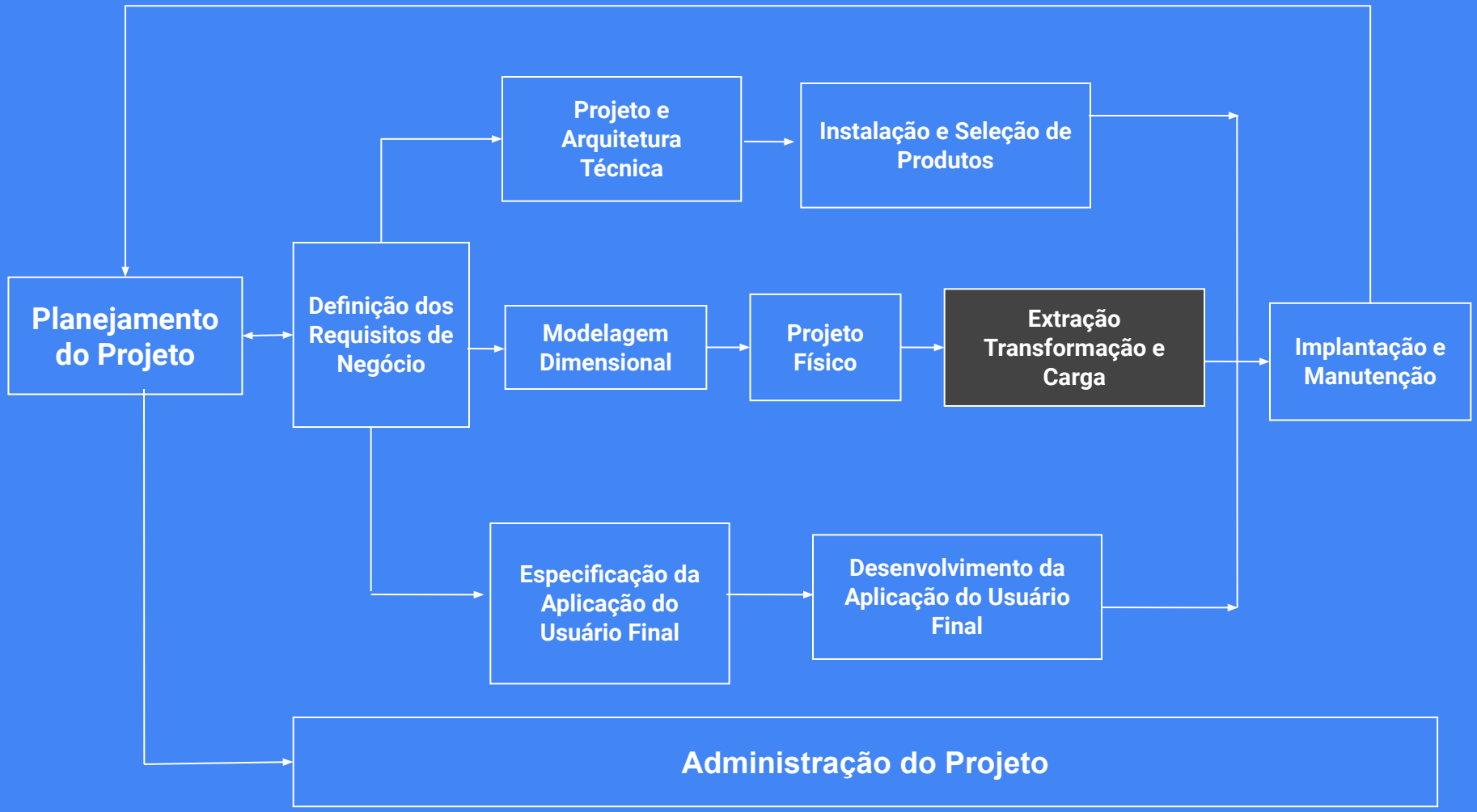
INE 5643
Data Warehouse
Aula 12 - Back Room - Carga

Prof. Mateus Grellert
Prof. Renato Fileto
Créditos: Prof. Tite Todesco **(slides originais, adaptados pelos professores atuais)**

Departamento de Informática e Estatística (INE)
Universidade Federal de Santa Catarina (UFSC)

Próxima Aula - Ciclo de Projeto DW



Processos ETL

- Serviços de Extração;
- Serviços de Transformação;
- Serviços de Carga de Dados;
- Serviços de Controle e Gerenciamento.

