Carregamento de Dados

Sistemas de Apoio à Decisão

Carregamento de Dados

- Sumário
  - Introdução
  - Processo de Carregamento
  - Técnicas SCD
  - Passos Típicos do Processo ETL

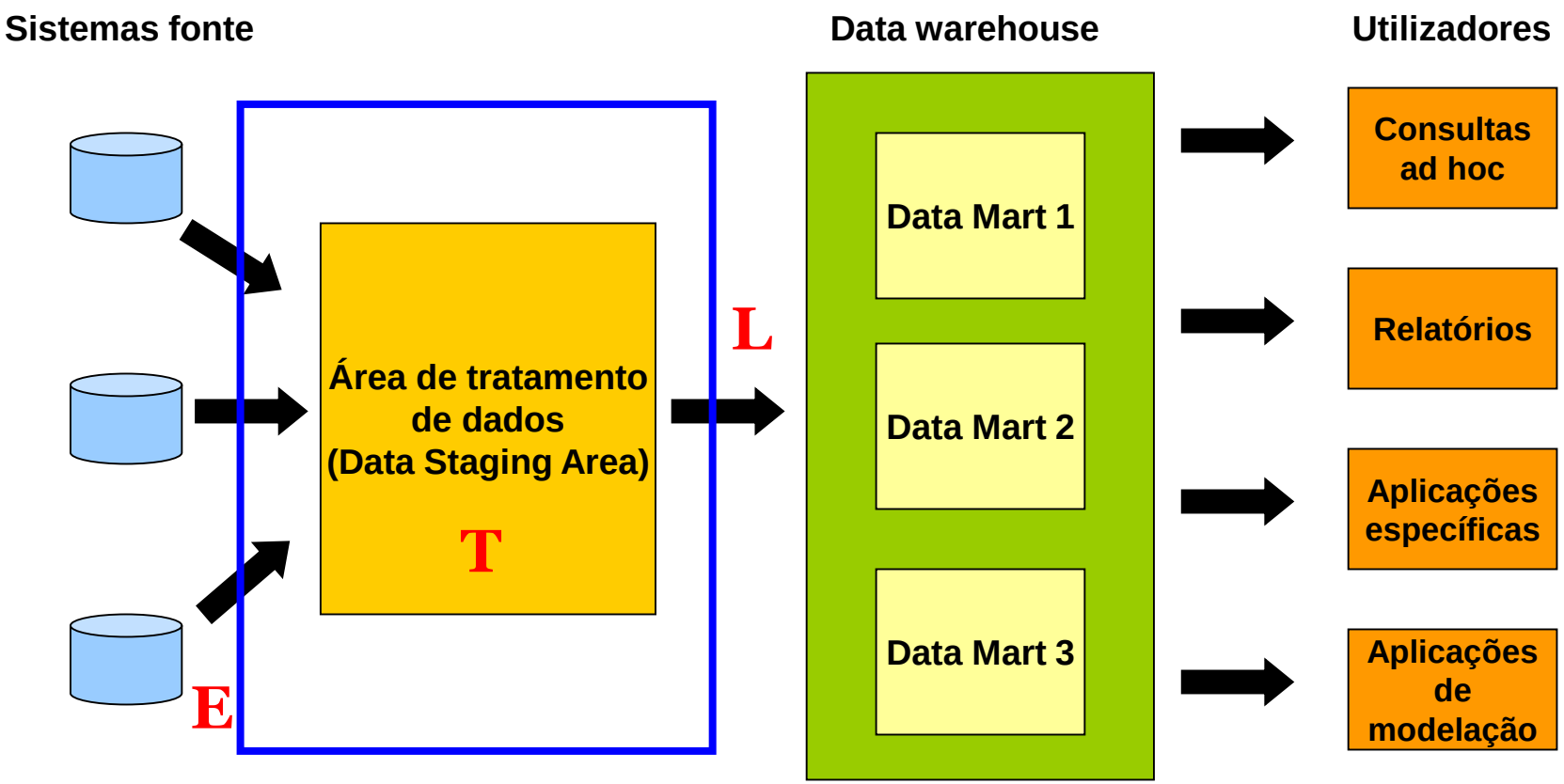
Introdução

- Processo ETL
  - Permite migrar dados dos sistemas fonte para o *Data Warehouse*, procedendo às necessárias transformações
    - Formato e conteúdo
  - 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*

