armazenamento.
- ▶ Solução: Colocar campos que trocam de valores mais rapidamente em outra tabela, sem alterar a versão do registro.

Dimensões Degeneradas

- ▶ É caracterizada por não ter a sua própria Dimensão e sim por estar presente na tabela Fato.
- ▶ Exemplo: Considere que uma tabela Fato com os itens das notas fiscais de uma empresa varejista. O que fazer com o número da nota fiscal? Ele em si, não representa nada, apenas serve para agrupar os itens de uma mesma compra. Não existe fisicamente uma Dimensão nota fiscal, embora exista uma coluna na tabela Fato com o número da nota propriamente dito.

Dimensão Ponte (Bridge Table)

- Uma tabela com chave composta capturando um relacionamento muitos-para-muitos que não pode ser acomodado pela granularidade natural de uma tabela Fato ou de Dimensão. Serve como uma ponte entre a tabela Fato e de Dimensão.



- Outros exemplos de dimensões multivaloradas: titulares de conta bancária.

Dimensões Hierárquicas

- ▶ Exemplo : País → Região → UF → Cidade
- ▶ As tabelas de País, Região, UF e Cidade são armazenadas separadamente no sistema fonte (normalizado). No Data Warehouse, elas compõem uma única tabela, a Dimensão Geografia. Cada nível da hierarquia deve ser representado individualmente.
- Usada para Permitir Drill:
 - **Drill Down:** Detalha a informação
 - **Drill Up:** Sumariza a informação
 - **Drill Across:** Muda de Dimensão, mantém Fato.
 - **Drill Through:** Vê registros do ambiente transacional que originaram aquela ocorrência

Drilling



Tipos de Métricas

Aditivas: Podem ser utilizadas para sumarizações.

Ex.: Valor de Venda

Semi-Aditivas: Podem ser utilizadas para sumarizações, com ressalvas.

Ex.: Saldo Bancário – Faz sentido somar o seu saldo caso ele tenha conta em mais de um banco, mas não faz sentido somar seu saldo todos os dias de uma mesma conta.

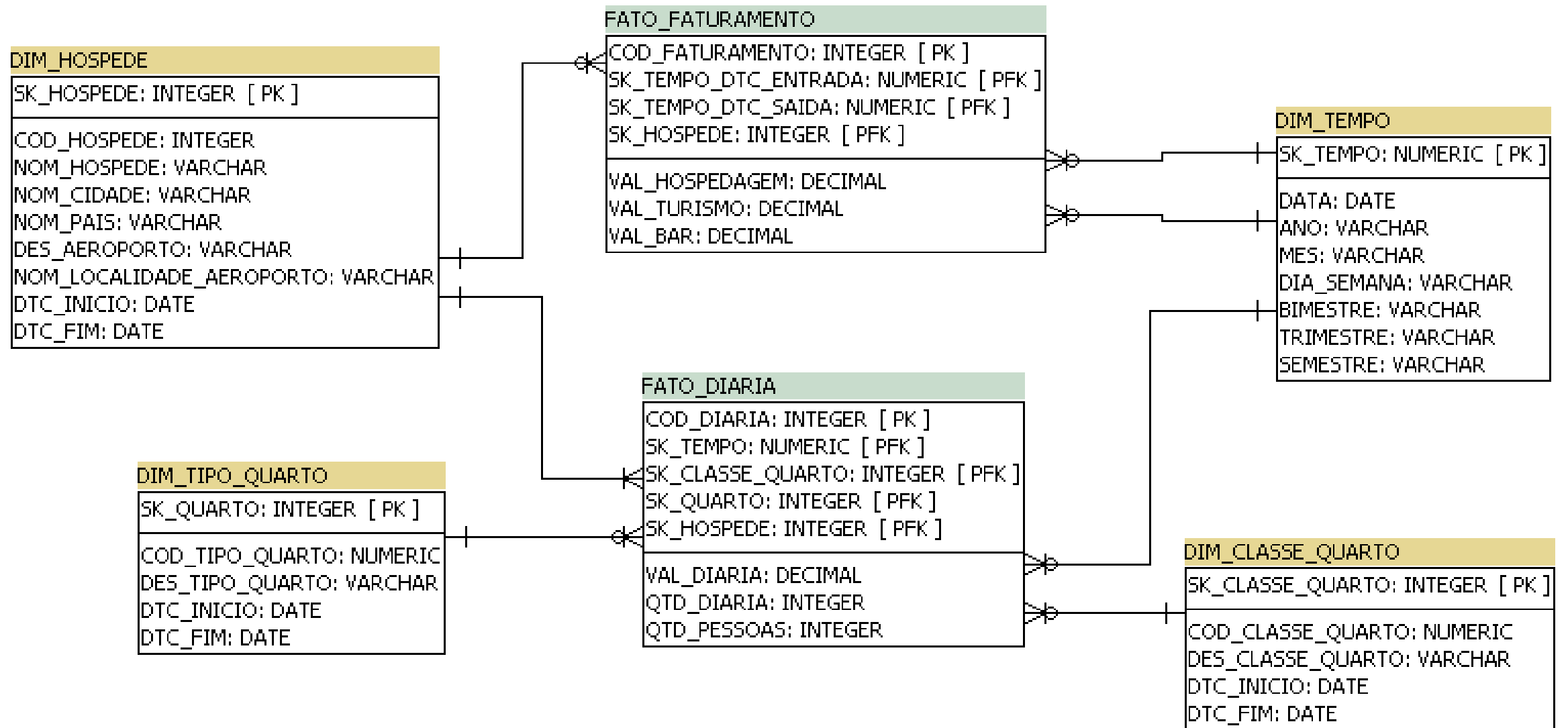
Não-Aditivas: Não podem ser utilizadas para sumarização.

