Processo ETL

Sistemas de Apoio à Decisão

Processo ETL

- Sumário
  - Introdução
  - Extração de Dados
  - Transformação de Dados
  - Carregamento de Dados
  - Mapa Lógico de Dados
  - Passos Típicos do Processo ETL

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*

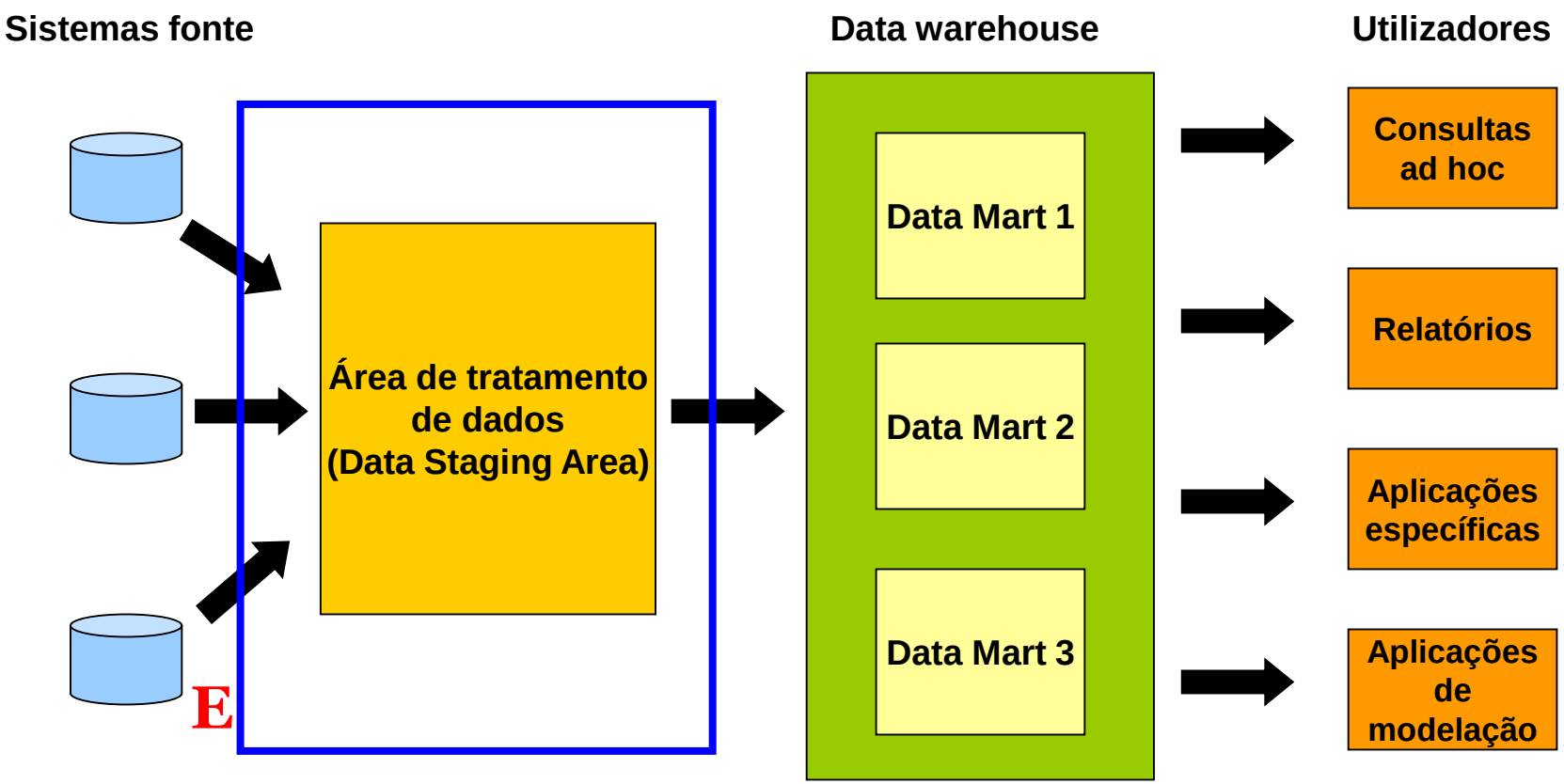
Introdução

- 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

Processo ETL

- Sumário
  - Introdução
  - Extração de Dados
  - Transformação de Dados
  - Carregamento de Dados
  - Mapa Lógico de Dados
  - Passos Típicos do Processo ETL

Arquitetura do *Data Warehouse*



Extração de Dados

- Introdução
  - A extração 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

Extração de Dados

- 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

Extração de Dados

- 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

Extração de Dados

- 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
      - 29-11-2021
      - 2021/11/29
      - novembro 29, 2021, etc.

