Extração de Dados

Sistemas de Apoio à Decisão

Extração de Dados

- Sumário
  - Introdução
  - Processo de Extração
  - Técnicas CDC
    - Técnicas de Extração Incremental
    - Técnicas de Extração Completa

Introdução

- Processo ETL
  - Permite migrar dados dos sistemas fonte para a BD do *Data Warehouse*, procedendo às necessárias transformações
    - Formato e conteúdo
  - Não é apenas a mera justaposição de três processos bem definidos:
    - Extração
    - Transformação
    - Carregamento

Introdução

- Processo ETL
  - Existe grande interdependência entre estes três processos
    - Numa perspetiva pedagógica podem ser abordados de forma independente
  - ETL é apontado como o grande problema escondido dos *Data Warehouses*
  - Normalmente consome cerca de 70% dos custos de construção e manutenção do *Data Warehouse*

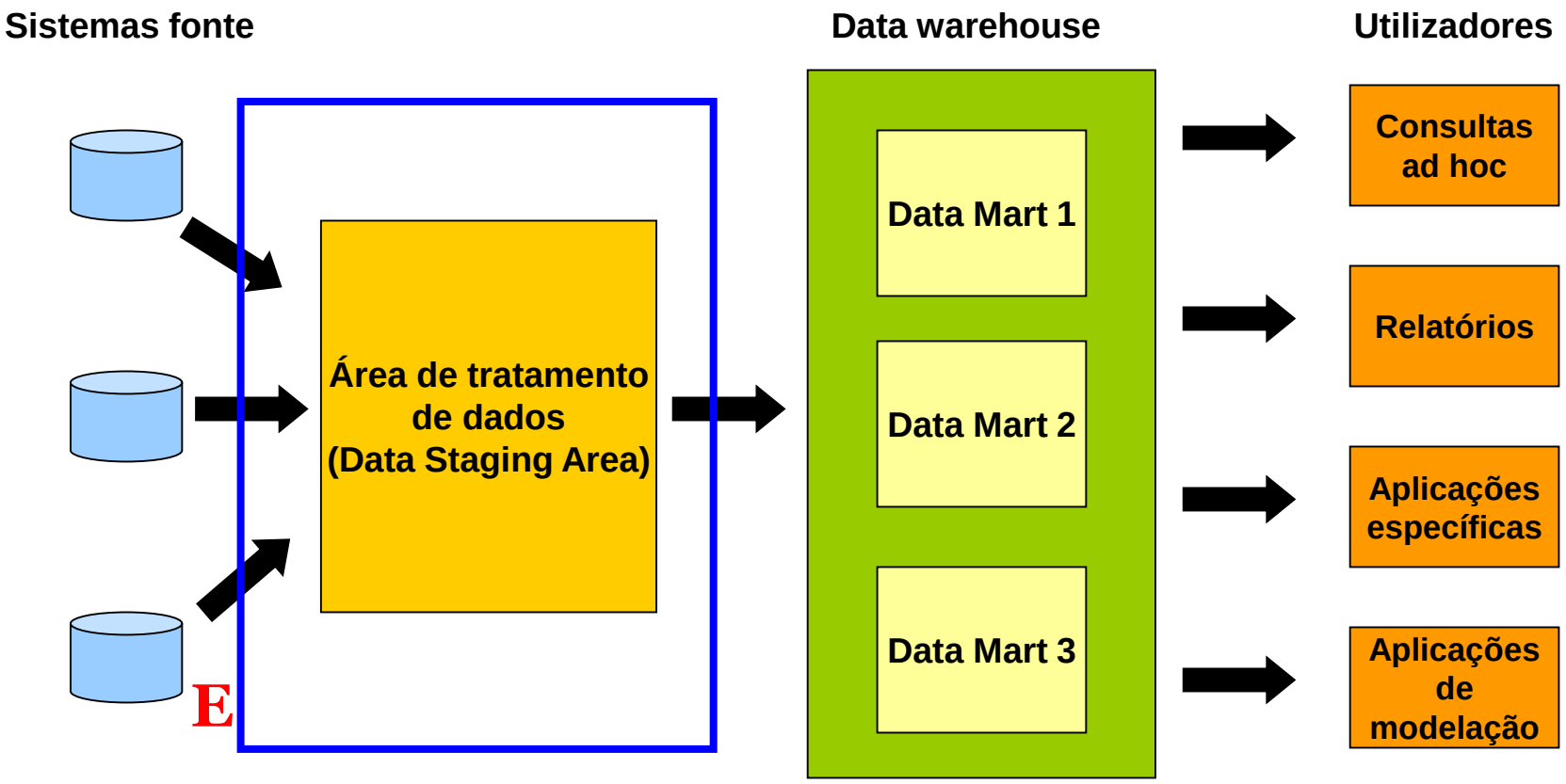
Introdução

- Processo ETL
  - Área de Tratamento de Dados (DSA)
  - Tem associado um conjunto de processos que permitem extrair, transformar e carregar os dados fonte para serem utilizados no *Data Warehouse*
  - Analogia entre um *Data Warehouse* e um restaurante
    - A área de tratamento de dados corresponde à cozinha do restaurante

Extração de Dados

- Sumário
  - Introdução
  - Processo de Extração
  - Técnicas CDC
    - Técnicas de Extração Incremental
    - Técnicas de Extração Completa

Processo de Extração



Processo de Extração

- Introdução
  - A extração de dados consiste no processo de compreender, selecionar e copiar os dados fonte para a área de tratamento de dados (DSA)
  - Duas abordagens principais
    - Exportação de dados
      - Os dados são convertidos num ficheiro que é depois lido para a DSA
    - Extração de dados
      - Utilização de código específico que transfere diretamente os dados para a DSA

Processo de Extração

- Introdução
  - O processo de extração precisa da cooperação dos sistemas fonte
  - No processo de extração existem duas situações bem distintas
    - Primeira extração de dados
    - Extrações incrementais
      - Novos dados
      - Dados que sofreram alterações

Processo de Extração

- Análise dos sistemas fonte
  - Começar pelo DER, se existir
  - Se não existir um DER fazer o *reverse engineering* da BD operacional
    - As ferramentas de modelação de dados e de ETL possuem esta funcionalidade
  - Procurar descrições das tabelas e dos campos da base de dados, mesmo que estas estejam desatualizadas
  - Falar com o "guru" da BD para perceber as modificações que ocorreram

Processo de Extração

- Análise do conteúdo dos dados
  - Detetar anomalias nos dados
    - Valores nulos em chaves estrangeiras
    - Valores nulos noutras colunas (regra de negócio para lidar com os valores a NULL)
  - Datas em campos que não representam datas
    - Existem vários formatos para as datas
      - 08-10-2025
      - 2025/10/08
      - 8 outubro, 2025
      - …

Processo de Extração

- Extração de diferentes plataformas
  - Integração de dados de fontes heterogéneas
    - Processo semelhante ao que ocorre quando há uma fusão entre empresas
    - Fontes de dados típicas:
      - BD operacionais
      - Ficheiros CSV, XML, JSON
      - Páginas Web, Web logs
      - ERPs, CRMs
      - ...

Processo de Extração

- Extração de dados que mudam
  - Técnicas CDC – *Change Data Capture*
    - Permitem identificar alterações nos dados
  - Na primeira iteração do processo ETL esta questão não se coloca
  - O planeamento para a extração de dados que mudam tem de ser feito antes do primeiro carregamento
  - Capturar as modificações nos dados fonte é crucial

Processo de Extração

- Extração de dados que mudam
  - Existem várias técnicas para deteção de dados que mudam
    - *Timestamps*
    - *Triggers*
    - Partições
    - Processo de eliminação
    - Outras técnicas
      - Análise de *logs*
      - Baseadas numa data
      - …

Extração de Dados

- Sumário
  - Introdução
  - Processo de Extração
  - Técnicas CDC
    - Técnicas de Extração Incremental
    - Técnicas de Extração Completa

Técnicas CDC: *Timestamps*

- *Timestamps*
  - Alguns SGBD permitem que as suas tabelas tenham uma coluna na qual é registada qual a data/hora da alteração de cada registo
    - Ferramentas integradas para deteção
  - Técnica de extração incremental

Técnicas CDC: *Triggers*

- Introdução
  - Permitem simular o mecanismo por *timestamps*
  - Quando um registo no sistema fonte for inserido/modificado, marcar esse registo com a data/hora do sistema
  - Técnica de extração incremental

Técnicas CDC: *Triggers*

- Implementação
  - No sistema fonte adicionar às tabelas uma coluna que guarde data e hora (DATE)
  - Criar *triggers* para "disparar" por cada INSERT/UPDATE
  - Para cada tabela do sistema fonte criar a tabela correspondente na DSA
  - Após a extração de uma fonte, guardar data/hora do último registo extraído dessa fonte

Técnicas CDC: *Triggers*

- Passos típicos do algoritmo
  1. Limpar a tabela destino
  2. Obter a data do último registo extraído da tabela fonte
  3. Obter a data do registo mais recente da tabela fonte
  4. Extrair os dados da tabela fonte
  5. Guardar a data do último registo extraído da tabela fonte

Técnicas CDC: *Triggers*

- Exemplo
  - Antes da 1ª Extração: Contexto

SISTEMA FONTE

T_PRODUCTS

serial_nr	name	src_last_changed
164425	Monitor 23", LG	2025-10-01 09:15:07
234234	Monitor 25 inch., Samsung	2025-10-01 19:40:01
233324	Pen Drive 64 Gigabytes	2025-10-01 20:20:21
552424	USB Pen 32GB	2025-10-02 09:10:45

DATA STAGING AREA

T_INFO_EXTRACTIONS

source_table_name	last_timestamp
T_PRODUCTS	NULL
...	...

