

# Arquitetura de Data Warehouse e Data Marts

## Aula 4: Projeto Físico do Data Warehouse

### Apresentação

---

Nesta aula serão apresentados pontos importantes para a aplicação do projeto físico do ciclo de vida do desenvolvimento do Data Warehouse/Data Mart (DW/DM). Será apresentada a área de armazenamento de dados e a importância desta estar preparada para o volume de dados que será tratado no DW/DM e a implementação do modelo no Sistema Gerenciador de Banco de Dados (SGBD) escolhido.

# Objetivos

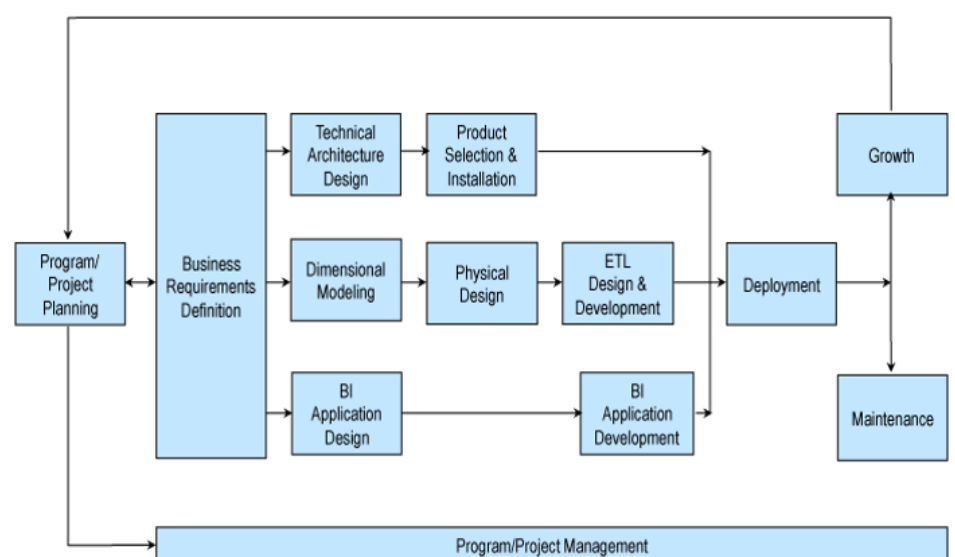
- Examinar padrões e restrições envolvidos no projeto físico do DW/DM;
- Explicar as necessidades existentes para a implantação do modelo de dados dimensional;
- Descrever a implementação do modelo de dados dimensional físico no SGBD.

## O Projeto Físico

Na aula passada conhecemos a Modelagem de Dados Dimensional, os esquemas Estrela e Floco-de-Neve e as tabelas Dimensão e Fato. Nesta aula daremos continuidade à trilha de dados contida no ciclo de vida de projetos de DW/DM (Data Warehouse/Data Mart). A trilha de dados, ilustrada pela Figura 1, dedica-se ao tratamento dos dados e encadeia as fases Modelagem Dimensional, a definição do projeto físico e a especificação de ETL.

Logo, como você já possui o conhecimento sobre a modelagem dimensional, agora verá o próximo passo, que é a definição do modelo de dados dimensional físico e o que é necessário para a implementação do modelo de dados dimensional de modo que ele responda às consultas do usuário com um bom desempenho.

Ciclo de Vida de um Projeto de Data Warehouse por Kimball



Fonte: Kimball (2013).

A implementação física do modelo de dados dimensional considera o Sistema Gerenciador de Banco de Dados (SGBD) escolhido para o projeto e alguns outros pontos que serão detalhados a seguir.

O modelo de dados dimensional físico parte do modelo lógico criado anteriormente e une os padrões estabelecidos, as regras de negócio, as características do SGBD e o envolvimento de alguns especialistas que darão suporte e irão aplicar soluções para que a implementação do modelo seja feita com sucesso, buscando um bom desempenho nas consultas analíticas.

**Para que o projeto do DW/DM continue seu desenvolvimento, o Modelo de Dados Dimensional lógico desenhado precisa ser transformado em um ambiente físico em que os dados possam ser acomodados. Nesse**

**momento, as características do SGBD devem ser observadas, pois o projeto físico utiliza essas informações para sua construção e isso pode variar entre os SGBDs.**

As informações de restrições e valores nulos devem ser avaliadas com atenção para que sejam aplicadas corretamente no projeto físico. Outra questão importante são os padrões utilizados para os nomes das tabelas, colunas, índices etc.