Extração de Dados

- 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:
      - *Mainframes*
      - Ficheiros
      - Fontes XML
      - Web logs
      - ERP's, ...

Extração de Dados

- Extração de dados que mudam
  - CDC – *Change Data Capture*
  - No primeiro carregamento 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

Extração de Dados

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

Extração de Dados

- Técnicas CDC: *Timestamps*
  - É feita a adição de uma coluna na qual é registada a hora/data da alteração de cada registo nos sistemas operacionais
  - Para preencher a coluna são utilizados *triggers* que disparam automaticamente sempre que são inseridos ou atualizados registos
  - Técnica de extração incremental

Extração de Dados

- 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

- Técnicas CDC: Processo de eliminação
  - Retém uma cópia da última extração na área de tratamento de dados (DSA)
  - Na próxima extração todos os dados fonte são carregados
  - Os dados são comparados e as diferenças são transformadas e carregadas para o *data warehouse*
  - Técnica de extração completa

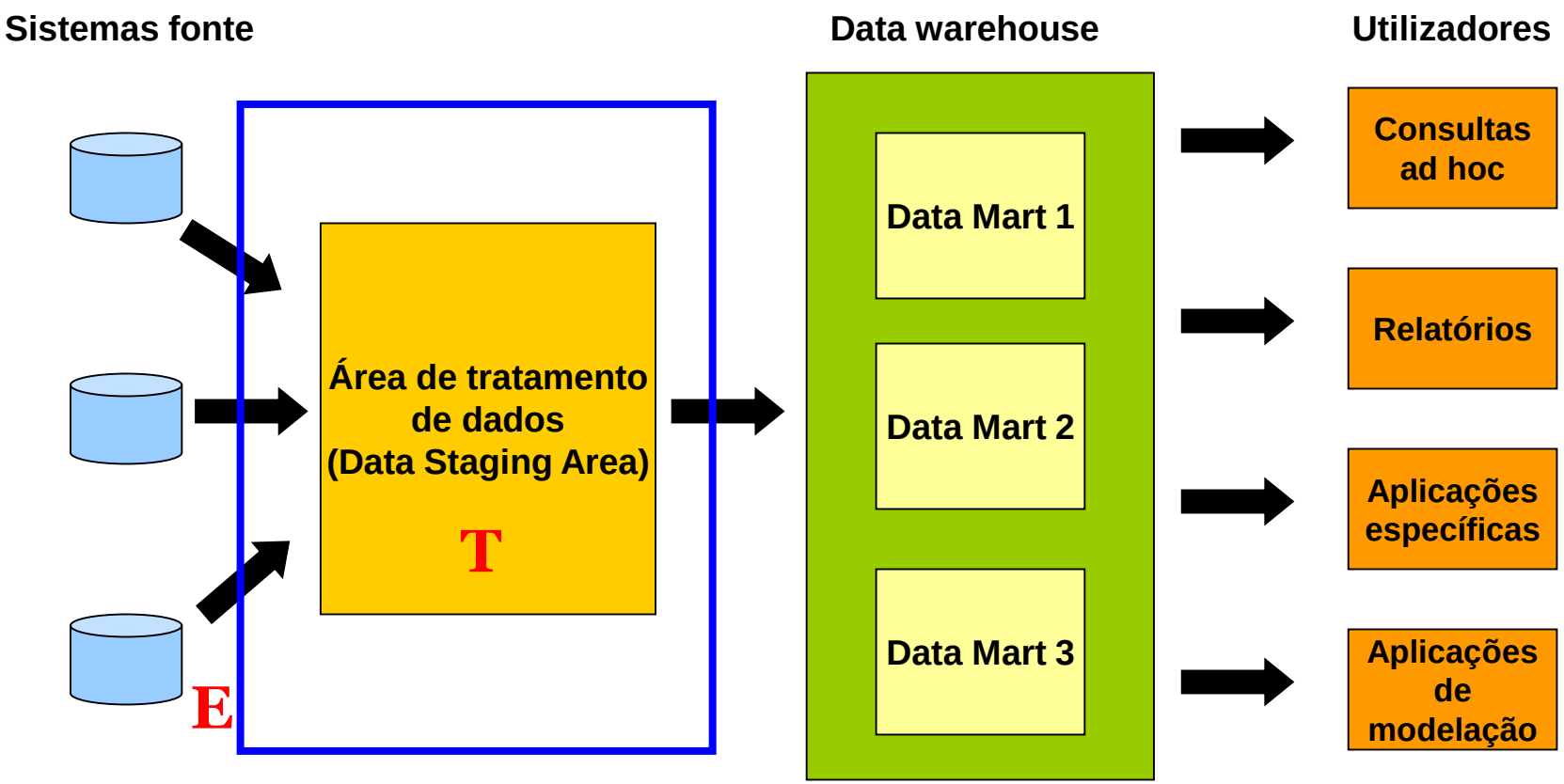
Extração de Dados

- Técnicas CDC: Processo de eliminação
  - Extração inicial e incrementais
    - Criam-se duas tabelas na DSA
      - *table_new* e *table_old*
    - Os dados extraídos vão para a *table_new*
    - Selecionar *table_new* *MINUS* *table_old*
    - Transformar e carregar o resultado da seleção para a BD do *Data Warehouse*
    - Por último na DSA fazer:
      - *drop table_old*
      - *rename table_new to table_old*
      - *create empty table_new*

Processo ETL

- Sumário
  - Introdução
  - Extração de Dados
  - Transformação de Dados
  - Carregamento de Dados
  - Mapa Lógico de Dados
  - Passos Típicos do Processo ETL

Arquitetura do *Data Warehouse*



Transformação de Dados