Processos ETL

- Serviços de Extração;
- Serviços de Transformação;
- Serviços de **Carga de Dados**;
- Serviços de **Controle e Gerenciamento**.

O Processo de Carga

Usar o utilitário de carga do gerenciador de banco de dados do DW:

- é o meio de carga mais eficiente de dados;
- seu uso é quase universal;
- guarda os logs facilitando o gerenciamento de erros da carga;
- facilita as conversões agilizando o processo.

Carga Histórica dos Fatos

Processamento da Tabela Fato:

- Para a carga de dados passados, verificar registro da dimensão correspondente ao período, através de campos **data inicial** e **data final** de uso;
- Importante se certificar de que será mantida a **integridade** referencial entre a tabela de fatos e as dimensões;

Carga Histórica dos Fatos

Valores Nulos:

- Podem ter dois significados:
  - O atributo não possuía valor no momento do registro da transação;
  - O atributo deveria ter valor, mas algum problema na transação ocorreu;
- Chaves operacionais (como datas) poderão estar nulas ocasionando problemas de **integridade**.

Carga Histórica dos Fatos

Valores Derivados:

- Se forem acessados **esporadicamente**, podem ser acessados através de uma visão (*view*), a qual mantém a regra de cálculo do valor derivado;
- Se utilizado de forma **freqüente** ou se utilizado como filtro, deverá ser armazenado fisicamente e indexado.

Migração Incremental para a Tabela Fato

Muitas tabelas fato tornam-se tão grandes que não podem ser carregadas de uma só vez. Uma das técnicas mais comuns para reduzir o tamanho do processo de migração é carregar somente dados novos ou alterados.

Migração Incremental para a Tabela Fato

Carga da Tabela de Fato:

- Carregamento mais frequente: sair de processo mensal ou semanal para um noturno;
- Arquivos e índices particionados;
- Execução em paralelo;
- Banco de dados em paralelo;
- Tabelas duplicadas.

Tabelas Agregadas e Carga MOLAP

Tabelas Agregadas:

- São o resultado de uma consulta agregada grande, armazenada como tabela;
- Problemáticas quando são muito grandes para serem processadas dentro da período de carga;
- Mesmo cenário da atualização das tabelas bases: questão da atualização incremental X *full refresh*.

Tabelas Agregadas e Carga MOLAP

Tabelas Agregadas:

- Frequentemente, o período de agregação é semanal ou mensal. Duas possibilidades:
  - Não incluir o período mais recente até que tenha acabado
  - Atribuir um status especial ao período atual e acrescentar seus dados aos já existentes.

Automação e Operação do DW

Abordagens de controle de tarefa:

- As opções disponíveis dependem da infraestrutura básica do DW, da experiência da equipe, da complexidade do processo de estagiamento e dos recursos disponíveis.

Automação e Operação do DW

Metadados da extração:

- Uma vez completada a extração o processo precisa salvar diversas peças de metadados. Há dois tipos:
  - Metadados de gerenciamento de processos;
  - Metadados de medida do processo.

Automação e Operação do DW

- Para manter o processo em funcionamento, a extração precisa notificar o processo seguinte para iniciar;
- O novo processo pode ser o processo de transformação associado ou outro processo de extração que depende do atual.
- Alguns monitoramentos de processos:
  - Existência de arquivos;
  - Configuração de flags no catálogo de metadado;
  - Mecanismo intrínseco da própria ferramenta de ETL.