O padrão da nomenclatura deve ser estabelecido antes mesmo de iniciar o desenho do modelo de dados dimensional físico para que todos os elementos sigam corretamente a definição. Não há um padrão obrigatório a ser usado e, normalmente, utiliza-se o padrão especificado pela organização.

O projeto físico envolve, além das tabelas do modelo de dados dimensional, algumas tabelas de suporte ao processo de ETL (Extract, Transform and Load) que veremos mais à frente. Essas tabelas são chamadas de tabelas temporárias e são a porta de entrada para a *staging area* ou área de manobras/preparação dos dados.

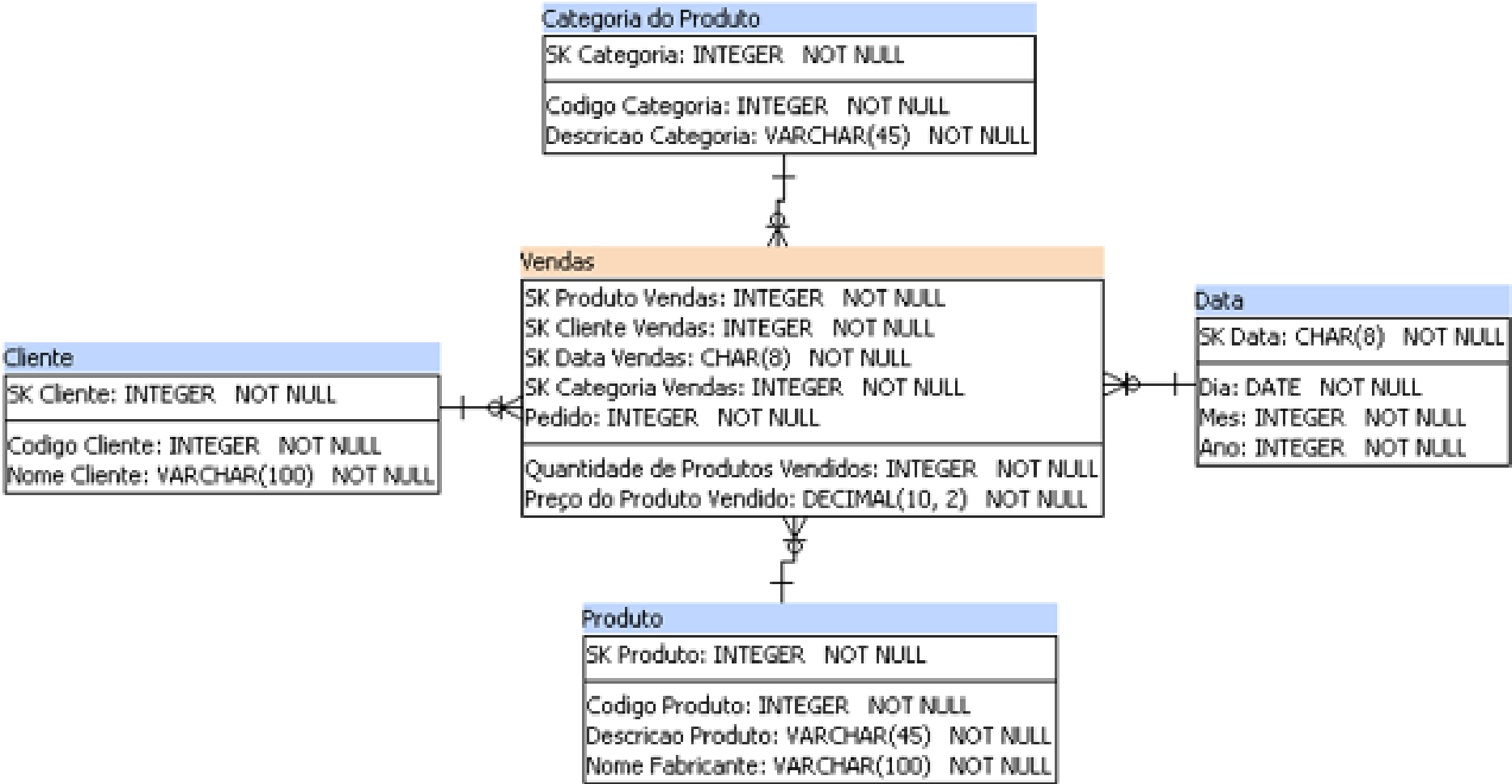
Saiba mais

A *staging area* é o conjunto de espaço e elementos que fica entre o sistema de origem e a área de apresentação dos dados. As tabelas temporárias recebem os dados extraídos do sistema origem para que eles possam ser tratados no processo ETL e, somente após os tratamentos, serem acomodados nas tabelas Dimensão e Fato.

Além das tabelas temporárias, outras tabelas de suporte à segurança, tabelas de “De para” de dados e tabelas de metadados podem ser construídas. A criação dessas tabelas depende da necessidade do projeto que está sendo desenvolvido.

A construção dos índices, partições e tabelas agregadas também é avaliada nessa etapa. Esses recursos melhoram o desempenho das consultas que serão submetidas no ambiente analítico e são muito importantes para o DW/DM que trabalha com um volume de dados muito grande.

A seguir, vamos explorar melhor esses pontos e aplicá-los ao projeto físico do DW Supermercado. Vamos utilizar o modelo de dados dimensional de Vendas a Varejo construído na aula anterior.



**Atenção!** Aqui existe uma videoaula, acesso pelo conteúdo online

# Padrão para a nomenclatura dos elementos do Modelo de Dados Dimensional Físico

Vamos adotar a seguinte nomenclatura para o desenvolvimento dos elementos do modelo de dados dimensional:

## Tabela Dimensão

Os nomes das tabelas dimensões serão iniciadas com dim\_;

## Tabela Fato

Os nomes das tabelas Fatos serão iniciadas com ft\_;

## Tabela Temporária

Os nomes das tabelas temporárias serão iniciados com tmp\_;