T_DATA_PRODUCTS

id	name

Técnicas CDC: *Triggers*

- Exemplo
  - 1ª Extração: Passo 1

SISTEMA FONTE

T_PRODUCTS

serial_nr	name	src_last_changed
164425	Monitor 23", LG	2025-10-01 09:15:07
234234	Monitor 25 inch., Samsung	2025-10-01 19:40:01
233324	Pen Drive 64 Gigabytes	2025-10-01 20:20:21
552424	USB Pen 32GB	2025-10-02 09:10:45

DATA STAGING AREA

T_INFO_EXTRACTIONS

source_table_name	last_timestamp
T_PRODUCTS	NULL
...	...

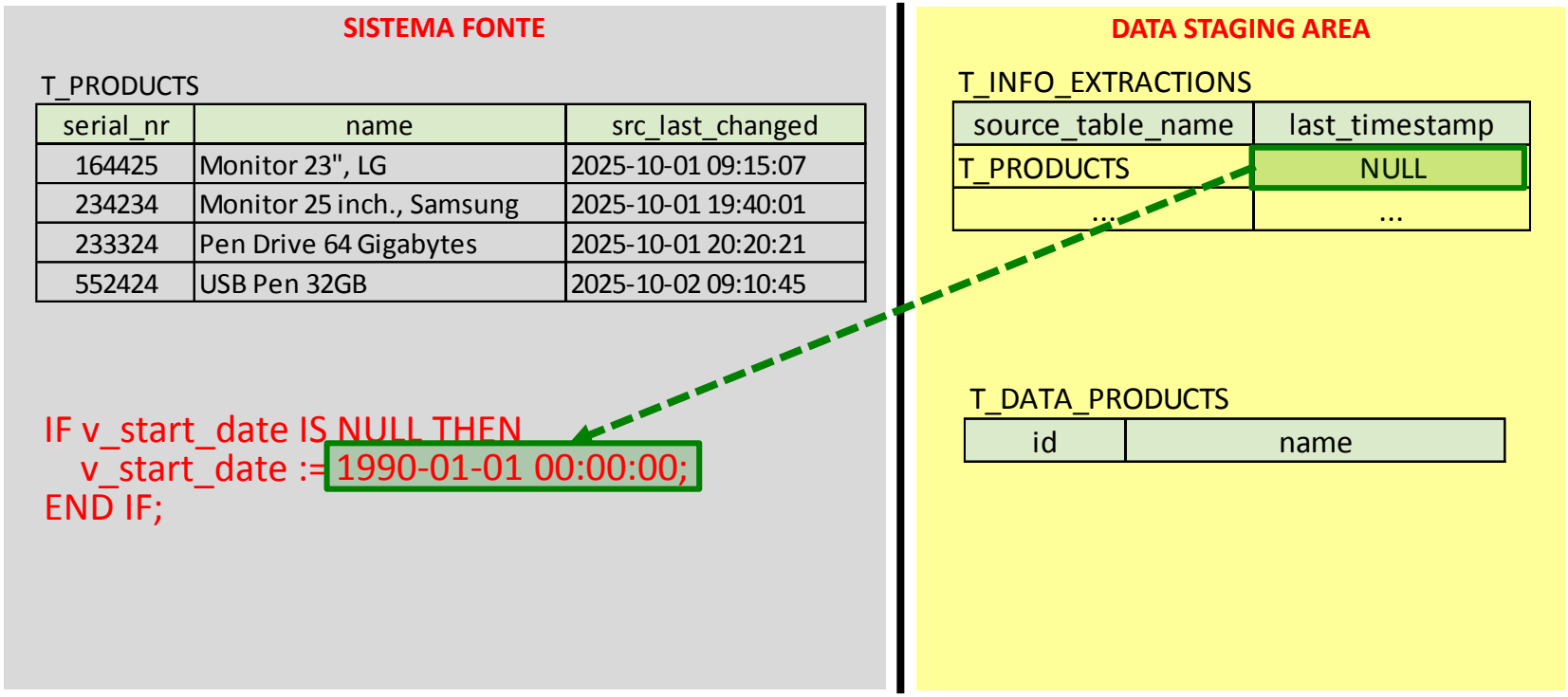
T_DATA_PRODUCTS

id	name

limpar a tabela destino

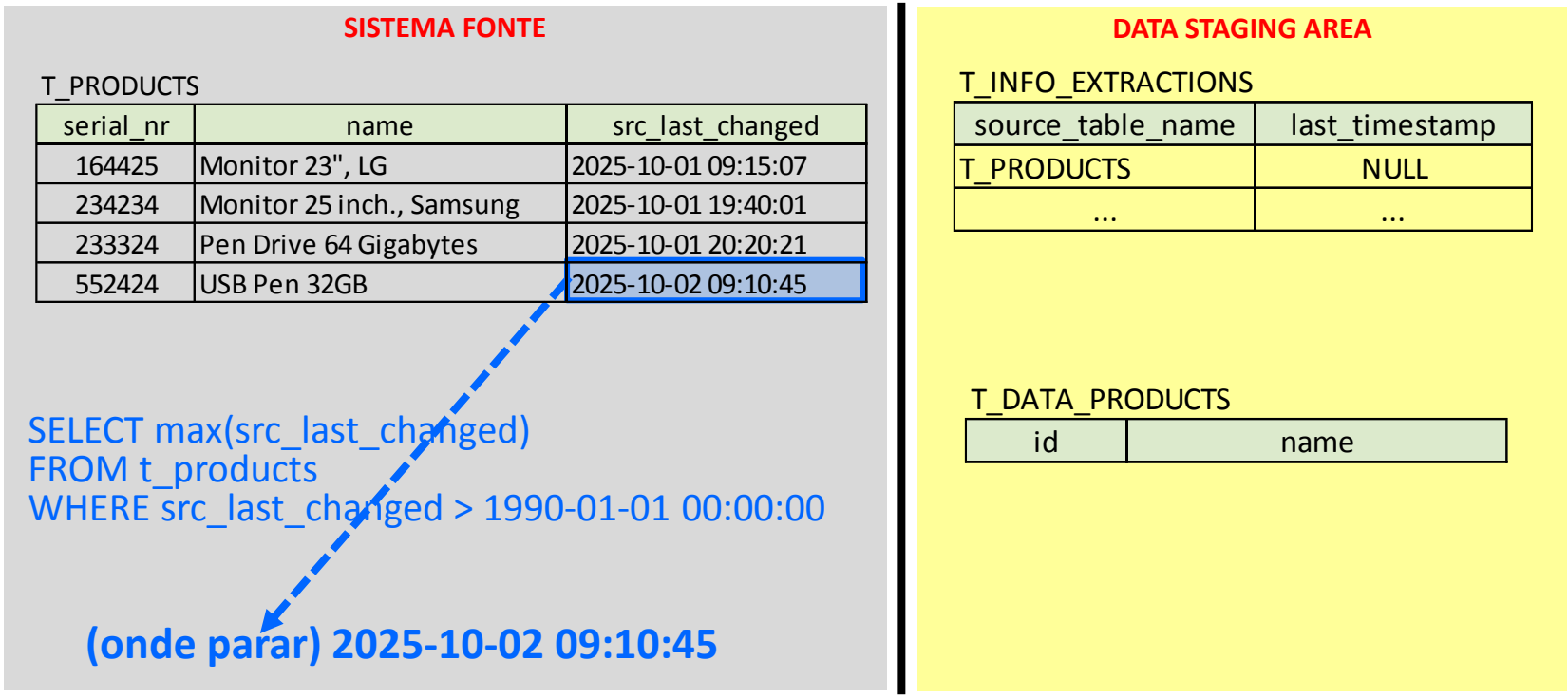
Técnicas CDC: *Triggers*

- Exemplo
  - 1ª Extração: Passo 2



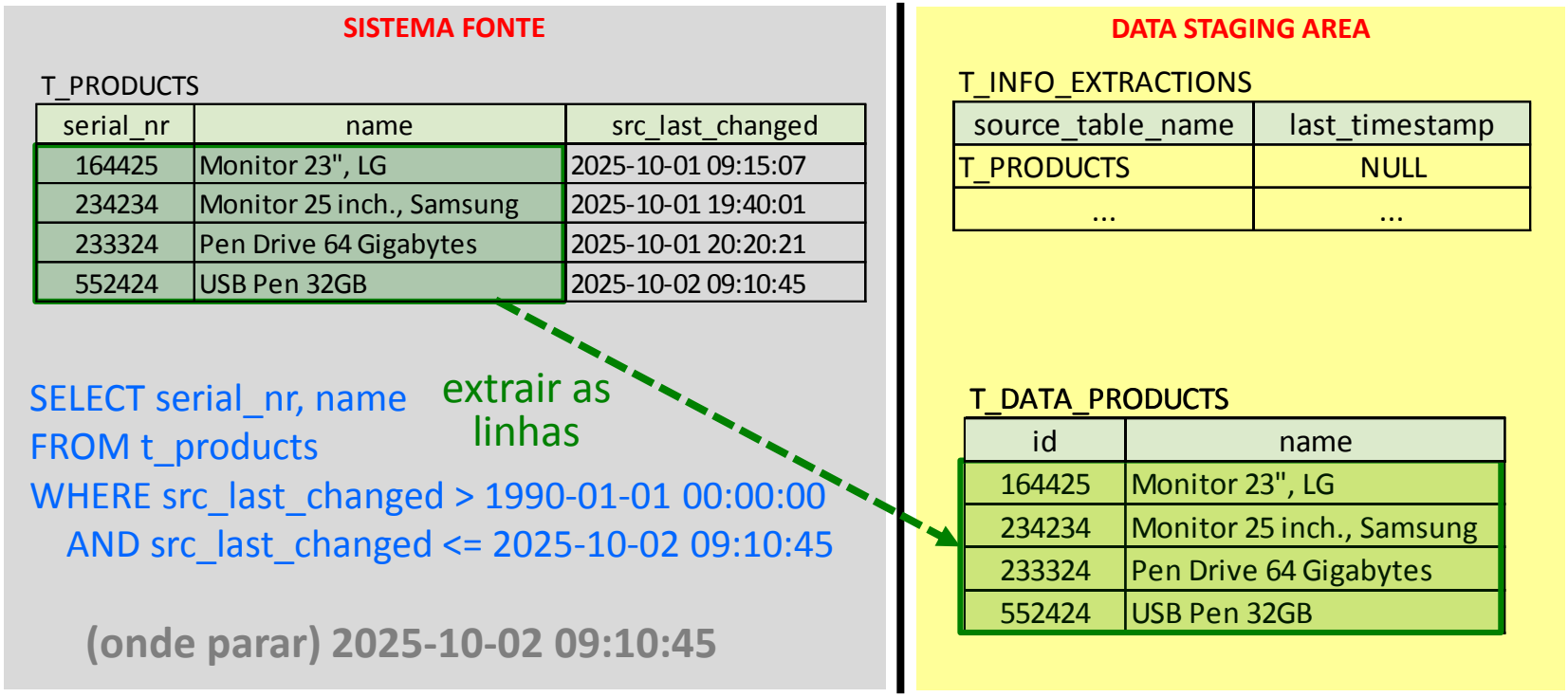
Técnicas CDC: *Triggers*

- Exemplo
  - 1ª Extração: Passo 3



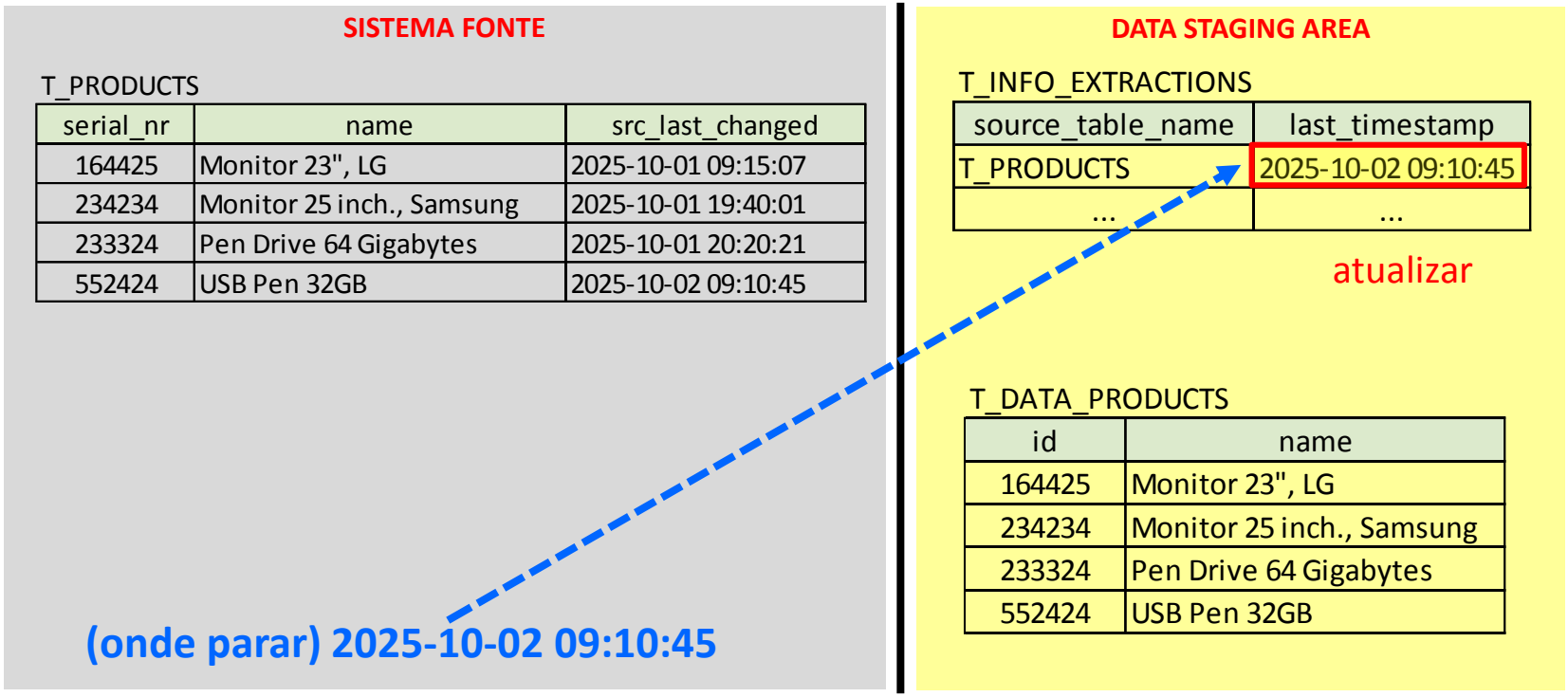
Técnicas CDC: *Triggers*

- Exemplo
  - 1ª Extração: Passo 4



Técnicas CDC: *Triggers*

- Exemplo
  - 1ª Extração: Passo 5