- Introdução
  - Ao contrário do processo de extração, onde geralmente os dados apenas são movidos e reformatados, no processo de transformação os dados são modificados
  - Após a extração de dados é crucial garantir a limpeza e conformidade dos mesmos

Transformação de Dados

- Introdução
  - O processo de transformação envolve duas atividades principais
    - Verificação da qualidade dos dados
    - Transformações de dados
  - Existem várias formas de transformar os dados provenientes dos sistemas fonte:
    - Limpeza dos dados
    - Eliminação de campos inúteis
    - Combinação de dados provenientes de fontes diferentes

Transformação de Dados

- Introdução
  - A limpeza e conformidade geram metadados que permitem um diagnóstico sobre o que está errado nos sistemas fonte
    - Estes metadados acompanham os dados até estes chegarem aos utilizadores finais do *Data Warehouse*
  - O objetivo final é garantir a qualidade dos dados

Transformação de Dados

- Qualidade dos dados
  - Correção
    - Os valores dos dados são genuínos
  - Clareza
    - Os dados só podem ter um significado
  - Consistência
    - Utilizar apenas uma convenção para a representação dos dados
  - Completude
    - Os valores dos campos existem

Transformação de Dados

- Qualidade dos dados
  - Verificação da qualidade dos dados
    - Deteção de erros ("*Screens*")
    - Registo de erros
    - Análise de erros
  - A correção de problemas nos dados deve ser feita nos sistemas fonte
    - As soluções adotadas na DSA são sempre temporárias
  - Dados sem qualidade comprometem o funcionamento do Data Warehouse

Transformação de Dados

- Transformações de dados
  - Limpeza de dados
  - Integração de dados
  - Outras transformações
    - Modificar códigos
    - Valores calculados
    - Agregações prévias
    - Introdução de referências temporais para casos excecionais
    - ...

