# Transformação de Dados


Sistemas de Apoio à Decisão

# Transformação de Dados

- Sumário
  - Introdução
  - Processo de Transformação
  - Qualidade dos Dados
  - Limpeza de Dados
  - Transformações de Dados
  - Mapa Lógico de Dados

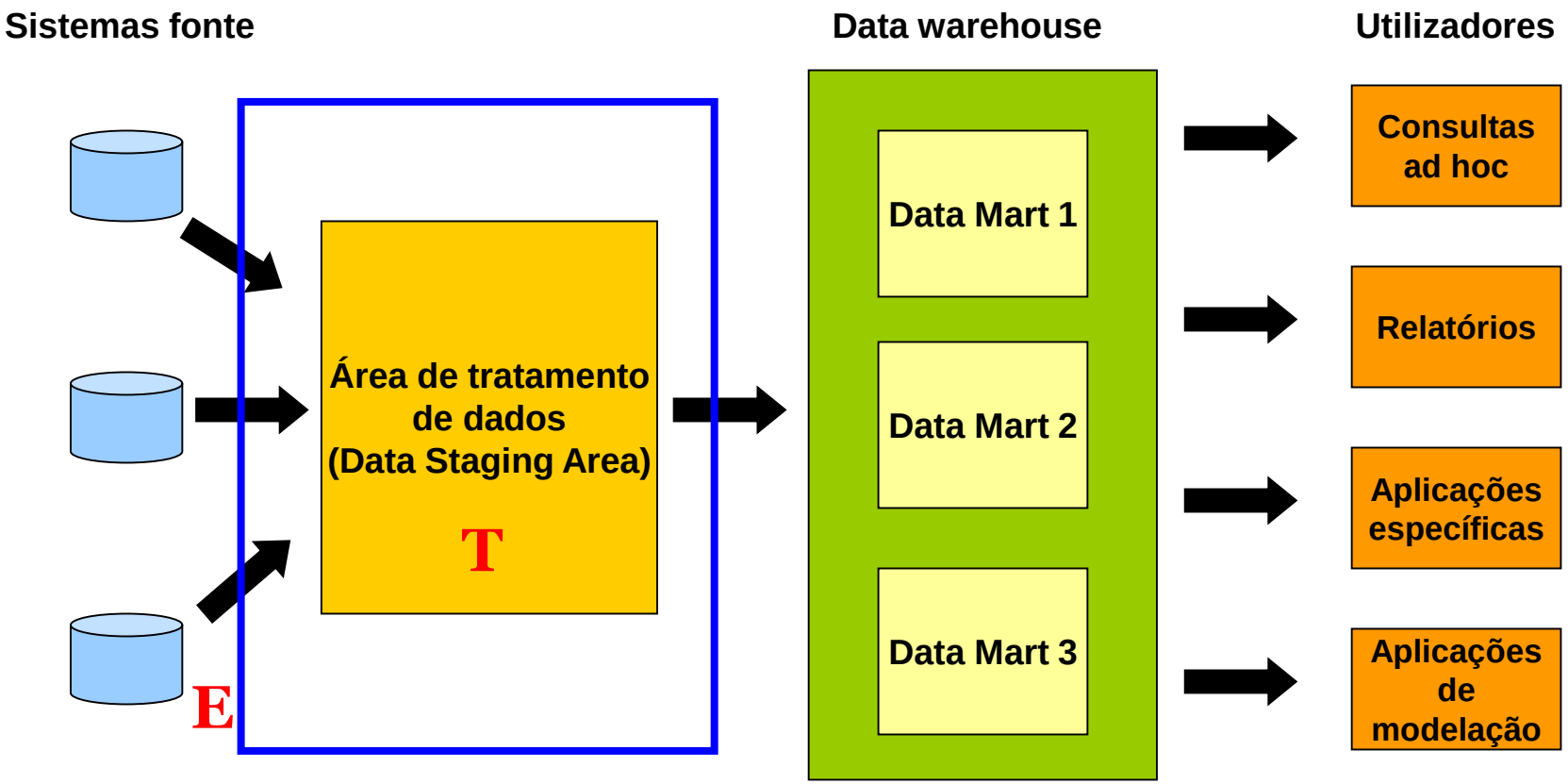
# Introdução

- 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*

# Transformação de Dados

- Sumário
  - Introdução
  - Processo de Transformação
  - Qualidade dos Dados
  - Limpeza de Dados
  - Transformações de Dados
  - Mapa Lógico de Dados

# Processo de Transformação



# Processo de Transformação

- 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

# Processo de Transformação

- 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
    - Integração de dados provenientes de fontes diferentes

# Processo de Transformação

- 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

- Sumário
  - Introdução
  - Processo de Transformação
  - Qualidade dos Dados
  - Limpeza de Dados
  - Transformações de Dados
  - Mapa Lógico de Dados

# Qualidade dos Dados

- Introdução
  - Pontos de garantia de qualidade
    - E->TL
    - ET->L
  - 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*