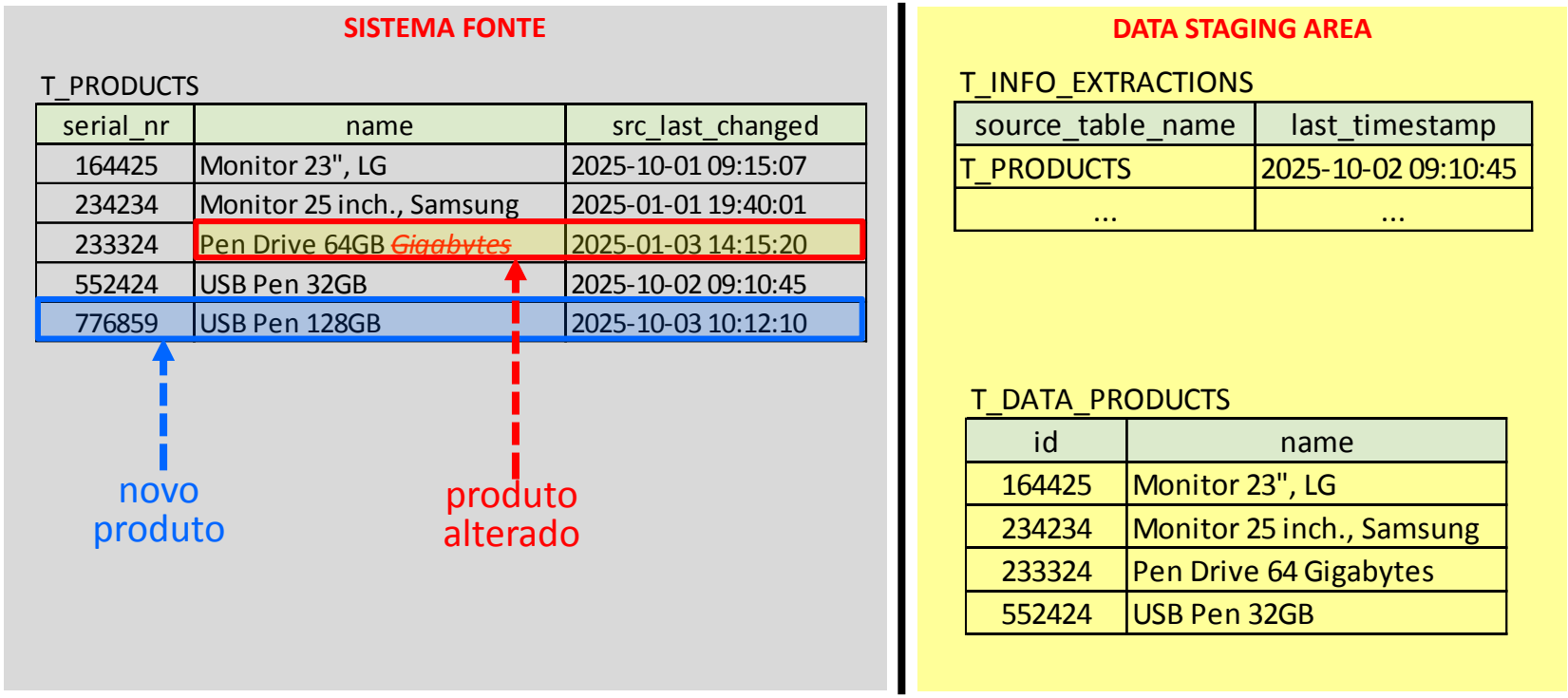


Técnicas CDC: *Triggers*

- Exemplo
  - O que é que acontece aos dados depois de extraídos para a DSA (1ª extração)?
    - Tabelas T_DATA_*
  - São transformados
    - Tabelas T_CLEAN_*
  - São carregados para o DW
    - Tabelas T_DIM_* e T_FACT_*
    - Antes do 1º carregamento as tabelas do DW estão vazias

Técnicas CDC: *Triggers*

- Exemplo
  - Alterações no Sistema Fonte



Técnicas CDC: *Triggers*

- Exemplo
  - 2ª Extração: Passo 1

SISTEMA FONTE

T_PRODUCTS

serial_nr	name	src_last_changed
164425	Monitor 23", LG	2025-10-01 09:15:07
234234	Monitor 25 inch., Samsung	2025-10-01 19:40:01
233324	Pen Drive 64GB	2025-10-03 14:15:20
552424	USB Pen 32GB	2025-10-02 09:10:45
776859	USB Pen 128GB	2025-10-03 10:12:10

DATA STAGING AREA

T_INFO_EXTRACTIONS

source_table_name	last_timestamp
T_PRODUCTS	2025-10-02 09:10:45
...	...