## Coluna de chave

Os nomes das colunas que representam identificadores serão iniciados com sk\_;

## Coluna de código

Os nomes das colunas que representam códigos serão iniciados com cd\_;

## Coluna numérica

Os nomes das colunas que representam dados numéricos serão iniciados com num\_;

## Coluna de descrição

Os nomes das colunas que representam descrições serão iniciados com ds\_;

## Coluna de nomes

Os nomes das colunas que representam nomes serão iniciados com nm\_;

## Coluna de data

Os nomes das colunas que representam datas serão iniciados com dt\_;

## Coluna de valor

Os nomes das colunas que representam os valores serão iniciados com vl\_.

Algumas ferramentas são sensíveis a letras maiúsculas e minúsculas. Assim, para minimizar problemas futuros é recomendado definir se os nomes serão criados todos em caixa alta ou em minúsculas.

## Tabela Dimensão

A Dimensão Produto contém os dados código do produto e descrição do produto. Apesar de a informação sobre o Fabricante do Produto estar armazenada na tabela Fabricante no sistema origem, ela foi adicionada à dimensão Produto como um atributo. A dimensão receberá o nome `dim_produto`, junção do prefixo definido na nomenclatura e da palavra produto que representa os elementos dessa dimensão. As colunas da dimensão devem seguir o mesmo critério para a formação dos nomes.

A Figura a seguir ilustra o desenho da dimensão Produto com os nomes físicos, o tipo de dados e a informação sobre se a coluna pode ou não ficar nula.

Conforme falamos na aula anterior, a dimensão contém uma coluna que identifica um registro na tabela, a Surrogate Key. Essa chave será inserida na tabela Fato (Foreign Key) para que o relacionamento entre elas seja realizado. Na dimensão Produto essa chave chama-se `sk_produto` e é identificada pela sigla PK de Primary Key.

Exemplo da tabela Dimensão Produto

dim_produto
sk_produto: INTEGER NOT NULL [ PK ]
cd_produto: INTEGER NOT NULL ds_produto: VARCHAR(45) NOT NULL nm_fabricante: VARCHAR(100) NOT NULL

 Fonte: próprio autor

### Dica

Vamos praticar? Construa as demais dimensões para o modelo.  
Nesta aula, vamos usar a ferramenta SQL Power Architect Community Edition (s. d.) para fazer a modelagem.