Transformação de Dados

- Limpeza de dados: Dados "sujos"
  - Valores sem sentido
    - Correção de erros ortográficos
  - Ausência de dados
    - Tratamento de campos vazios
  - Dados duplicados
    - Eliminação de duplicações
  - Dados cujo significado não é claro (e que os metadados não esclarecem)

Transformação de Dados

- Limpeza de dados: Dados "sujos"
  - Dados contraditórios
    - Resolução de conflitos (Exemplo: cidade incompatível com código postal)
  - Dados que violam regras de integridade
    - Referencial
    - Temporal
    - Domínio
  - Colocar os dados em formatos standard

Transformação de Dados

- Limpeza de dados: Eliminação de inconsistências
  - Devidas à recolha dos mesmos dados em mais do que um sistema ou plataforma
  - Devido a insuficiências no processo de extração
  - Causadas por alterações nos sistemas operacionais
  - Devidas a problemas técnicos nos sistemas operacionais
    - Situações de falha

Transformação de Dados

- Limpeza de dados: Exemplo

CUST #	NAME	ADDRESS	TYPE
90328574	Digital Equipment	187 N. PARK St. Salem NH 01458	OEM
90328575	DEC	187 N. Pk. St. Salem NH 01458	OEM
90238475	Digital	187 N. Park St Salem NH 01458	$#%
90233479	Digital Corp	187 N. Park Ave. Salem NH 01458	Comp
90233489	Digital Consulting	15 Main Street Andover MA 02341	Consult
90234889	Digital Info Service	PO Box 9 Boston MA 02210	Mail List
90345672	Digital Integration	Park Blvd. Boston MA 04106	SYS INT

No Unique Key

Anomalies

No Standardization

Spelling

Noise in Blank Fields

Transformação de Dados

- Processo de limpeza de dados
  - Processos automáticos permitem resolver normalmente com eficácia:
    - Problemas relacionados com formatos dos dados, conversões, etc.
    - Falta de estandardização
    - Preenchimento de valores em falta, etc.
  - Processos manuais
    - Necessários quando a correção é semântica
    - Apoiados por ferramentas

Transformação de Dados

- Integração de dados: Conformidade
  - Dados que deviam estar relacionados mas que não podem ser relacionados corretamente
    - Devido à ausência de chaves primárias nos dados ou a chaves não unívocas
  - Dados que estão relacionados, mas que na verdade não devem ter qualquer relacionamento entre eles
    - Quando se utilizam atributos ou registos para vários fins