Carregamento de Dados

- Sumário
  - Introdução
  - Processo de Carregamento
  - Técnicas SCD
  - Passos Típicos do Processo ETL

Processo de Carregamento



Processo de Carregamento

- Introdução
  - Depois de transformados, é necessário carregar os dados para a BD do *Data Warehouse*
  - Geralmente são carregados muitos registos de uma só vez
    - Técnicas de *bulk loading*
  - Ordem do carregamento
    - Tabelas de minidimensão
    - Tabelas de dimensão
    - Tabelas de factos

Processo de Carregamento

- Introdução
  - Criação de chaves primárias independentes das chaves utilizadas nos sistemas fonte
  - Criação de registos especiais para situações de exceção
    - Evitar a interrupção do carregamento
  - Construção de agregados de modo a acelerar as pesquisas
  - Depois de carregados, os dados são indexados

Processo de Carregamento

- Carregamento das Tabelas
  - É uma fase crítica em que eventuais falhas podem levar a recuperações complexas
  - Quase tudo o que é feito para otimizar o desempenho do *Data Warehouse* tende a atrasar o carregamento:
    - Índices
    - Agregados
    - Particionamento de tabelas, …

Processo de Carregamento

- Carregamento Inicial
  - Disponibilização no *Data Warehouse* dos dados extraídos das fontes e validados na DSA
  - Geralmente o primeiro carregamento corre sempre bem
  - Importa minimizar ao máximo a janela de carregamento

Processo de Carregamento

- Carregamentos Periódicos
  - Para além do carregamento inicial é necessário resolver os carregamentos periódicos, com características diferentes
    - Atualizações de dimensões
    - Agregados, etc.
  - Questões a considerar:
    - Duração estimada do carregamento
    - Impacto na coerência do *Data Warehouse* caso o processo tenha de ser interrompido

Carregamento de Dados

- Sumário
  - Introdução
  - Processo de Carregamento
  - Técnicas SCD
  - Passos Típicos do Processo ETL

Técnicas SCD

- Introdução
  - Necessidade de preservar o histórico de dados num *Data Warehouse*
  - Os atributos das dimensões podem mudar ao longo do tempo
  - Necessária uma estratégia para lidar com as alterações
    - Técnicas SCD (*Slowly Changing Dimensions*)

Técnicas SCD

- Introdução
  - Técnicas para lidar com as alterações nas dimensões
    - Tipo 0 (SCD0)
    - Tipo 1 (SCD1)
    - Tipo 2 (SCD2)
    - Tipo 3 (SCD3)
    - Tipo 4 (SCD4)
    - Tipo 6 (SCD6)

Técnicas SCD

- Alterações Tipo 1 (SCD1)
  - Consiste em alterar diretamente o valor de um ou mais campos
    - Sobreposição de valores
  - É útil, sobretudo quando se detetam erros nos dados
  - Nem sempre os dados estão disponíveis quando se cria um registo numa dimensão

Técnicas SCD

- Alterações Tipo 1 (SCD1)
  - Exemplo

SISTEMA FONTE

T_CLIENTES

ID	Nome	Telemóvel	…
12345	João Lima	991234567	
54321	Maria Violeta	981234567	
…	…		…