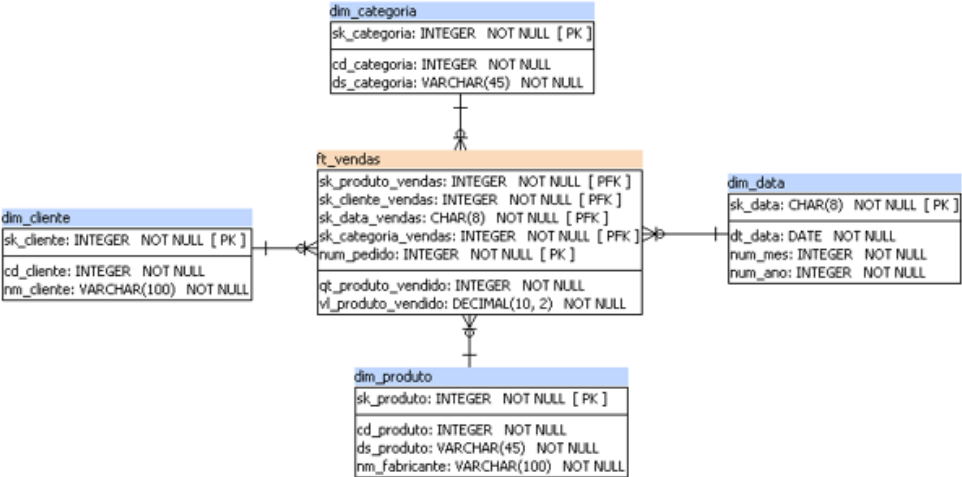
## Tabela Fato

Já sabemos que a tabela Fato reúne as métricas que serão analisadas pelas dimensões e que ela é relacionada às tabelas dimensões do modelo. Agora, vamos entender os efeitos dessa relação!

### Saiba mais

A tabela Fato recebe todas as chaves primárias das dimensões que estão ligadas a ela. Geralmente, a tabela Fato possui uma chave primária e essa chave é uma composição das chaves estrangeiras das dimensões. A chave composta garante que um registro na tabela Fato seja único e, caso haja dois registros com a mesma combinação de chaves, haverá uma exceção que deverá ser tratada no processo ETL.

Observe o modelo a seguir: a tabela ft\_vendas no centro do modelo, recebeu, além das métricas, as chaves estrangeiras correspondentes às chaves primárias das tabelas dimensões.



Exemplo do modelo de dados dimensional físico


Fonte: próprio autor

Cada registro da tabela Fato representa a venda de um produto que pertence a uma categoria, para um cliente, em um determinado dia. Se um mesmo cliente comprar vários produtos em um mesmo dia, haverá várias linhas para esse cliente relacionadas aos diversos produtos comprados.

Comentário

O campo num\_pedido é um dado numérico que não pode ser sumarizado. Ele é o número que identifica o pedido no sistema transacional. Alguns dados podem ser importantes para as análises, mas não possuem características que o definam como uma dimensão. Nesse caso, eles são adicionados na tabela Fato e são denominadas de dimensões degeneradas.

### Restrições

 Clique no botão acima.

As restrições de integridade servem para garantir que os dados cumpram corretamente as regras estabelecidas para a carga na base de dados. Por exemplo, no cenário Supermercado, diariamente ocorrem muitas vendas de produtos. Obrigatoriamente, temos que informar o produto que está sendo vendido, pois ele possui o preço que deverá ser pago pelo cliente.

No entanto, os dados do cliente podem não ser informados no ato da venda na loja física, diferente da venda realizada pela loja on-line, em que a identificação do cliente é obrigatória. Com essa particularidade, a informação do cliente pode ser preenchida ou não, e quando a informação não for preenchida devemos considerar um tratamento adequado para ela.

As dimensões do DW/DM podem receber os elementos Não Informado e Não se Aplica para solucionar problemas desse tipo. O elemento Não Informado é utilizado quando um dado apresenta o valor nulo na área de preparação dos dados e o elemento Não se Aplica é utilizado quando o preenchimento de um dado para o contexto do registro não se aplica.