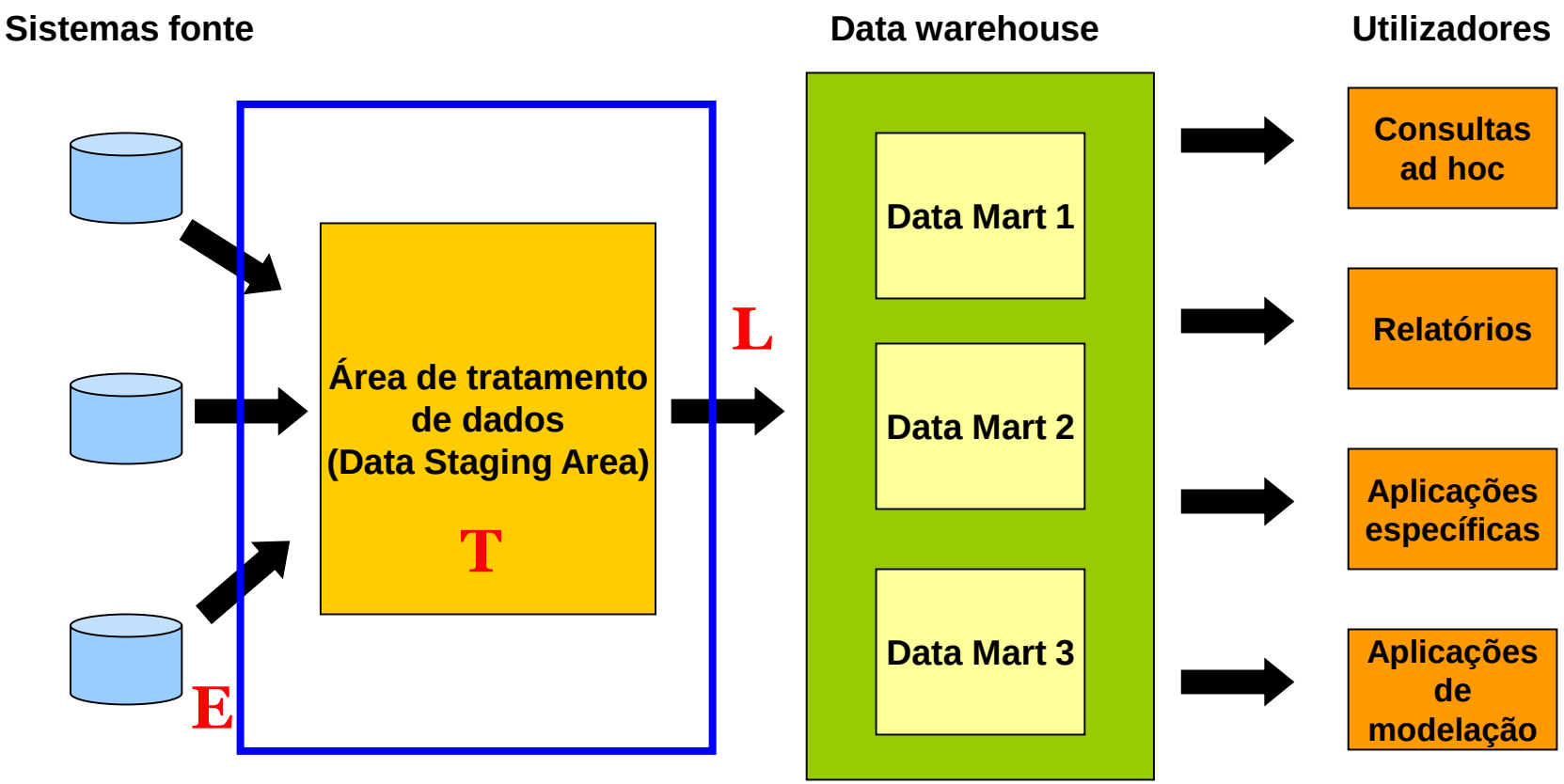
Transformação de Dados

- Tipos de transformações
  - Ao nível do registo
    - Seleção: particionamento dos dados
    - Junção: combinação dos dados
    - Agregação: resumo dos dados
  - Ao nível dos campos
    - Envolvendo um único campo: de um campo para outro campo
    - Envolvendo múltiplos campos: de muitos campos para um ou de um campo para muitos

Processo ETL

- Sumário
  - Introdução
  - Extração de Dados
  - Transformação de Dados
  - Carregamento de Dados
  - Mapa Lógico de Dados
  - Passos Típicos do Processo ETL

Arquitetura do *Data Warehouse*



Carregamento de Dados

- 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

Carregamento de Dados

- 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

Carregamento de Dados

- 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, …

Carregamento de Dados

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

Carregamento de Dados

- 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

Processo ETL

- Sumário
  - Introdução
  - Extração de Dados
  - Transformação de Dados
  - Carregamento de Dados
  - Mapa Lógico de Dados
  - Passos Típicos do Processo ETL

Mapa Lógico de Dados

- Definição do mapa lógico de dados
  - Essencial para o sucesso do Processo ETL
  - Descreve os relacionamentos entre as fontes de dados e os campos destino no *Data Warehouse*
  - Este documento permite estabelecer uma ligação entre o ponto inicial e o ponto final do Processo ETL

Mapa Lógico de Dados

- Definição do mapa lógico de dados
  - Antes de se implementar o Processo ETL é necessário
    - Ter um plano (Mapa Lógico de Dados)
    - Identificar as fontes de dados candidatas
    - Analisar os sistemas fonte (qualidade dos dados, etc.)
    - Percorrer a linhagem dos dados e regras de negócio
    - Percorrer o modelo físico de dados do DW
    - Validar cálculos e fórmulas

Mapa Lógico de Dados

- Estrutura do Mapa
  - É geralmente apresentado na forma de uma tabela ou folha de cálculo e inclui três componentes principais:
    - Destino
    - Origem
    - Transformação
  - Para cada um dos componentes principais são definidas várias colunas

Mapa Lógico de Dados

- Estrutura do Mapa: Destino
  - Nome da tabela destino
  - Nome da coluna destino
  - Tipo de dados da coluna destino
  - Tamanho
  - Tipo de tabela
    - Tabela de Dimensão
      - Tipo de alteração (SCD: Tipo 1, 2 ou 3)
    - Tabela de Factos
      - Tipo de facto (Aditivo, Semi-aditivo ou Não aditivo)