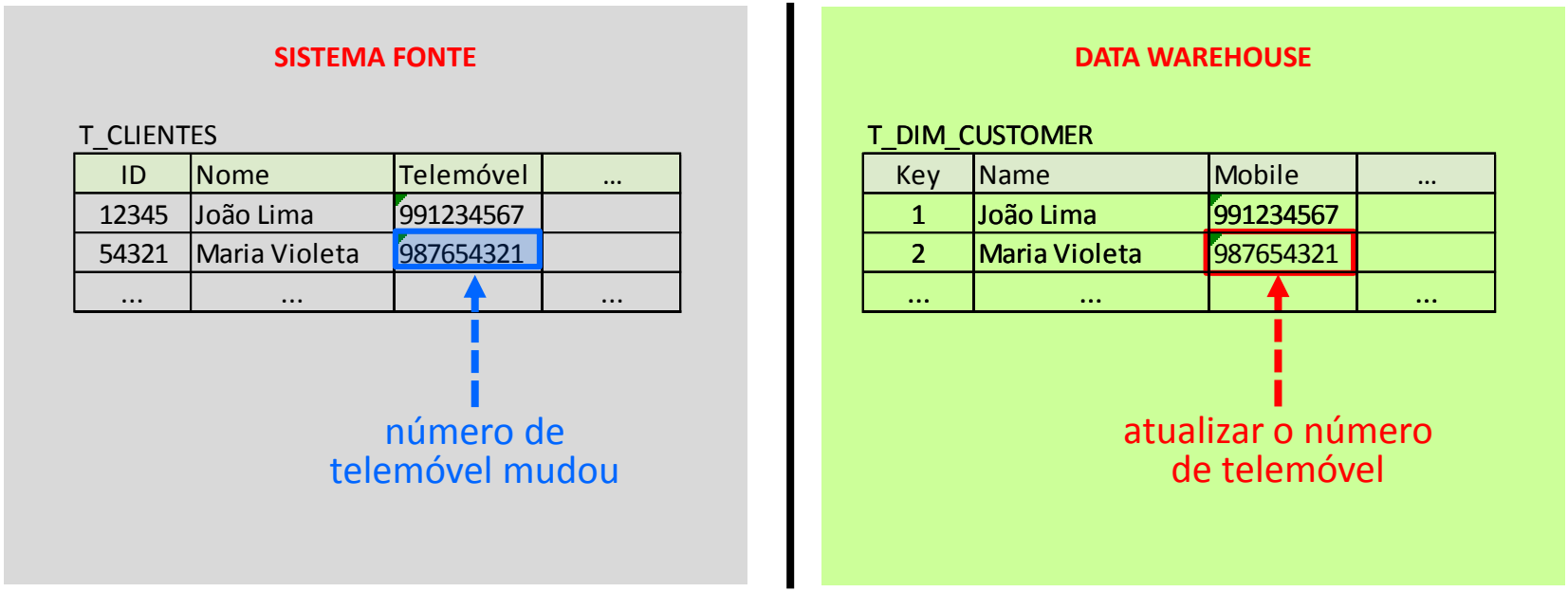
DATA WAREHOUSE

T_DIM_CUSTOMER

Key	Name	Mobile	…
1	João Lima	991234567	
2	Maria Violeta	981234567	
…	…		…

Técnicas SCD

- Alterações Tipo 1 (SCD1)
  - Exemplo



Técnicas SCD

- Alterações Tipo 2 (SCD2)
  - Criação de novo registo na dimensão com a alteração pretendida
  - Necessária uma coluna adicional para indicar qual a versão ativa do registo
  - Permite preservar todo o histórico de dados

Técnicas SCD

- Alterações Tipo 2 (SCD2)
  - Exemplo

SISTEMA FONTE

T_CLIENTES

ID	Nome	Telemóvel	…
12345	João Lima	991234567	
54321	Maria Violeta	981234567	
…	…		…