A figura a seguir ilustra um exemplo sobre a unicidade da chave primária na tabela Fato e o caso do cliente não informado. O código sk\_cliente igual a 1 representa o dado Não Informado. que, nas linhas 1 e 3 do exemplo, o cliente está preenchido com o elemento 1 - Não informado e como eles compraram o mesmo produto no mesmo dia, a restrição de unicidade será violada.

	sk_produto	sk_cliente	sk_data	sk_categoria	qt_produto_vendido	vl_produto_vendido
1	1	1	20200915	6	10	R\$ 10,52
2	1	100	20200915	6	2	R\$ 23,15
3	1	1	20200915	6	5	R\$ 10,52

Exemplo de violação de integridade na tabela Fato Vendas a Varejo. Fonte: próprio autor

Para resolver esse problema o número do pedido deve ser adicionado à chave primária da tabela Fato. Veja o resultado.

ft\_vendas

sk\_produto\_vendas: INTEGER NOT NULL [ PFK ]  
sk\_cliente\_vendas: INTEGER NOT NULL [ PFK ]  
sk\_data\_vendas: CHAR(8) NOT NULL [ PFK ]  
sk\_categoria\_vendas: INTEGER NOT NULL [ PFK ]  
num\_pedido: INTEGER NOT NULL [ PK ]

qt\_produto\_vendido: INTEGER NOT NULL  
vl\_produto\_vendido: DECIMAL(10, 2) NOT NULL

Alteração da PK da tabela Fato Vendas a Varejo. Fonte: próprio autor

Com essa alteração, o problema da unicidade do dado será contornado e o resultado obtido será conforme ilustrado a seguir.

	num_pedido	sk_produto	sk_cliente	sk_data	sk_categoria	qt_produto_vendido	vl_produto_vendido
1	10003	1	1	20200915	6	10	R\$ 10,52
2	10004	1	100	20200915	6	2	R\$ 23,15
3	10008	1	1	20200915	6	5	R\$ 10,52

Resolução da violação PK na tabela Fato Vendas a Varejo. Fonte: próprio autor

As restrições pertinentes às características do SGBD, como o preenchimento das chaves primárias e estrangeiras, são facilmente observadas na construção do modelo de dados dimensional. Contudo, restrições por parte do negócio, como o cliente Não Informado, são variadas e devem ser analisadas com atenção para evitar problemas futuros.

**Atenção!** Aqui existe uma videoaula, acesso pelo conteúdo online

## Tabelas Temporárias

As tabelas temporárias dão suporte ao processo de ETL. Elas recebem os dados que são extraídos dos sistemas de origem e auxiliam os tratamentos que devem ser aplicados aos dados.

### Saiba mais

Nessas tabelas não há restrições de chaves, o dado é copiado e carregado sem qualquer crítica. Após a carga dos dados, a transformação deles pode ocorrer para o conteúdo armazenado. Nesse momento são aplicadas as validações dos dados, a checagem de existência dos elementos e chaves, ocorrendo a integração de dados de sistemas diferentes, entre outros. O resultado das validações é armazenado nessas tabelas, assim como os dados informativos a respeito da limpeza dos registros, quando será possível informar que ele deverá ser carregado na tabela definitiva ou descartado pelo processo. O processo de ETL será detalhado mais à frente.

Assim, normalmente, para cada uma das tabelas Dimensões e tabelas Fato há uma tabela temporária que irá suportar o processo de validação dos dados.

## Tabela Fato Estoque

Conforme o levantamento de requisitos, para a construção das consultas será necessário que o modelo de dados dimensional contenha o desenho apropriado para acomodar os dados referentes ao estoque dos produtos. Com isso, complete o modelo de dados dimensional com a tabela Fato Estoque (ft\_estoque) e os relacionamentos com as dimensões dim\_produto, dim\_data e dim\_cliente.