Ex.: % margem de lucro

Grão

- ▶ Conceito que identifica a unidade de medida das métricas.
- ▶ Nível de detalhe dos dados.
- ▶ Menor Grão: Mais detalhe -> Mais dados -> Análise mais longa -> Informação mais detalhada.
- ▶ Maior Grão: Menos detalhe -> Menos dados -> Análise mais rápida -> Informação menos detalhada.

Modelo Multidimensional: Hotel Dallas



Processo completo

Finalidade: Coleta de dados com os gestores para a construção do BI.

Fatos		Diária
Dimensões		
Hóspede		✓
Tipo Quarto		✓
Código Tipo Quarto		
Tipo Quarto	HISTÓRICO	
Classe Quarto		✓
Tempo (Data Registro Primeira Diária)		✓

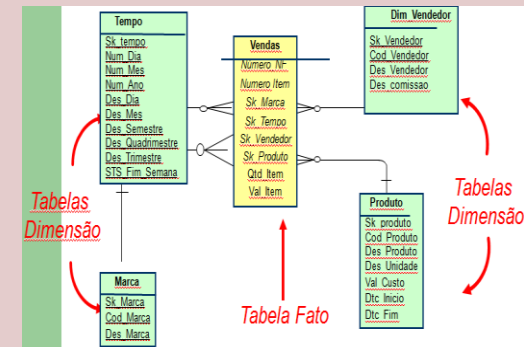
Matriz de
Necessidades

Finalidade: Levantamento dos relacionamentos e objetos que armazenam os dados da empresa.

DIMENSÕES	ORIGEM	
	TABELA/VISÃO	CAMPO
Hóspede		
Nome Hóspede	HOSPEDE	NOM_HOSPEDE
Cidade Hóspede	CIDADE_ORIGEM	NOM_CIDADE
País Hóspede	PAIS_ORIGEM	NOM_PAIS
Aeroporto Hóspede	AEROPORTO_SAIDA	DES_AEROPORTO
Local Aeroporto Saída	AEROPORTO_SAIDA	NOM_LOCALIDADE
Código Hóspede	HOSPEDE	COD_HOSPEDE

Fonte de Dados

Finalidade: Modelo adequado para realizar as consultas nas bases que servirão ao BI



Modelagem
Multidimensional

Obrigado!
grimaldo_lopes@hotmail.com