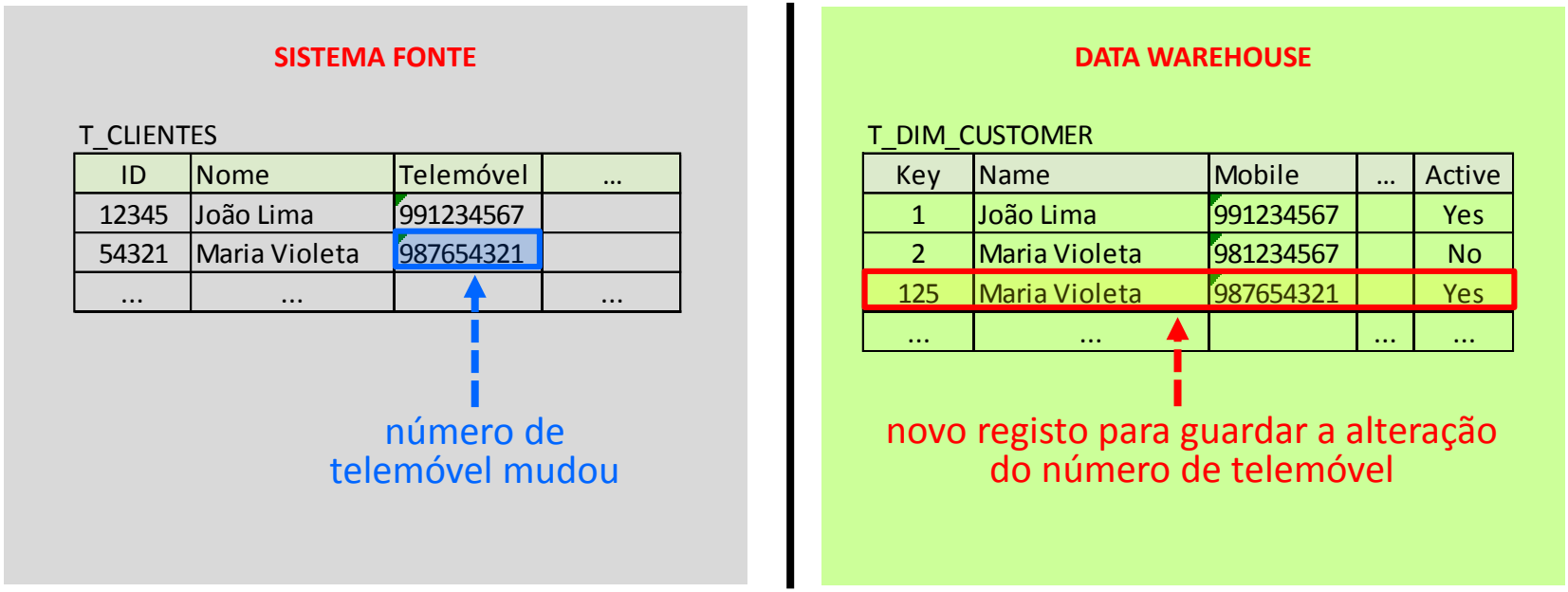
DATA WAREHOUSE

T_DIM_CUSTOMER

Key	Name	Mobile	…
1	João Lima	991234567	
2	Maria Violeta	981234567	
…	…		…

Técnicas SCD

- Alterações Tipo 2 (SCD2)
  - Exemplo



Técnicas SCD

- Alterações Tipo 3 (SCD3)
  - Criação de um campo adicional que permitem registar a alteração do valor do atributo
  - Permite apenas guardar o valor atual e a alteração anterior ou o valor inicial
  - Não é preservado todo o histórico de dados

Técnicas SCD

- Alterações Tipo 3 (SCD3)
  - Exemplo

SISTEMA FONTE

T_CLIENTES

ID	Nome	Telemóvel	…
12345	João Lima	991234567	
54321	Maria Violeta	981234567	
…	…		…