### Comentário

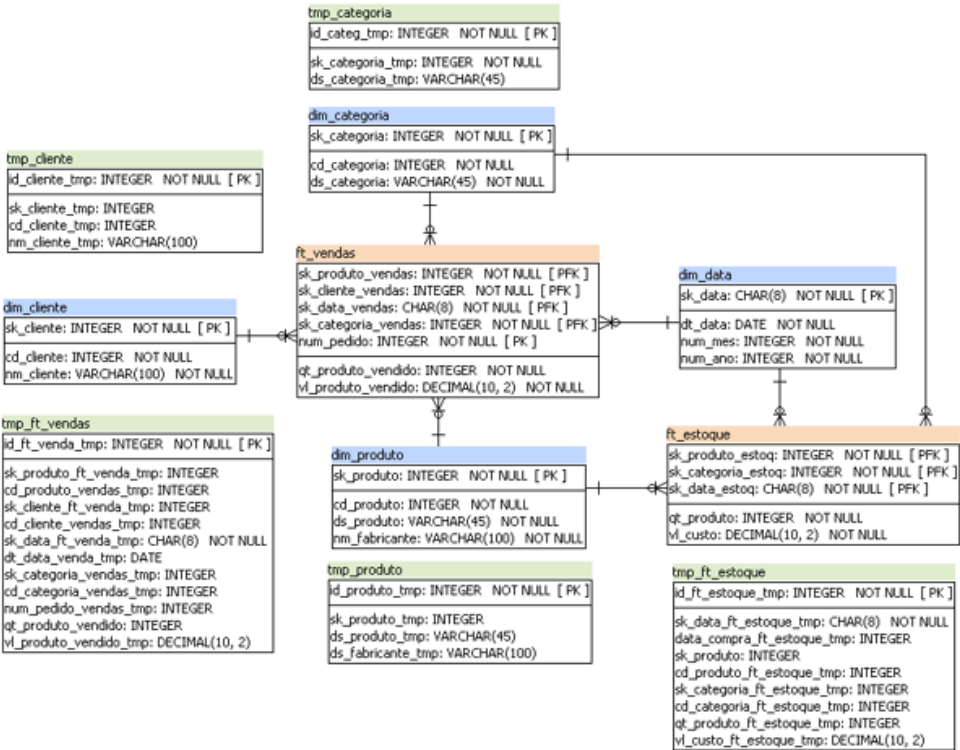
Uma observação importante é que, para relacionar a tabela Fato Estoque com as dimensões não é necessário duplicar as tabelas Dimensões, porque elas devem ser apenas relacionadas com a nova tabela Fato.

Acrescente também as tabelas temporárias ao modelo. Elas não devem ser relacionadas a nenhuma das tabelas do modelo de dados dimensional. Como falado anteriormente, elas darão suporte ao processo ETL.

Veja o resultado do modelo com a inclusão da tabela Fato Estoque e das tabelas temporárias!

Na cor azul estão as Dimensões, na cor laranja as tabelas Fato e na cor verde estão as tabelas Temporárias.

Modelo de Dados Dimensional DW Supermercado.



Fonte: próprio autor

### Armazenamento dos dados

Clique no botão acima.

A estrutura de armazenamento dos dados de um DW/DM conta com espaço em disco disponível, processos de backup e deve ser apoiada por um grupo de atividades importantes para o bom desempenho do DW/DM, como a estrutura correta da criação dos elementos com os nomes padronizados.

Em caso de extensão do DW/DM deve ser verificado se os elementos estão adequados, se não estão sendo criados repetidos ou se os dados com conceitos já existentes estão sendo inseridos nas tabelas corretas; manutenção da documentação, entre outros. O Administrador de Dados (AD) é o responsável por essas atividades e está presente no desenvolvimento de projetos de DW.

Outra atividade importante está relacionada ao Administrador de Banco de Dados (DBA), que é responsável pela manutenção da base de dados, sua integridade e sua criação. Além disso, ele se preocupa com o desempenho da base, muito importante para o DW que possui grandes volumes de dados armazenados.

O particionamento das tabelas Fato e a criação de índices são tarefas realizadas pelo DBA para que o DW tenha um desempenho melhor nas consultas. O particionamento de tabelas e índices é usado para facilitar o gerenciamento de grandes volumes de dados armazenados. O particionamento divide a tabela em várias tabelas e essa divisão pode ser feita verticalmente ou horizontalmente. No particionamento horizontal, a quantidade de linhas é reduzida nas tabelas e, no particionamento vertical, a quantidade de colunas.

Geralmente, a partição é baseada no tempo. Por exemplo, podem ser criadas partições por mês ou ano, e quando uma consulta for submetida para o ano 2020, apenas a partição que está com o conjunto de dados para 2020 será consultada.

Quando os dados são agrupados nas partições, a busca fica restrita apenas à partição em que os dados requeridos estão armazenados. Isso minimiza o tempo de consulta, pois evita que a tabela seja totalmente verificada para trazer os dados solicitados.