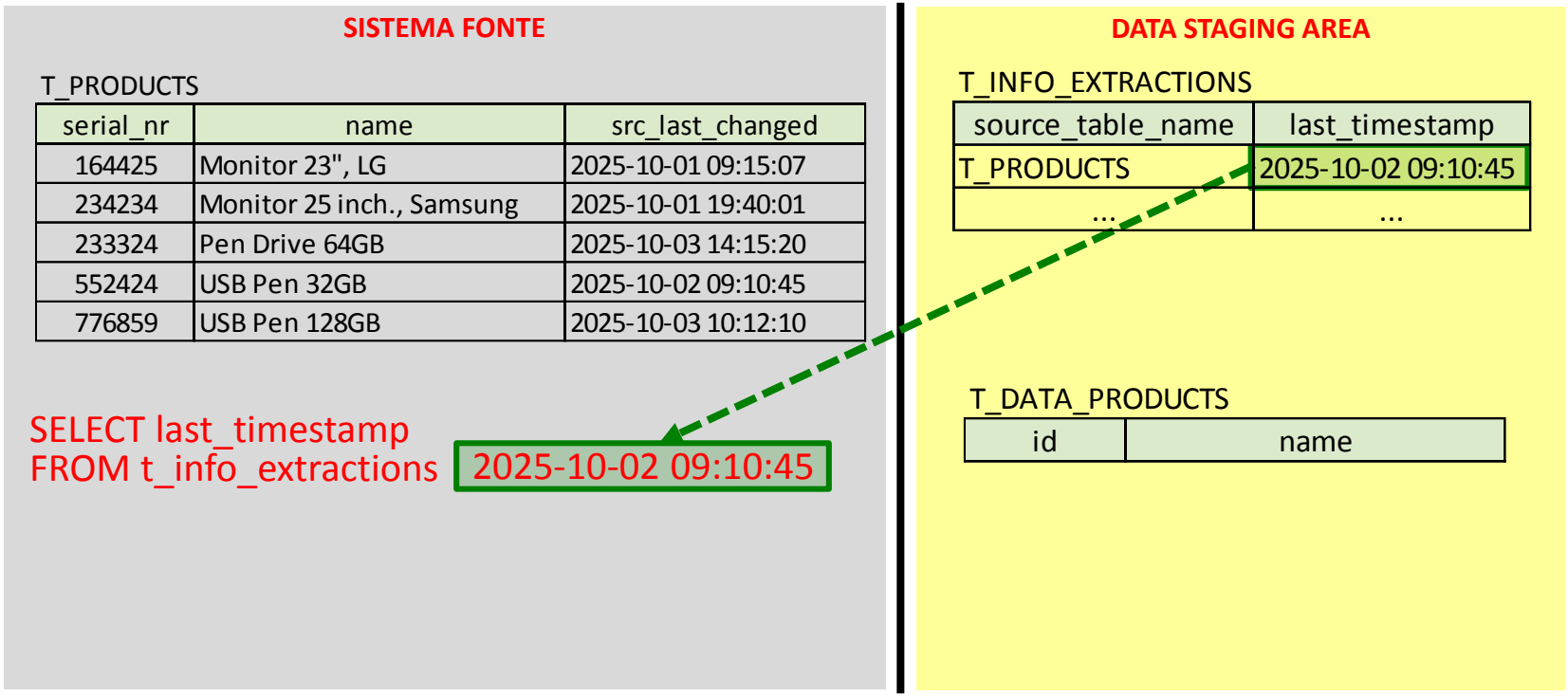
T_DATA_PRODUCTS

id	name
164425	Monitor 23", LG
234234	Monitor 25 inch., Samsung
233324	Pen Drive 64 Gigabytes
552424	USB Pen 32GB

limpar a tabela destino

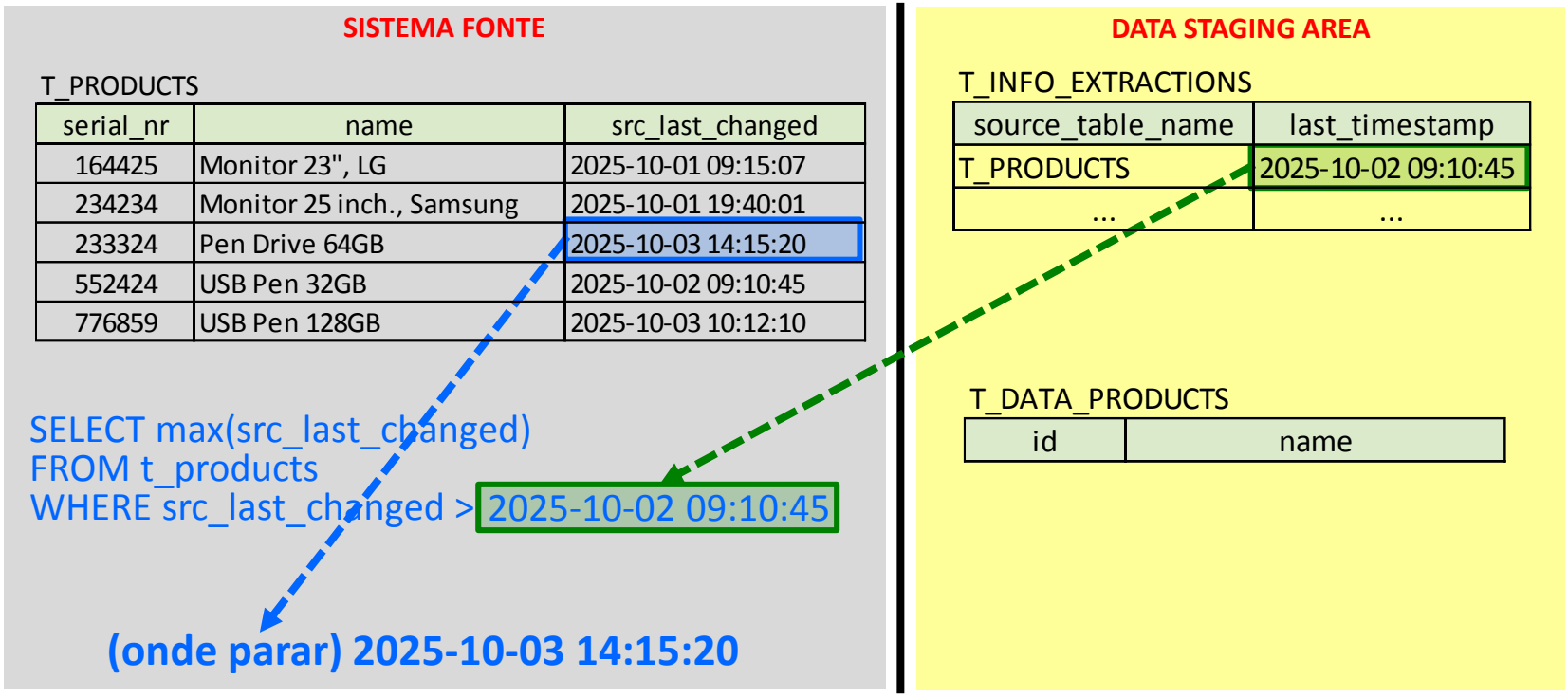
Técnicas CDC: *Triggers*

- Exemplo
  - 2ª Extração: Passo 2



Técnicas CDC: *Triggers*

- Exemplo
  - 2ª Extração: Passo 3



Técnicas CDC: *Triggers*

- Exemplo
  - 2ª Extração: Passo 4

SISTEMA FONTE

T_PRODUCTS

serial_nr	name	src_last_changed
164425	Monitor 23", LG	2025-10-01 09:15:07
234234	Monitor 25 inch., Samsung	2025-10-01 19:40:01
233324	Pen Drive 64GB	2025-10-03 14:15:20
552424	USB Pen 32GB	2025-10-02 09:10:45
776859	USB Pen 128GB	2025-10-03 10:12:10

```
SELECT serial_nr, name
FROM t_products
WHERE src_last_changed > 2025-10-02 09:10:45
   AND src_last_changed <= 2025-10-03 14:15:20
```

(onde parar) 2025-10-03 14:15:20

DATA STAGING AREA

T_INFO_EXTRACTIONS

source_table_name	last_timestamp
T_PRODUCTS	2025-10-02 09:10:45
...	...