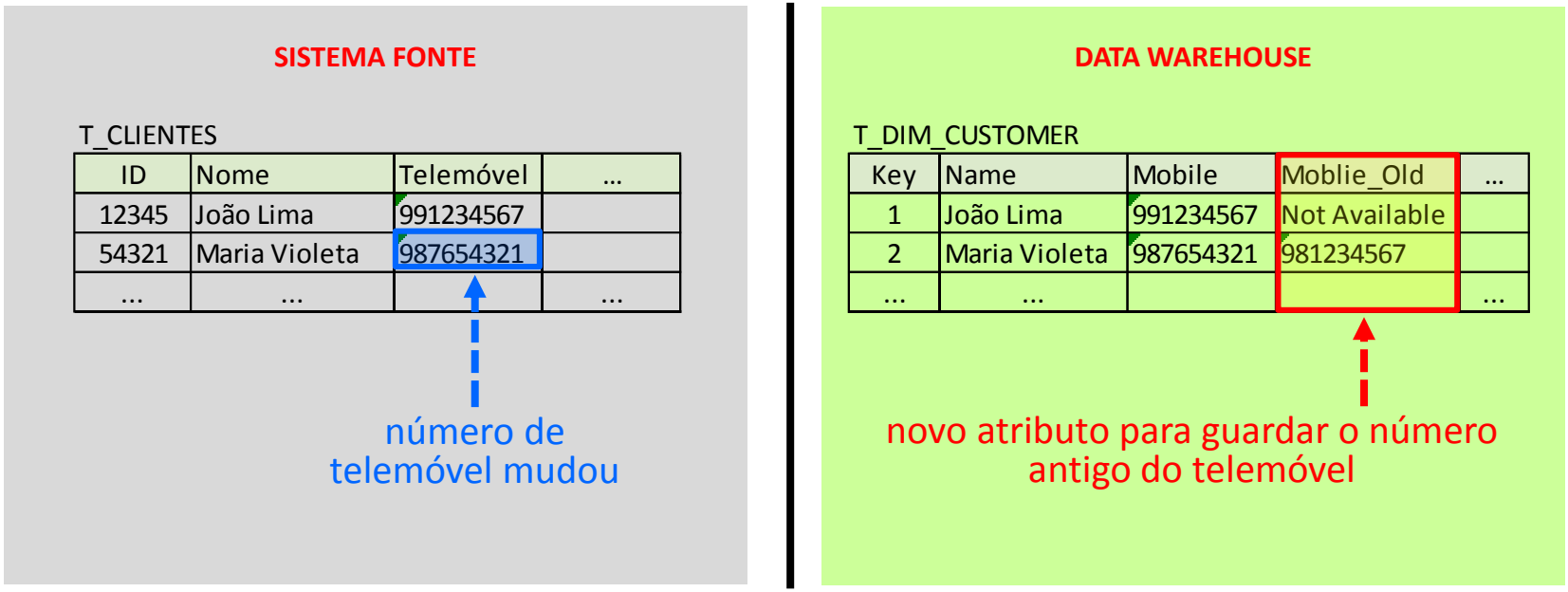
DATA WAREHOUSE

T_DIM_CUSTOMER

Key	Name	Mobile	…
1	João Lima	991234567	
2	Maria Violeta	981234567	
…	…		…

Técnicas SCD

- Alterações Tipo 3 (SCD3)
  - Exemplo



Carregamento de Dados

- Sumário
  - Introdução
  - Processo de Carregamento
  - Técnicas SCD
  - Passos Típicos do Processo ETL

Processo ETL

- Passos Típicos
  - Planeamento
  - Carregamento de dimensões
  - Carregamento de factos
  - Automatizar o processo ao máximo
  - Infraestrutura para a área de tratamento de dados
  - Carregamento inicial e periódicos
  - Administração

Processo ETL

- Planeamento
  - Definir um plano geral (tipo *end-to-end*)
    - Mapa Lógico de Dados
  - Definir infraestrutura para a área de tratamento de dados
  - Escolher as ferramentas de ETL
  - Fazer plano detalhado analisando todos os problemas que é necessário resolver para carregar cada tabela destino
    - Fontes, transformações, etc.

Processo ETL

- Carregamento de Dimensões
  - Elaborar, testar e executar planos ETL para as dimensões estáticas e simples
    - Permite testar toda a infraestrutura
  - Elaborar, testar e executar planos ETL para as dimensões que mudam
  - Tratar todos os restantes casos
    - Dimensões geradas com dados manuais, dimensões especiais, etc.