Mapa Lógico de Dados

- Estrutura do Mapa: Origem
  - Base de Dados
    - Base de dados origem
    - Nome da tabela origem
    - Nome da coluna origem
    - Tipo de dados da coluna origem
  - Ficheiro
    - Ficheiro origem
    - Nome da folha/elemento origem
    - Nome da coluna origem
    - Tipo de dados da coluna origem

Mapa Lógico de Dados

- Estrutura do Mapa: Transformação
  - Descrição exata da forma como é feita a manipulação dos dados fonte de forma a corresponder ao formato destino que é esperado
    - Código SQL
    - Pseudocódigo

Processo ETL

- Sumário
  - Introdução
  - Extração de Dados
  - Transformação de Dados
  - Carregamento de Dados
  - Mapa Lógico de Dados
  - 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

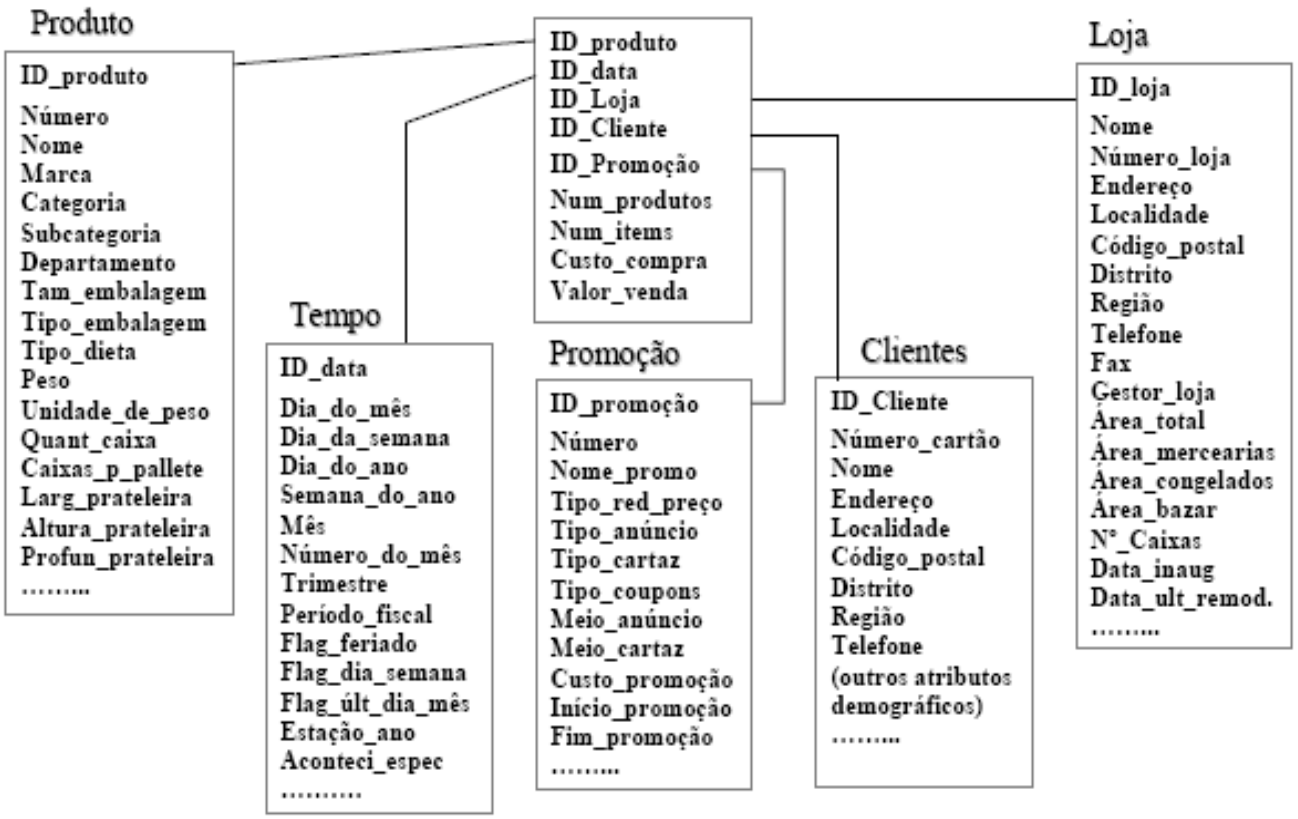
- Carregamento de factos
  - Elaborar, testar e executar planos ETL para tabelas de factos
  - Elaborar e testar processo de carregamentos periódicos