Automação e Operação do DW

Agenda de Tarefas Típicas:

- Processar fatos
  - Transformações de dados;
  - *Surrogate key lookup*;
  - Checar integridade referencial (*RI check*);
  - Salvar registros falhos no log;
  - Processar registros falhos.

Automação e Operação do DW

Agenda de Tarefas Típicas:

- Processar agregados;
  - Validar a carga junto aos metadados;
  - Carregar dimensões antes da tabela de fatos, forçando integridade referencial;
  - Carregar fatos;
  - Carregar agregados;
  - Rever o processo de carga.

Automação e Operação do DW

Limpeza e Qualidade de Dados:

- Nos DWs em implementação, os dados são limpos por dois processos:
  - Entrada limpa de dados;
  - Corrigir problemas assim que os dados são inseridos.
- Praticamente não existe sistema que contenha somente dados perfeitos.

Automação e Operação do DW

Garantia de Qualidade do Dado:

- O dado que está prestes a ser carregado está correto?
  - Auditoria garante que o número de linhas está correto;
  - A inserção correta dos registros (mantendo restrições, integridade referencial, etc.) garante a consistência;
  - E o conteúdo dos dados?
- *Cross-Footing*
  - Cruzamento das informações do DW com o sistema fonte;

Automação e Operação do DW

Outros Aspectos:

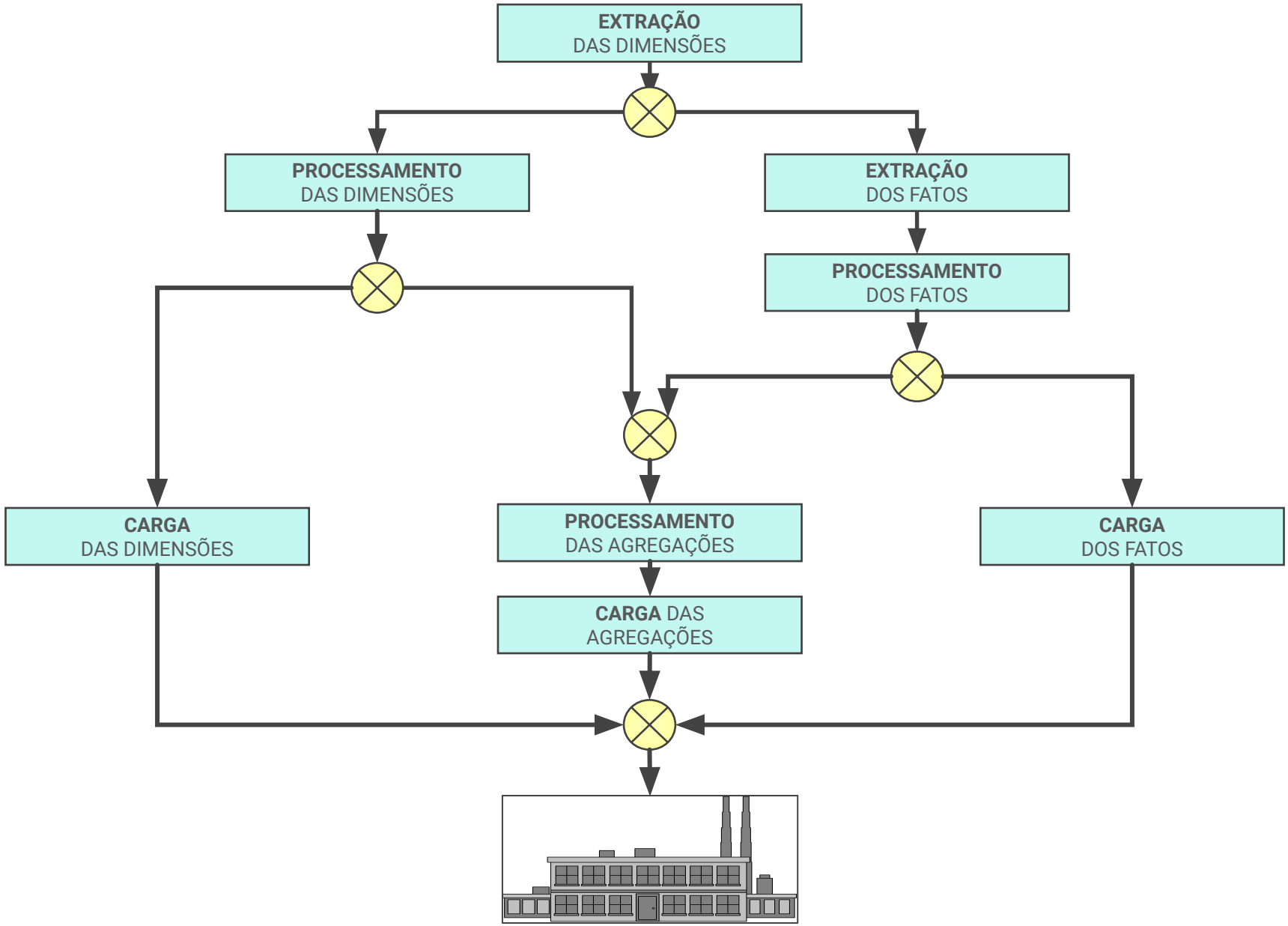
- Armazenamento na área de estagiamento
  - Uma função importante da área de estagiamento de dados é armazenar o dado tanto em sua forma mais granular como na forma mais agregada;
  - Os arquivos da área de estagiamento deveriam ser capazes de restaurar um data mart.