Processo ETL

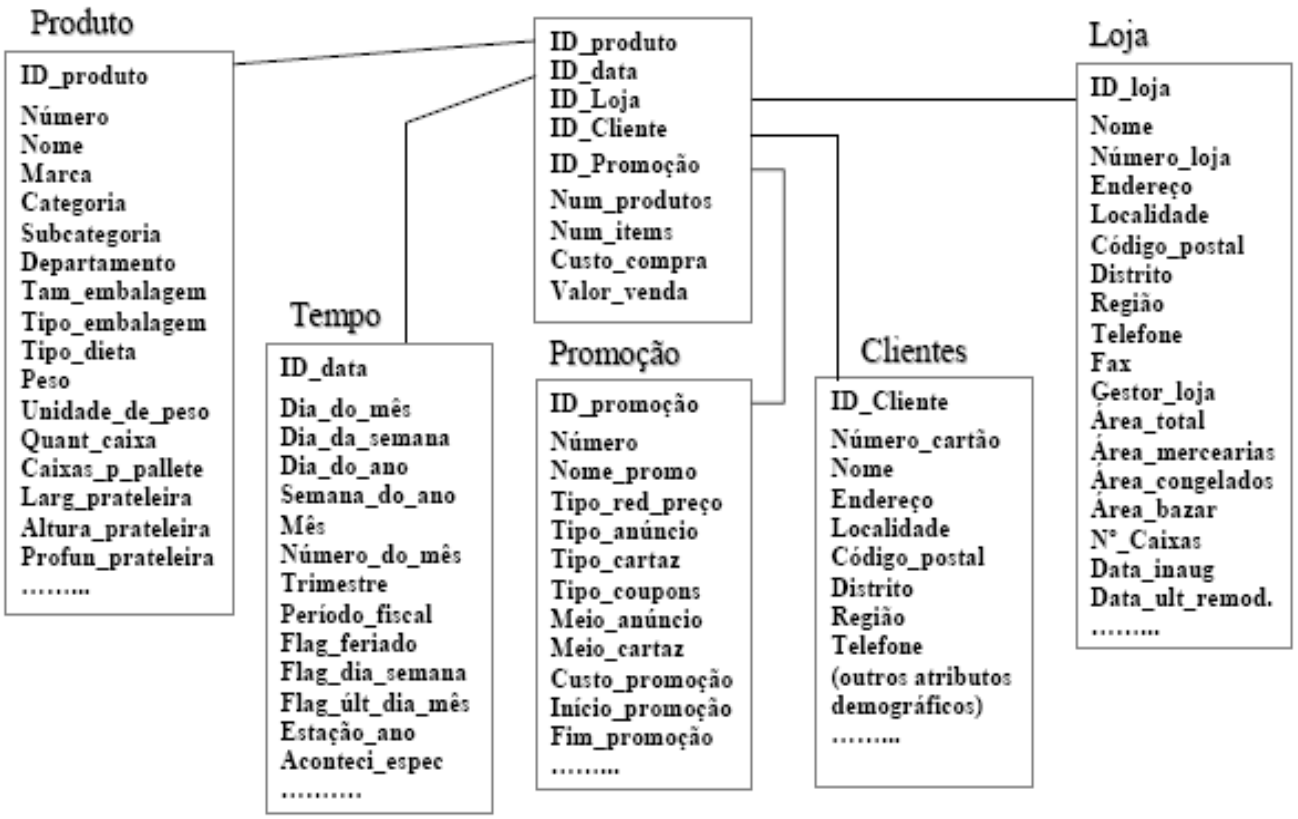
- Carregamento de Factos
  - Elaborar, testar e executar planos ETL para tabelas de factos
  - Elaborar e testar processo de carregamentos periódicos

Processo ETL

- Automatizar o Processo
  - Utilização de ferramentas sofisticadas de suporte
    - Escalonamento dos processos
    - Execução automática