extrair linhas

T_DATA_PRODUCTS

id	name
233324	Pen Drive 64GB
776859	USB Pen 128GB

Técnicas CDC: *Triggers*

- Exemplo
  - 2ª Extração: Passo 5

SISTEMA FONTE

T_PRODUCTS

serial_nr	name	src_last_changed
164425	Monitor 23", LG	2025-10-01 09:15:07
234234	Monitor 25 inch., Samsung	2025-10-01 19:40:01
233324	Pen Drive 64GB	2025-10-03 14:15:20
552424	USB Pen 32GB	2025-10-02 09:10:45
776859	USB Pen 128GB	2025-10-03 10:12:10

(onde parar) 2025-10-03 14:15:20

DATA STAGING AREA

T_INFO_EXTRACTIONS

source_table_name	last_timestamp
T_PRODUCTS	2025-10-03 14:15:20
...	...

atualizar

T_DATA_PRODUCTS

id	name
233324	Pen Drive 64GB
776859	USB Pen 128GB

Técnicas CDC: *Triggers*

- Exemplo
  - O que é que acontece aos dados depois de extraídos para a DSA (2ª extração)?
    - Tabelas T_DATA_*
  - São transformados
    - Tabelas T_CLEAN_*
  - São carregados para o DW
    - O DW já tem os dados da 1ª extração
    - Tabelas T_DIM_* e T_FACT_*

Técnicas CDC

- Partições
  - As tabelas de dados são divididas em partições
  - Cada partição representa um horizonte temporal
    - 1 partição = 1 dia
  - Técnica de extração incremental

Extração de Dados

- Sumário
  - Introdução
  - Processo de Extração
  - Técnicas CDC
    - Técnicas de Extração Incremental
    - Técnicas de Extração Completa

Técnicas CDC

- Processo de Eliminação
  - Em cada extração todos os dados fonte são extraídos
  - Retém uma cópia da última extração na área de tratamento de dados (DSA)
  - Os dados são comparados e as diferenças são depois transformadas e carregadas para o *data warehouse*
  - Técnica de extração completa

Técnicas CDC

- Processo de Eliminação
  - Utilizado na extração de dados a partir de ficheiros
    - Ficheiros CSV, JSON, etc.
  - Pode ser utilizado também para extrair dados de tabelas de BDs operacionais
    - Na impossibilidade de aplicar as técnicas de extração incremental
    - Exportação de dados das tabelas para ficheiros

Técnicas CDC

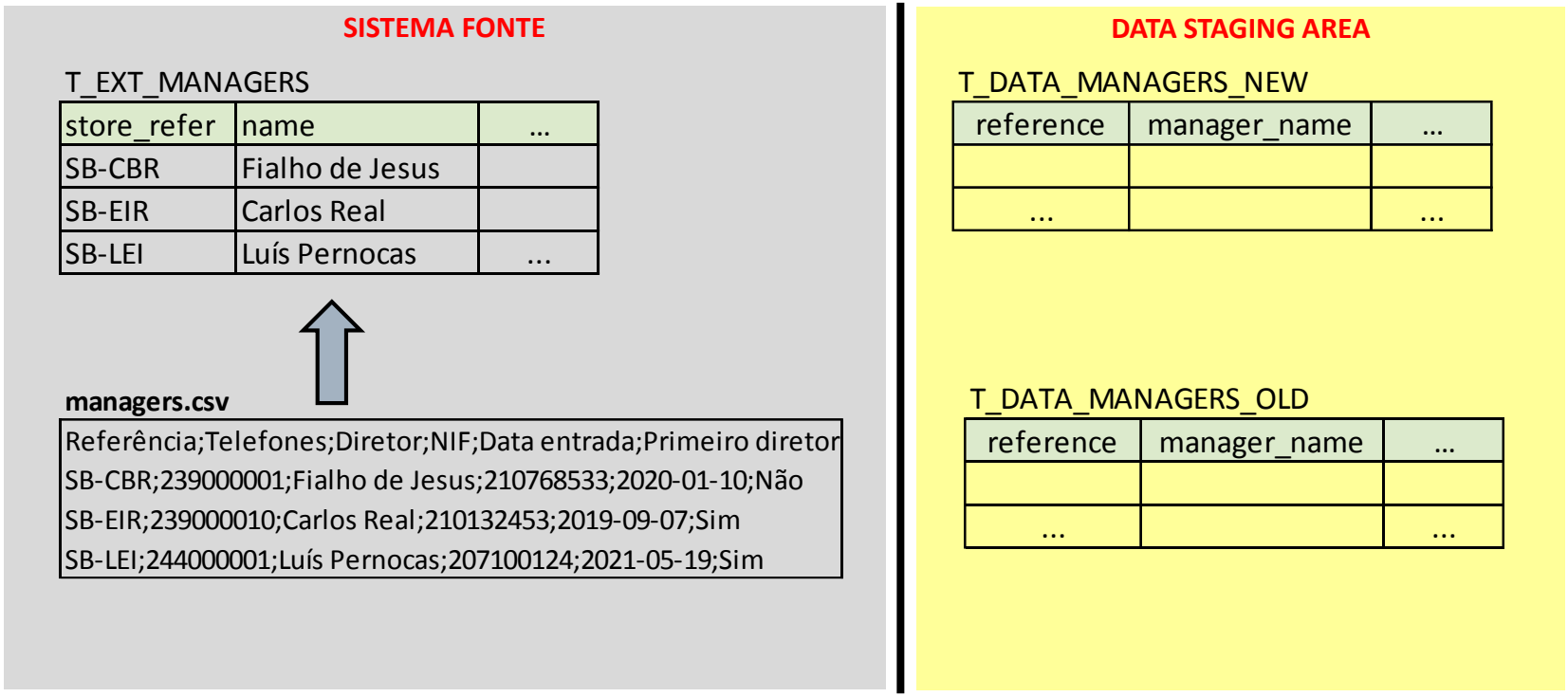
- Processo de Eliminação
  - Implementação
    - Criar duas tabelas na DSA
      - *table_new* e *table_old*
    - Os dados extraídos da fonte vão para a *table_new*
    - Os dados da extração anterior estão na *table_old*
    - Se necessário, criar objeto adicional para fazer a ligação entre o objeto fonte e as tabelas da DSA

Técnicas CDC

- Processo de Eliminação
  - Passos típicos do algoritmo
    1. Limpar a *table_old*
    2. Copiar os dados da *table_new* para a *table_old*
    3. Limpar a *table_new*
    4. Extrair os dados da fonte para a *table_new*