Automação e Operação do DW

Outros Aspectos:

- Segmento de *rollback* do sistema fonte
  - Extrações obtidas junto a fontes relacionais podem ser problemáticas quando demoram, por vezes devido a tranca (*lock*) nos registros;
- Gerenciamento de espaço em disco
  - Um dos problemas mais comuns no processo de estagiamento é o estouro de espaço em disco no meio do processo.

Overview dos Processos ETL



Vamos praticar!

Prática

- No Exercício de hoje vamos carregar dados de novos pedidos (*orders_export.csv*) e gerar uma tabela fato **Vendas**
- Como esse arquivo de dados vem de um processo de negócio da empresa, cada pedido contém as chaves de todas as dimensões de interesse (IDs dos clientes)
  - ORDER_ID
  - **ORDER_DATE**
  - **CUSTOMER_ID**
  - ORDER_STATUS
  - ORDER_TOTAL
  - **SALES_REP_ID**
  - **PROMO_ID**
  - LINE_ITEM_ID
  - **PRODUCT_ID**
  - UNIT_PRICE
  - QUANTITY

Dimensões Degeneradas (Degenerate Dimensions)

- Digamos que nosso grão no tempo é o **1 dia**
- O que acontece se o mesmo cliente comprou o mesmo produto na mesma loja com o mesmo representante no mesmo dia?

Entradas **duplicadas** na tabela Fato

CDIM_ID	PRODDIM_ID	DATEDIM_ID	STOREDIM_ID	EMPDIM_ID	DollarSales
3	6000	877	9	2	$45
3	6000	877	9	2	$45

Dimensões Degeneradas (Degenerate Dimensions)

- Digamos que nosso grão no tempo é o **1 dia**
- O que acontece se o mesmo cliente comprou o mesmo produto na mesma loja com o mesmo representante no mesmo dia?

Entradas **duplicadas** na tabela Fato

CDIM_ID	PRODDIM_ID	DATEDIM_ID	STOREDIM_ID	EMPDIM_ID	DollarSales
3	6000	877	9	2	$45
3	6000	877	9	2	$45

- **Solução: adicionar uma dimensão extra para termos controle sobre o ID do pedido, evitando duplicação**

CDIM_ID	PRODDIM_ID	DATEDIM_ID	STOREDIM_ID	EMPDIM_ID	OrderNum	DollarSales
3	6000	877	9	2	ON962334	$45
3	6000	877	9	2	ON962356	$45

Por hoje é só! 💪