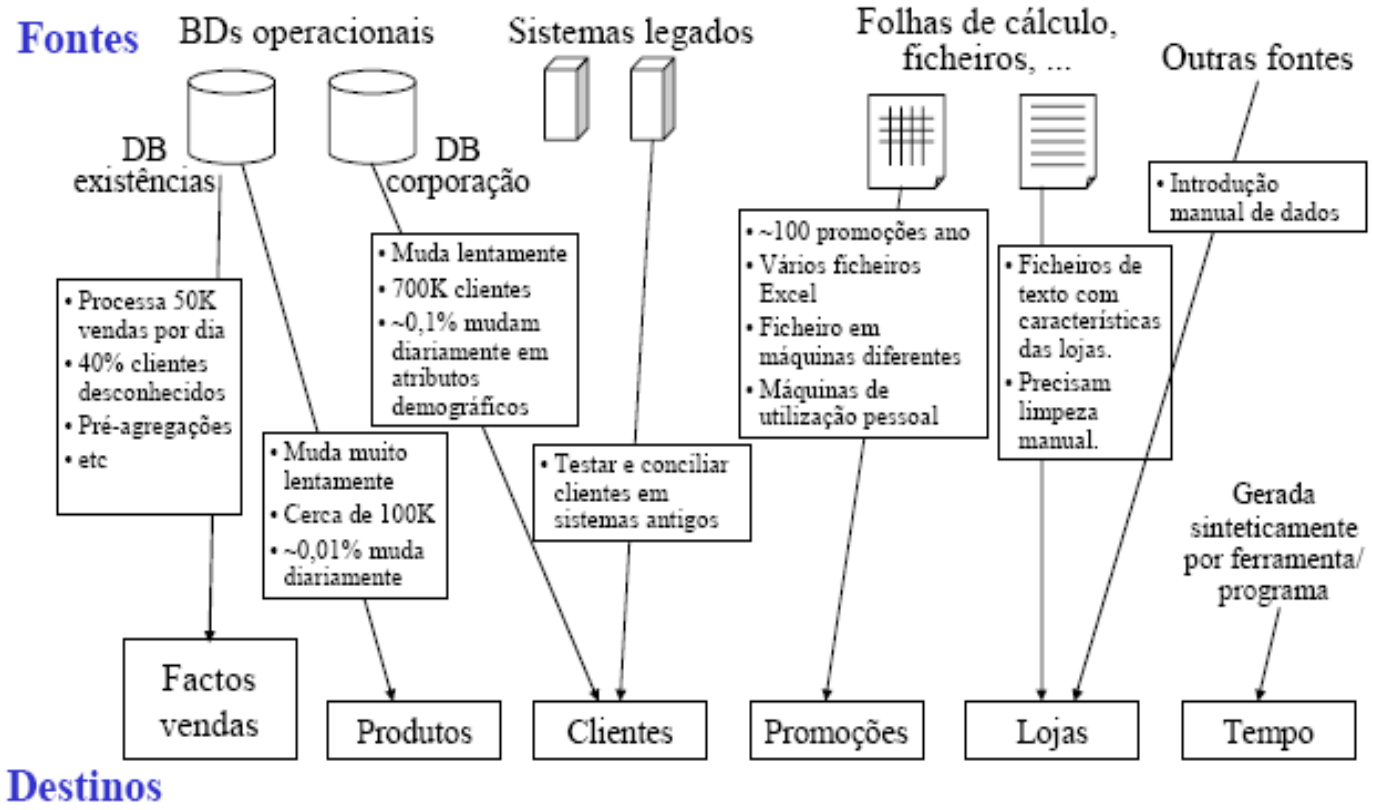
Processo ETL

- Exemplo: Cadeia de Lojas



Processo ETL

- Exemplo: Cadeia de Lojas



Processo ETL

- Infraestrutura da DSA
  - Pode ir de uma simples conta no servidor do *Data Warehouse* a máquinas dedicadas de grande capacidade
  - A decisão depende do volume de dados e da complexidade das operações a fazer nos dados antes de os carregar
  - Tipicamente, para cada dimensão e tabela de factos, prepara-se tudo na área de tratamento de dados para depois fazer um carregamento direto

Processo ETL

- Carregamento Inicial
  - Feito diretamente da área de tratamento de dados para as tabelas do *Data Warehouse* (depois dos dados preparados)
  - Alguns cuidados:
    - Desligar sistemas de *logging*
    - Ordenar previamente os dados a carregar pela chave primária
    - Fazer, eventualmente, algumas agregações básicas durante o carregamento