Outro recurso que pode ser aplicado pelo DBA são os índices. Os índices são estruturas que auxiliam a recuperação dos dados de maneira mais rápida. No DW/DM, que possui alto volume de dados, é recomendado criar os índices para otimizar as consultas submetidas à base de dados. Para dados com baixa cardinalidade, normalmente são usados índices do tipo bitmap, mas cada caso deve ser examinado para que a melhor ação seja tomada.

Além das partições e índices que podem ser criados pelo DBA para melhorar o desempenho das consultas no ambiente analítico, há também as agregações de dados que são armazenadas em tabelas. Esse ponto será melhor explorado na próxima aula.

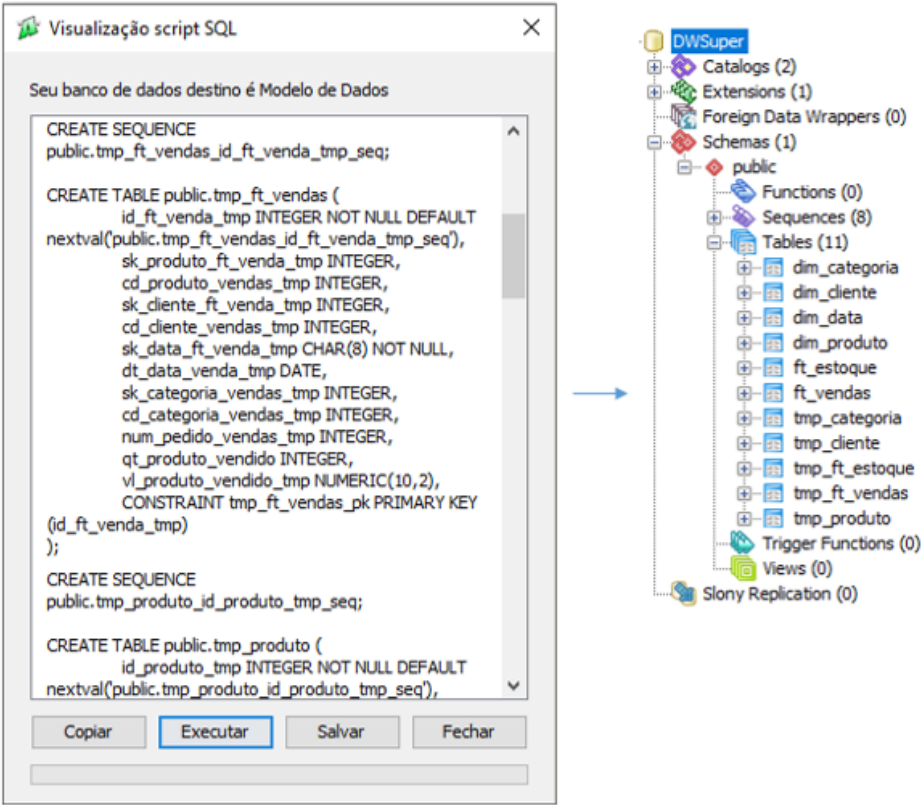
**Atenção!** Aqui existe uma videoaula, acesso pelo conteúdo online

## Implementação do Modelo de Dados Dimensional

Após a correta preparação da base de dados, a implementação do Modelo de dados Dimensional pode ser realizada. Algumas ferramentas de modelagem geram o script para a criação dos elementos tabelas, chaves, etc. Esse recurso facilita a criação dos elementos, mas também podem ser criados diretamente no SGBD seguindo as definições do modelo de dados dimensional físico. As ferramentas MySQL Workbenck e SQL Power Architect possuem esse recurso.

O SQL Power Architect permite que você escolha o banco de dados no qual o modelo será fisicalizado. No nosso exemplo, vamos criar no SGBD PostgreSQL. A ferramenta usa a conexão criada no início da criação do modelo e executa o script criando a base de dados.

Criação da base de dados



Nesse momento, a base de dados para o DW Supermercado está criada. No entanto, com o andamento do projeto e até mesmo depois da conclusão, novas necessidades podem surgir e então o modelo criado pode sofrer alterações para atender às novas demandas. Esse trabalho deve ser feito com cautela para assegurar que o modelo criado e os dados nele contidos não sofram perdas devido ao crescimento do ambiente.