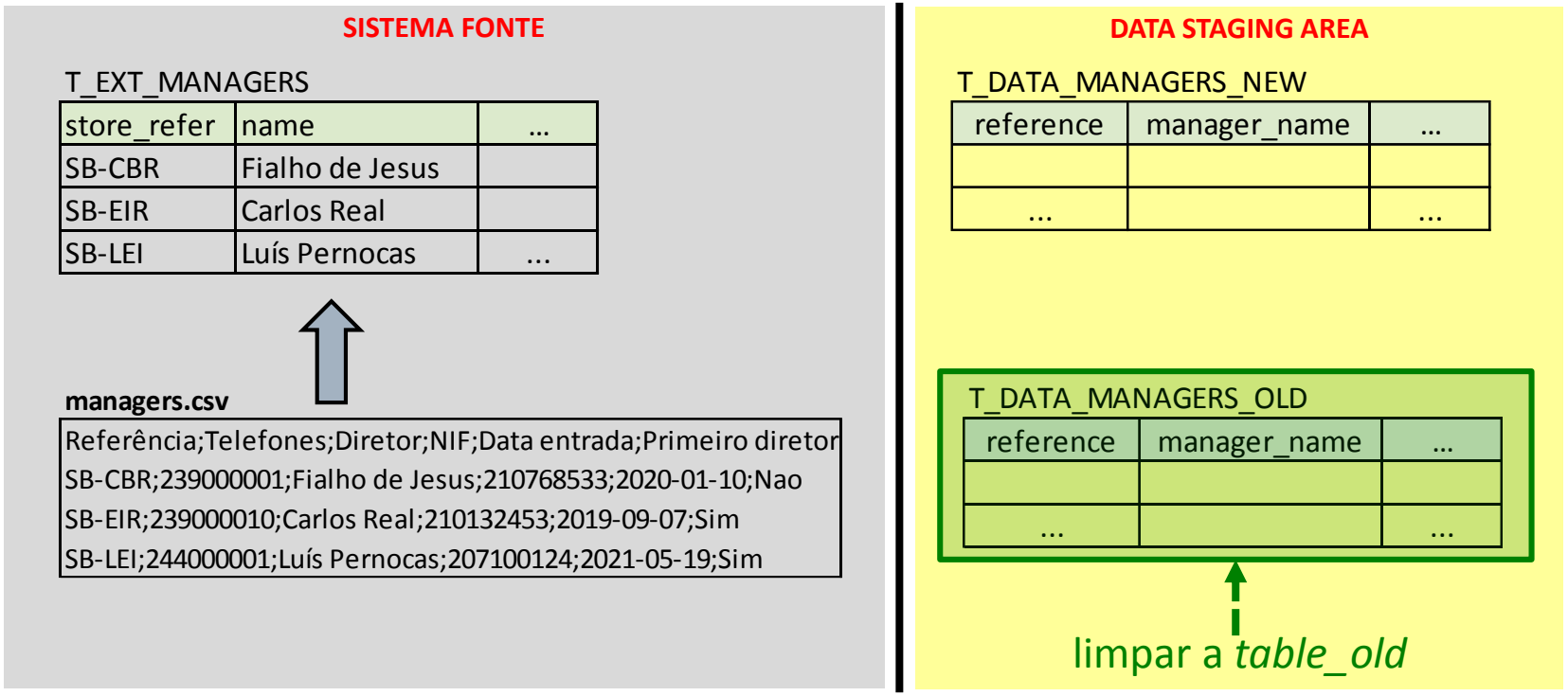
Técnicas CDC

- Processo de Eliminação: Exemplo
  - Antes da 1ª Extração: Contexto



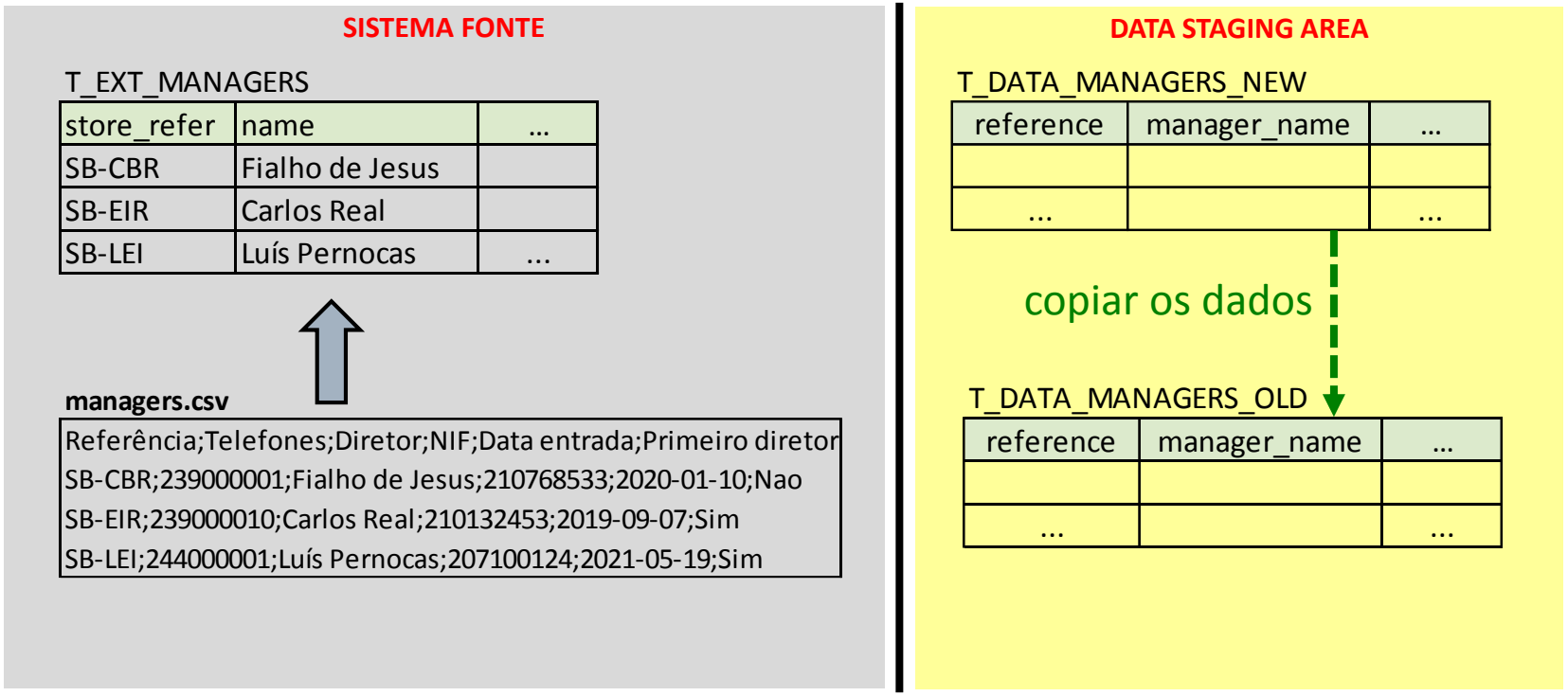
Técnicas CDC

- Processo de Eliminação: Exemplo
  - 1ª Extração: Passo 1



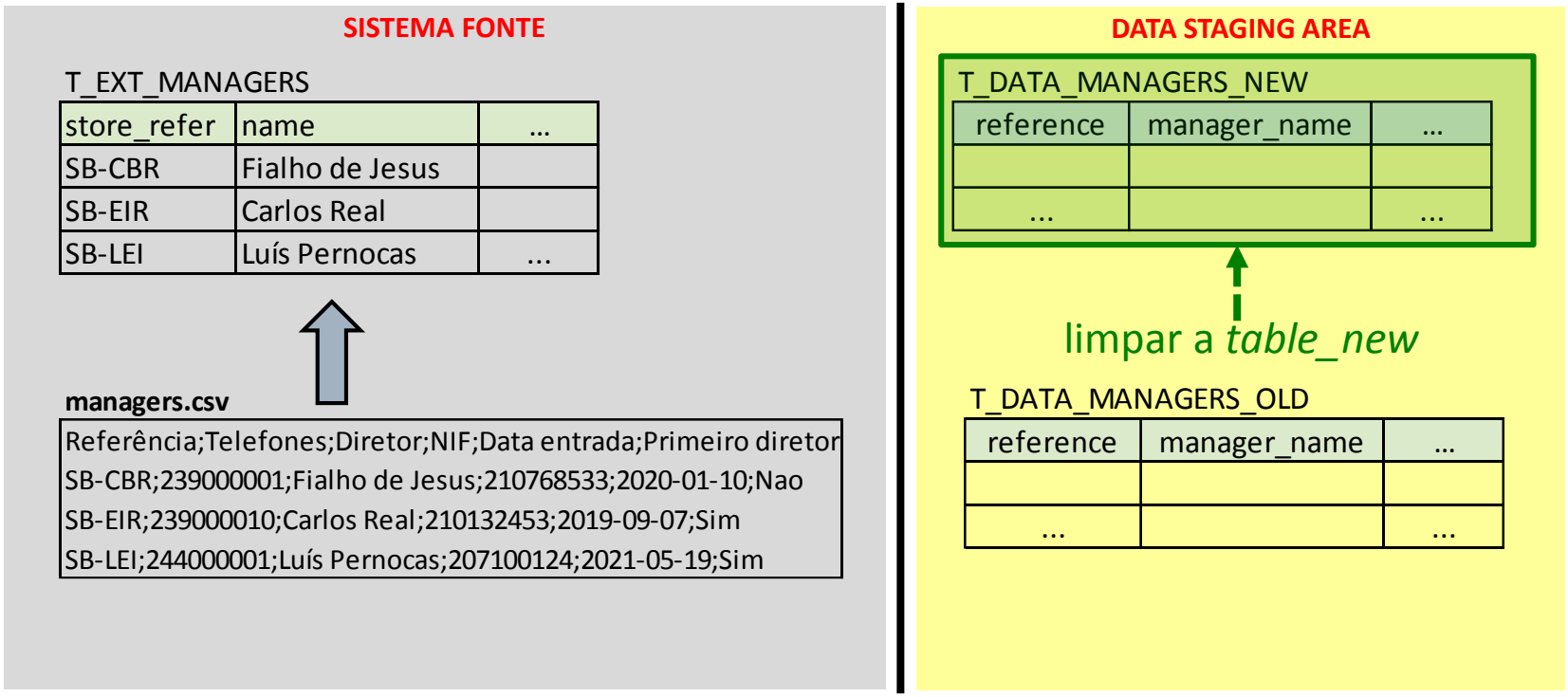
Técnicas CDC

- Processo de Eliminação: Exemplo
  - 1ª Extração: Passo 2



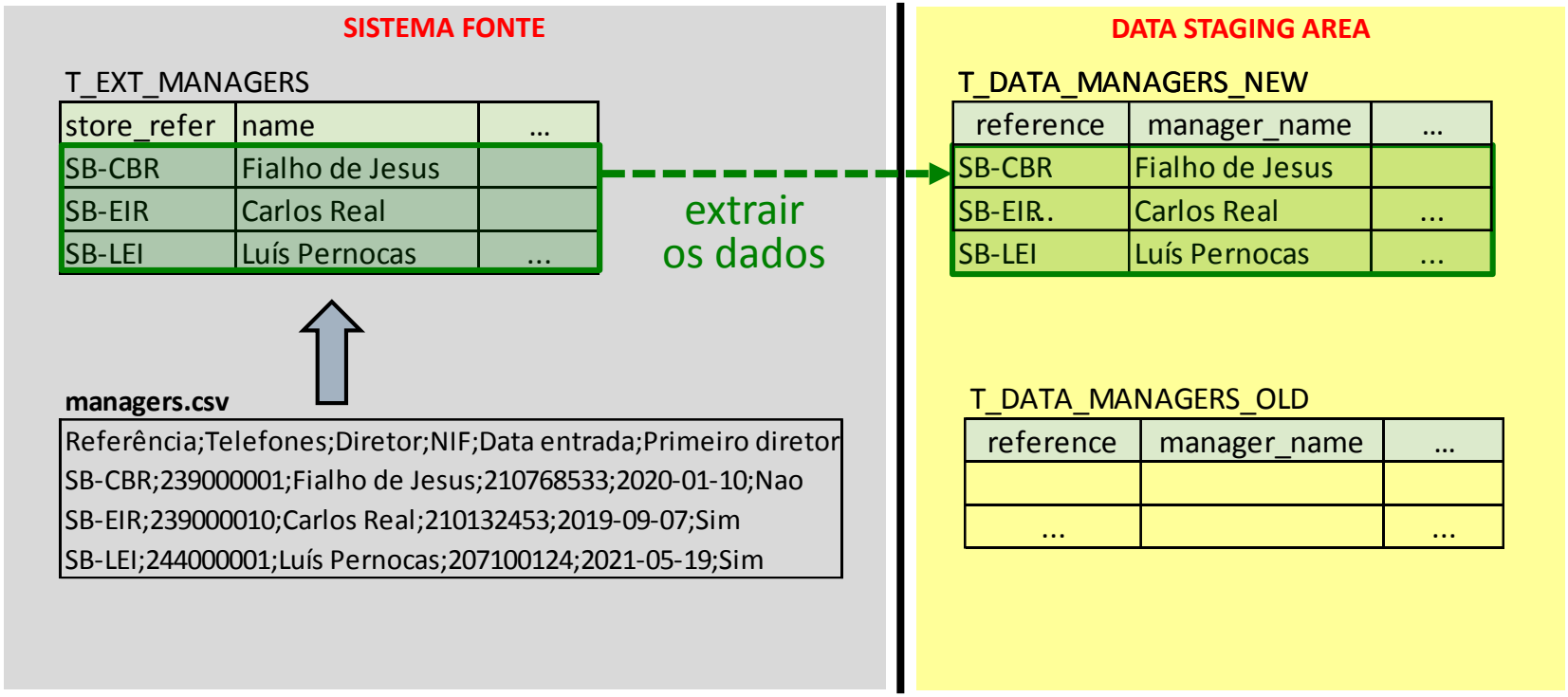
Técnicas CDC

- Processo de Eliminação: Exemplo
  - 1ª Extração: Passo 3