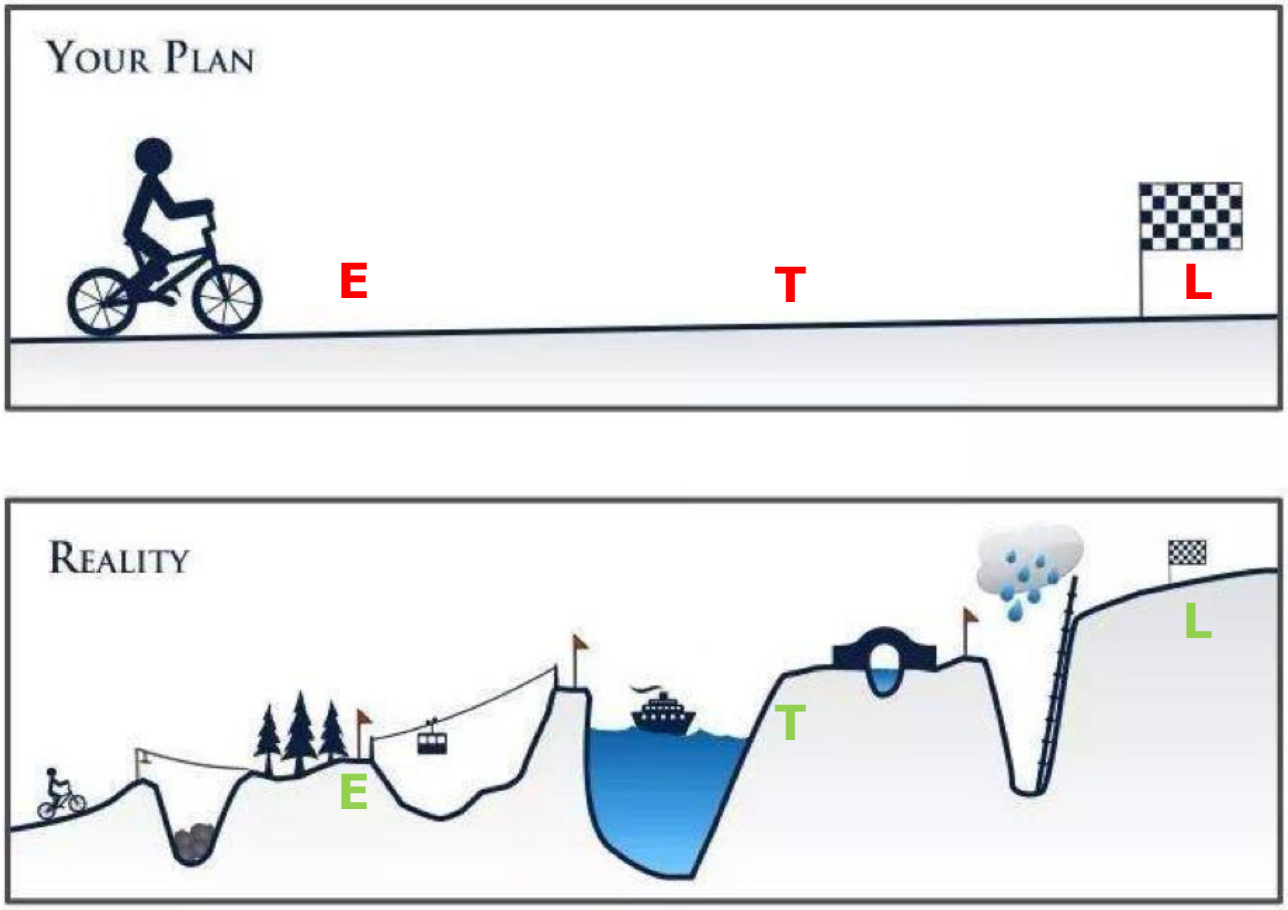
Processo ETL

- Carregamentos Periódicos
  - Definir estratégia para identificar novos dados nos sistemas fonte
    - Novas transações
    - Atualizações a dados de transações anteriores
  - Identificar:
    - Registos novos para cada dimensão
    - Atualizações de atributos de dimensões e como estas vão ser tratadas
    - Novos factos ou medidas numéricas

Processo ETL

- Administração
  - Construir, utilizar e manter as ferramentas de extração de dados
  - Garantir a qualidade dos dados, após cada extração
  - Construir e manter agregados
  - Vigiar e afinar o desempenho do sistema
  - Fazer cópias de segurança e recuperar o estado da BD do *Data Warehouse* em caso de falha

Processo ETL



Carregamento de Dados

- Referências
  - The Data Warehouse ETL Toolkit, R. Kimball e J. Caserta, John Wiley & Sons, 2004
    - Capítulos 1, 2, 5 e 6
  - Sistemas de Suporte à Decisão, B. Cortes, FCA, 2005
    - Capítulo 3