Comentário

Normalmente, e é altamente recomendado, as tarefas são construídas no ambiente de desenvolvimento em que os testes são realizados, e somente após esses passos as alterações são efetivadas no ambiente de produção. Em empresas maiores, há ainda um terceiro ambiente chamado Homologação, onde os elementos desenvolvidos e as alterações feitas no processo são testadas pelo usuário. Somente após esse passo a alteração pode ser refletida no ambiente de produção.

Estudamos nesta aula o projeto físico do modelo de dados dimensional, algumas restrições que devem ser consideradas, alguns aspectos relevantes para o armazenamento dos dados e para a implementação do modelo de dados dimensional físico no SGBD.

Agora, vamos fixar o entendimento!

## Atividades

---

1. Sobre o modelo de dados dimensional físico:

- a) É construído com base nos padrões estabelecidos, nas regras de negócio e considera as características do SGBD.
  - b) É independente do SGBD utilizado e considera as regras de negócio estabelecidas.
  - c) É construído com base nas regras de negócio e não considera as características do SGBD utilizado.
  - d) Não segue padrões específicos, mas é apoiado nas regras de negócio e considera as características do SGBD.
  - e) É construído com base nos padrões estabelecidos, sem regras de negócio e considera as características do SGBD.
- 

2. As restrições de integridade garantem as regras estabelecidas para a carga na base de dados. Sobre essa afirmação:

- a) Não se aplica a modelos de dados dimensionais.
  - b) Podem ser aplicadas aos modelos de dados dimensionais, mas podem ser desconsideradas.
  - c) É aplicada ao modelo de dado dimensional e garante que as relações importantes sejam respeitadas.
  - d) Aplicam-se apenas às tabelas dimensões de Data e Fato.
  - e) É aplicada somente às tabelas temporárias.
- 

3. Sobre o armazenamento dos dados de um projeto de DW/DM:

- a) Não ocupa muito espaço em disco; logo, basta criar o modelo de dados físico na base de dados disponível na organização.
  - b) A única medida a ser tomada para melhorar o desempenho das consultas analíticas é o aumento do espaço em disco.
  - c) Não existem especialistas que possam dedicar-se às tarefas envolvidas em um armazenamento de dados volumoso como em um DW.
  - d) Não é possível realizar particionamento devido ao volume de dados armazenado.
  - e) Deve ser avaliado o espaço em disco disponível, processos de backup, particionamento, entre outros.
- 

## Notas

Lorem Ipsum é simplesmente uma simulação de texto da indústria tipográfica e de impressos. Lorem Ipsum é simplesmente uma simulação de texto da indústria tipográfica e de impressos. Lorem Ipsum é simplesmente uma simulação de texto da indústria tipográfica e de impressos.

## Título modal <sup>1</sup>

Lorem Ipsum é simplesmente uma simulação de texto da indústria tipográfica e de impressos. Lorem Ipsum é simplesmente uma simulação de texto da indústria tipográfica e de impressos. Lorem Ipsum é simplesmente uma simulação de texto da indústria tipográfica e de impressos.

## Referências

KIMBALL, M. R. R. **The Data Warehouse Toolkit - The Definitive Guide to Dimensional Modeling**. 3. ed. Indianapolis, Indiana: John Wiley Sons, 2013. Disponível em: <https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/books/data-warehouse-dw-lifecycle-toolkit/>. Acesso em: 10 set. 2020.

PITON, R. **Data Warehouse Passo a Passo – O guia prático de como construir um Data Warehouse do zero**. Porto Alegre: Edição do Autor, 2018.

SQL Power Architect Community Edition. **Best of BI Productivity Tools**. Disponível em: [http://www.bestofbi.com/page/architect\\_download\\_os](http://www.bestofbi.com/page/architect_download_os). Acesso em: 5 set. 2020.

## Próxima aula

- Hierarquias;
- Agregações.

## Explore mais

- Você já possui um SGBD instalado? A criação de uma base de dados pode ser feita em qualquer SGBD, mas, caso ainda não tenha um preferido, você pode visitar os sites dos SGBDs PostgreSQL e MySQL, e escolher um deles para realizar os exercícios. Eles são simples de instalar e utilizar.