Técnicas CDC

- Processo de Eliminação: Exemplo
  - 1ª Extração: Passo 4

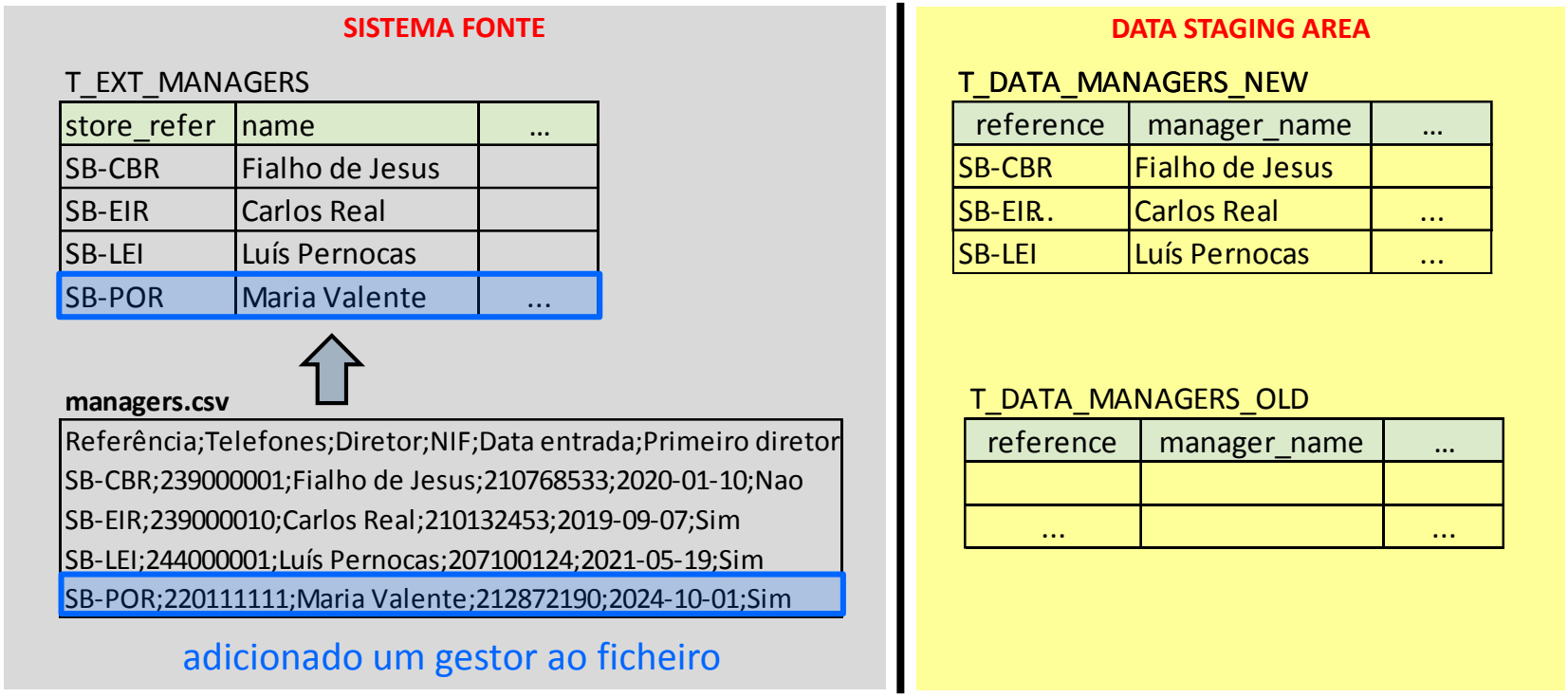


Técnicas CDC

- Processo de Eliminação: Exemplo
  - O que é que acontece aos dados depois de extraídos para a DSA (1ª extração)?
    - Tabelas T_DATA_*
  - São transformados
    - MINUS das tabelas *_NEW e *_OLD
    - Tabelas T_CLEAN_*
  - São carregados para o DW
    - Tabelas T_DIM_* e T_FACT_*
    - Antes do 1º carregamento as tabelas do DW estão vazias