Processo ETL

- Automatizar o processo ao máximo
  - Utilização de ferramentas sofisticadas de suporte
    - Escalonamento dos processos
    - Execução automática

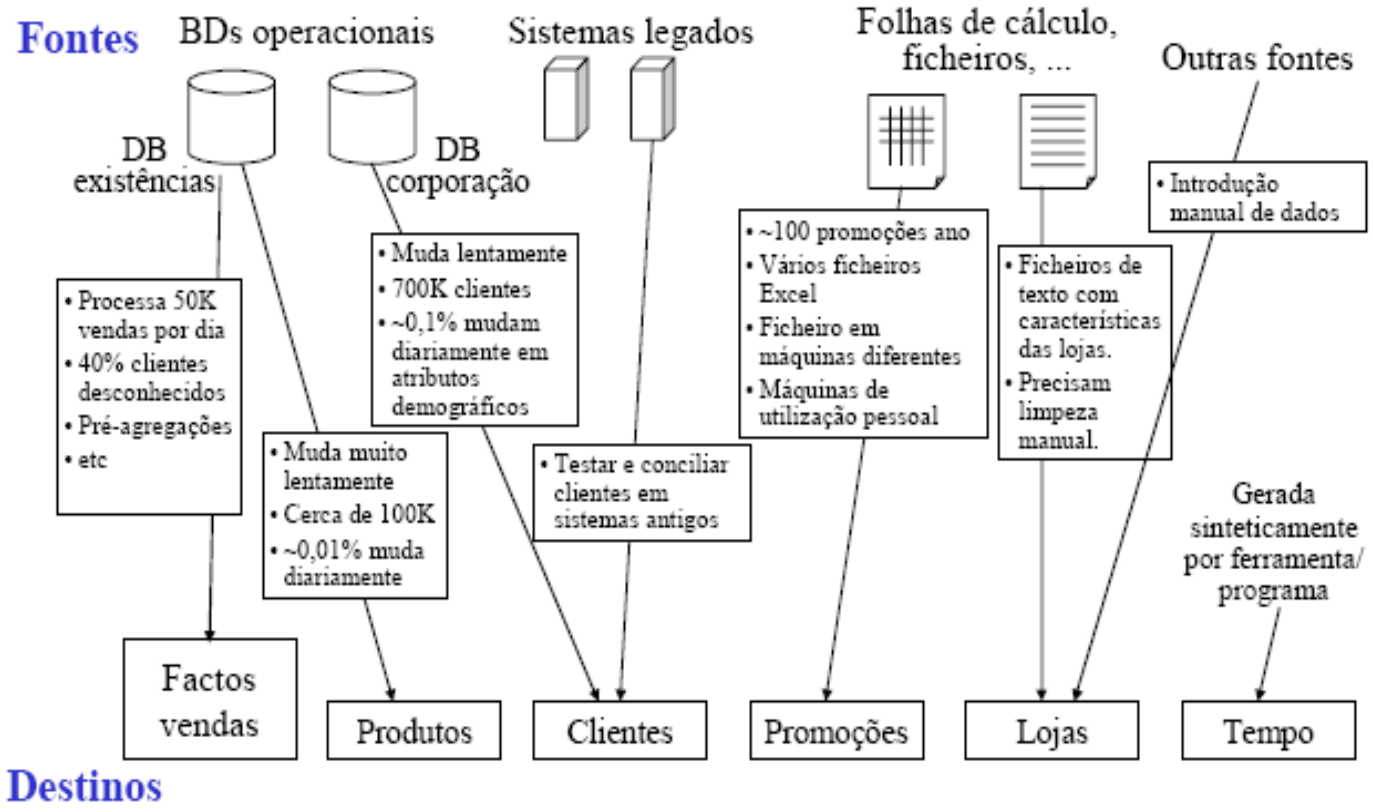
Processo ETL

- 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

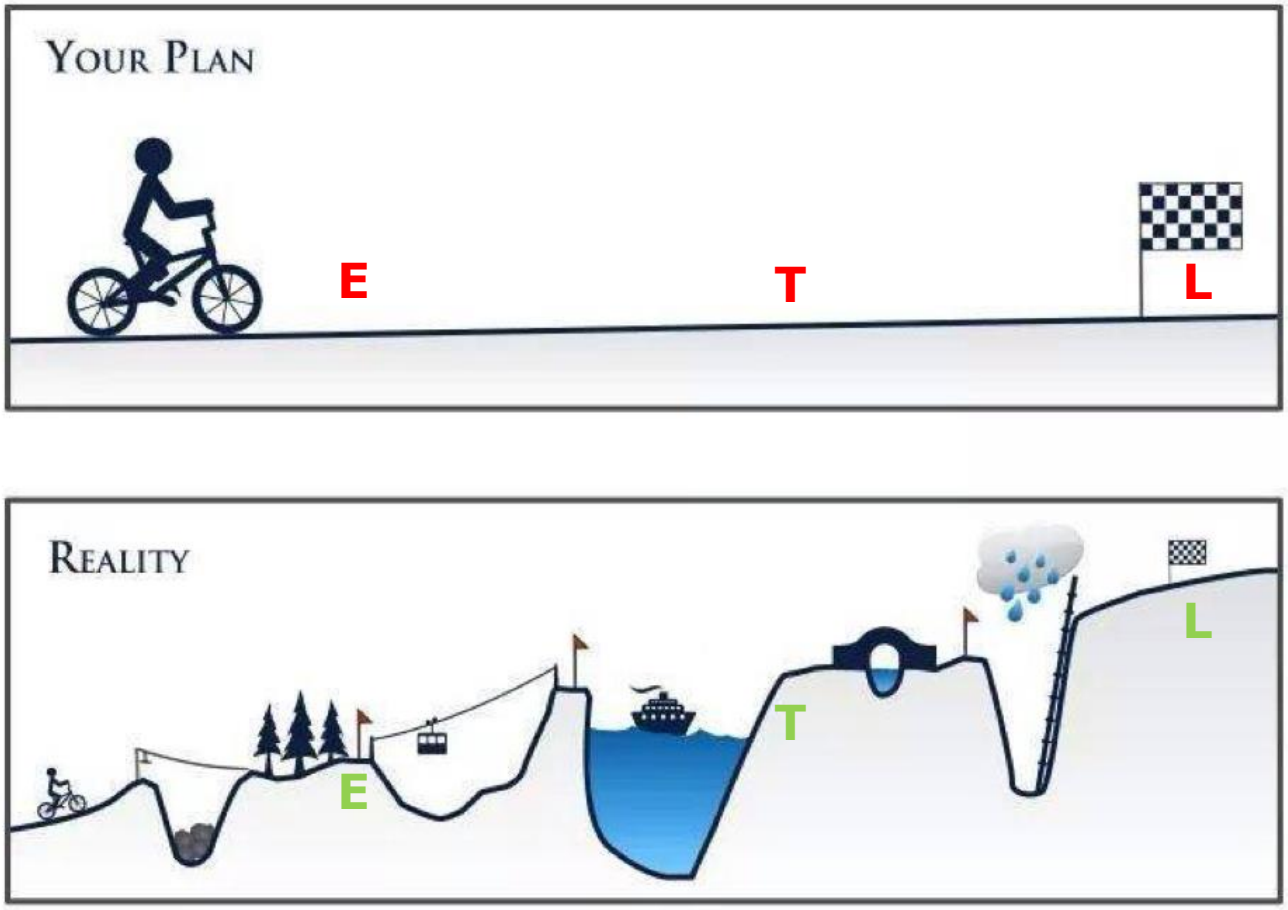
Processo ETL

- 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



Processo ETL

- Referências
  - The Data Warehouse ETL Toolkit, R. Kimball e J. Caserta, John Wiley & Sons, 2004
    - Capítulos 1, 2, 3, 4, 5 e 6
  - Sistemas de Suporte à Decisão, B. Cortes, FCA, 2005
    - Capítulo 3