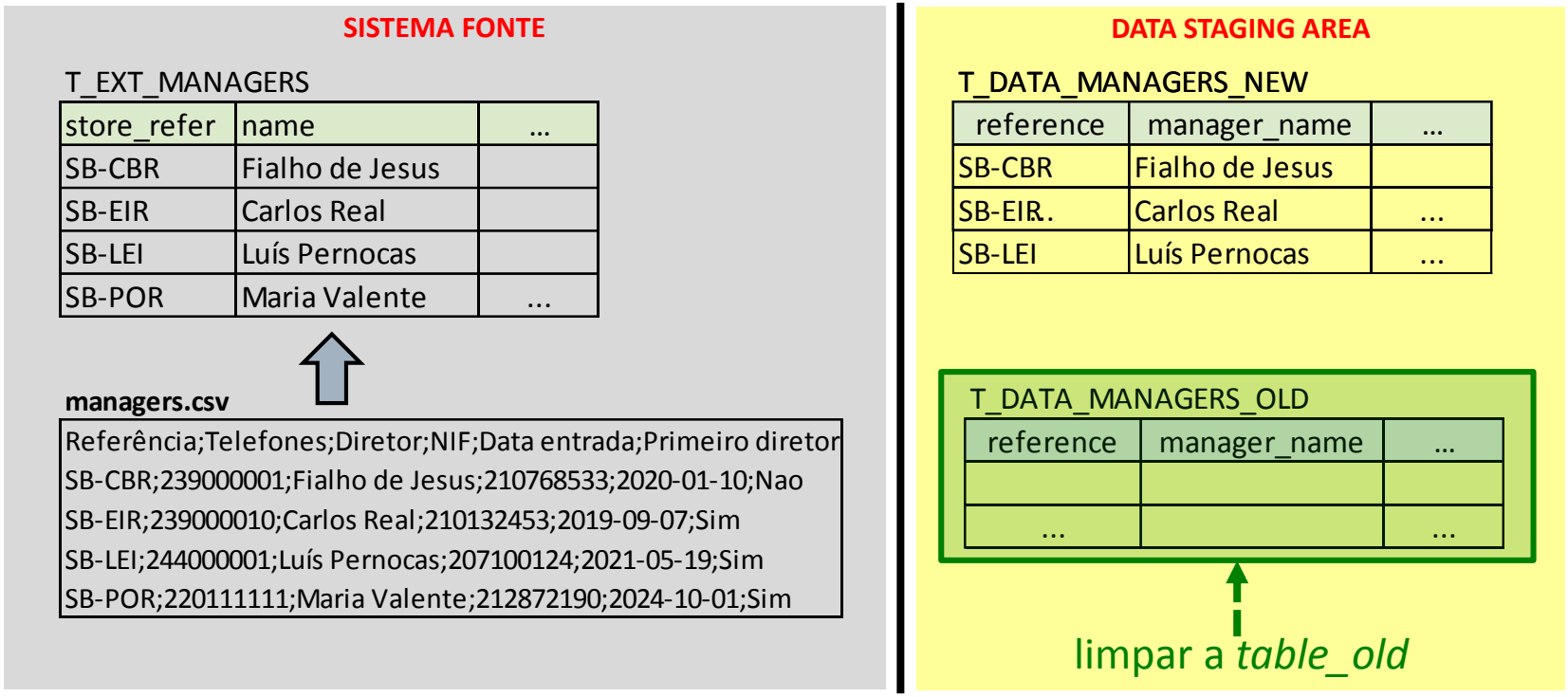
Técnicas CDC

- Processo de Eliminação: Exemplo
  - Alterações no Sistema Fonte



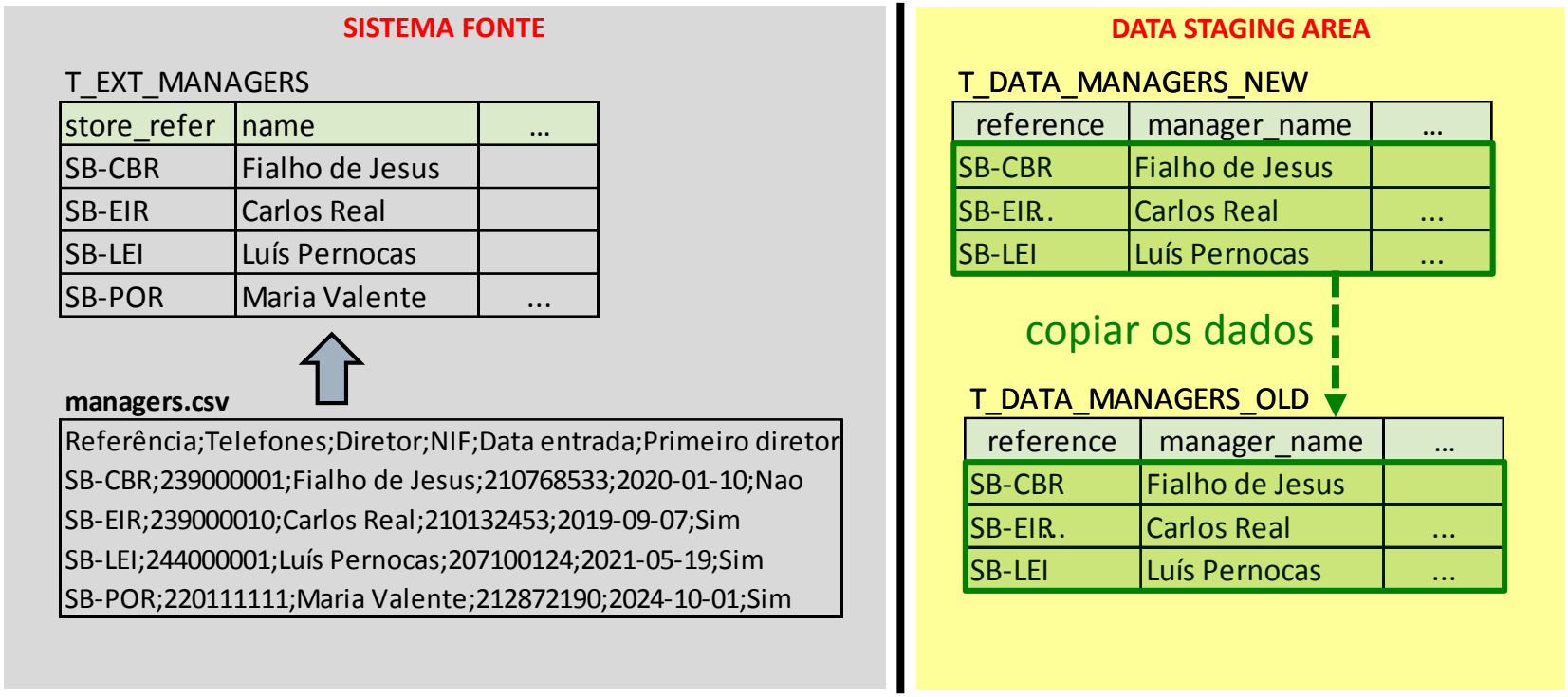
Técnicas CDC

- Processo de Eliminação: Exemplo
  - 2ª Extração: Passo 1



Técnicas CDC

- Processo de Eliminação: Exemplo
  - 2ª Extração: Passo 2



Técnicas CDC

- Processo de Eliminação: Exemplo
  - 2ª Extração: Passo 3

SISTEMA FONTE

T_EXT_MANAGERS

store_refer	name	…
SB-CBR	Fialho de Jesus	
SB-EIR	Carlos Real	
SB-LEI	Luís Pernocas	
SB-POR	Maria Valente	…

managers.csv

```
Referência;Telefones;Diretor;NIF;Data entrada;Primeiro diretor
SB-CBR;239000001;Fialho de Jesus;210768533;2020-01-10;Nao
SB-EIR;239000010;Carlos Real;210132453;2019-09-07;Sim
SB-LEI;244000001;Luís Pernocas;207100124;2021-05-19;Sim
SB-POR;220111111;Maria Valente;212872190;2024-10-01;Sim
```

DATA STAGING AREA

T_DATA_MANAGERS_NEW

reference	manager_name	…
SB-CBR	Fialho de Jesus	
SB-EIR	Carlos Real	…
SB-LEI	Luís Pernocas	…

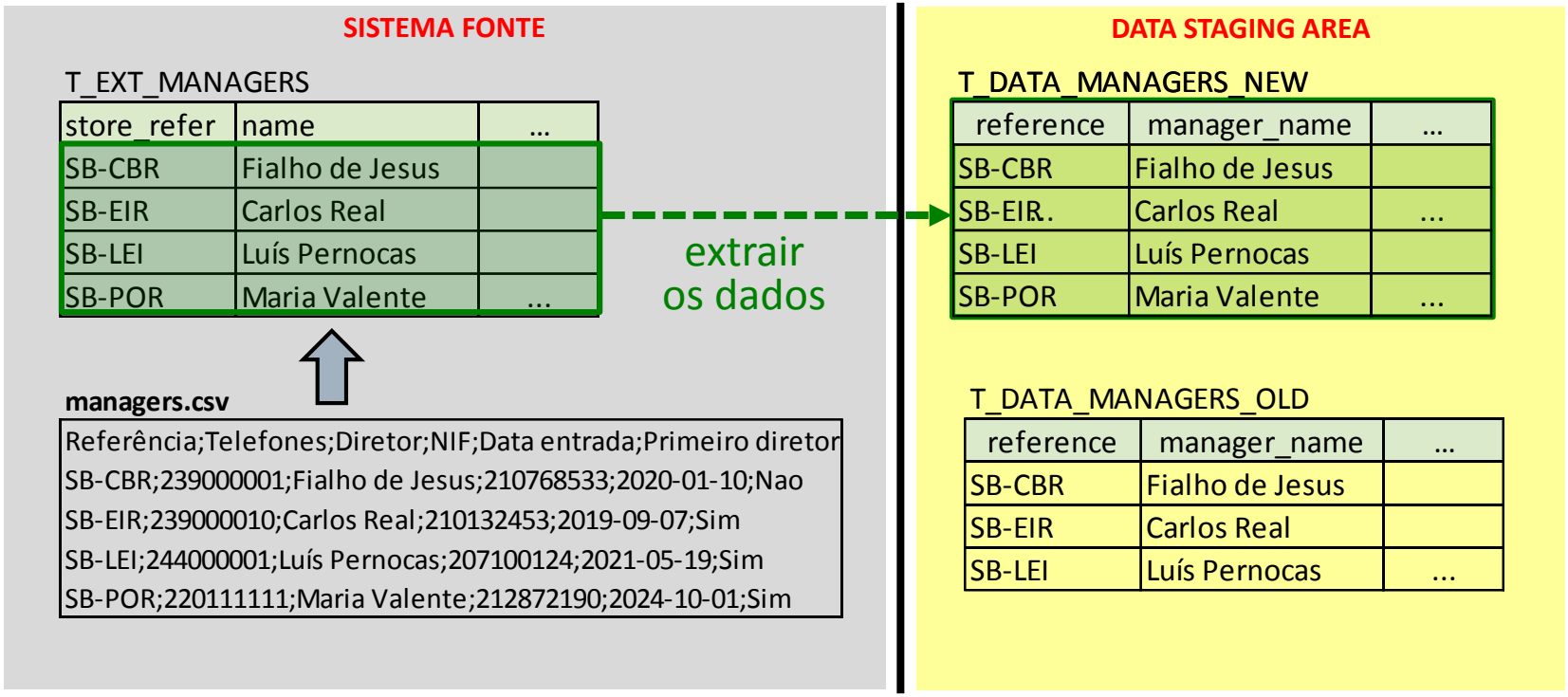
limpar a *table_new*

T_DATA_MANAGERS_OLD

reference	manager_name	…
SB-CBR	Fialho de Jesus	
SB-EIR	Carlos Real	
SB-LEI	Luís Pernocas	…

Técnicas CDC

- Processo de Eliminação: Exemplo
  - 2ª Extração: Passo 4



Técnicas CDC

- Processo de Eliminação: Exemplo
  - O que é que acontece aos dados depois de extraídos para a DSA (2ª extração)?
    - Tabelas T_DATA_*
  - São transformados
    - MINUS das tabelas *_NEW e *_OLD
    - Tabelas T_CLEAN_*
  - São carregados para o DW
    - O DW já tem os dados da 1ª extração
    - Tabelas T_DIM_* e T_FACT_*

Extração de Dados

- Referências
  - The Data Warehouse ETL Toolkit, R. Kimball e J. Caserta, John Wiley & Sons, 2004
    - Capítulos 1, 2, 3
  - Sistemas de Suporte à Decisão, B. Cortes, FCA, 2005
    - Capítulo 3