# Qualidade dos Dados

- Introdução
  - Critérios de qualidade dos dados
    - Correção
    - Clareza
    - Consistência
    - Completude
  - Tarefas relacionadas com a qualidade dos dados
    - Deteção de erros
    - Registo de erros
    - Análise de erros

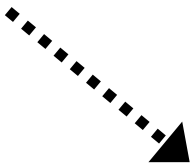
# Qualidade dos Dados

- Correção
  - Os valores dos dados respeitam uma versão oficial
    - Atenção: correto ≠ verdadeiro

Clientes

| Id | dataEntrada | Morada |
|---|---|---|
| 2000 | 12-Jun-2024 | Rua Filipe Andrade |
| 2001 | 12-Jun-2024 | Avenida Maria Faria Filhos |
| 2003 | 15-Jun-2024 | Estrada Conto do Vigário |
| 2006 | 15-Jun-2024 | Alto do Vieiro |

Fontes
(dados *não CORRETOS*)



**DW**
**(dados CORRETOS)**

Clientes

| Id | dataEntrada | Morada |
|---|---|---|
| 2000 | 12-Jun-2024 | Rua Filipe **de** Andrade |
| 2001 | 12-Jun-2024 | Avenida Maria Faria **e** Filhos |
| 2003 | 15-Jun-2024 | Estrada C**a**nto do Vigário |
| 2006 | 15-Jun-2024 | Alto do Vieiro |

# Qualidade dos Dados

- Clareza
  - Os dados só podem ter um significado
  - Dados são claros se não há dúvidas sobre o seu significado

Dados *não claros*

Locais

| Id | Nome |
|---|---|
| 1 | New York |
| 2 | New York |
| 3 | Lisboa |
| 4 | Lisboa |

Dados **CLAROS**

Locais

| Id | Nome | Tipo |
|---|---|---|
| 1 | New York | Estado |
| 2 | New York | Cidade |
| 3 | Lisboa | Cidade |
| 3 | Lisboa | Distrito |

# Qualidade dos Dados

- Consistência
  - Utilizar apenas uma convenção para a representação dos dados

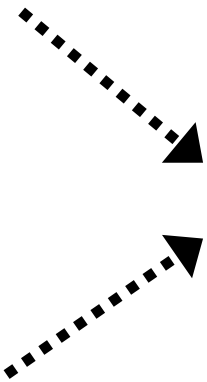
Fontes
(dados *não CONSISTENTES*)

Clientes

| Id | NIF | País |
|---|---|---|
| 192 | 210116324 | USA |
| 119 | 211075757 | PT |
| 186 | 211009929 | PT |
| 100 | 211009929 | ES |

Clientes

| Id | NIF | País |
|---|---|---|
| 192 | 210116324 | EUA |
| 119 | 211075757 | POR |
| 208 | 207446644 | POR |
| 15 | 208747475 | SPA |



**DW**
**(dados CONSISTENTES)**

Clientes

| Id | NIF | País |
|---|---|---|
| 192 | 210116324 | USA |
| 119 | 211075757 | PT |
| 186 | 211009929 | PT |
| 100 | 211009929 | ES |
| 208 | 207446644 | PT |
| 15 | 208747475 | ES |

# Qualidade dos Dados

- Completude
  - Os valores dos campos existem quando é esperado que existam

**DW**
**(dados *não completos*)**

**Fontes**

...

T_DATA_ENTREGAS

| Id | data | codPostal |
|---|---|---|
| 771616252 | 12-Jun-2025 | 3080-**000** |
| 771616253 | 12-Jun-2025 | 3090-881 |
| 771616254 | 15-Jun-2025 | **2400** |
| 771616255 | 15-Jun-2025 | |
| 771616257 | 16-Jun-2025 | 3090-354 |

# Qualidade dos Dados

- Completude
  - O número total de linhas é o esperado

**DW**
**(dados *não completos*? Por semana há habitualmente 7200 entregas)**

**Fontes**

...

T_DATA_ENTREGAS

| Id | data | codPostal |
|---|---|---|
| 771616252 | 12-Jun-2025 | 3080-122 |
| 771616253 | 12-Jun-2025 | 3090-881 |
| 771616254 | 15-Jun-2025 | 2400-100 |
| 771616255 | 15-Jun-2025 | 2400-102 |

# Qualidade dos Dados

- Gestão de Erros
  - Atitudes perante o erro
    - Ignorar?
    - Abortar o processo?
    - Corrigir os erros?
  - Escolher atitude após uma decisão informada
  - O controlo de qualidade implica uma série de *checkpoints* e relatórios (*screens*)
  - Registar todos os erros detetados

# Qualidade dos Dados

- Deteção de Erros
  - Testes de qualidade
  - Deve existir uma forte componente de dados e metadados acerca das fontes
  - *Screens*
    - Têm sempre associado um ou mais critérios de qualidade dos dados (4Cs)
    - Exemplos de testes
      - As colunas estão preenchidas?
      - Os valores extraídos são os esperados?
      - O total de linhas extraído é o esperado?

# Qualidade dos Dados

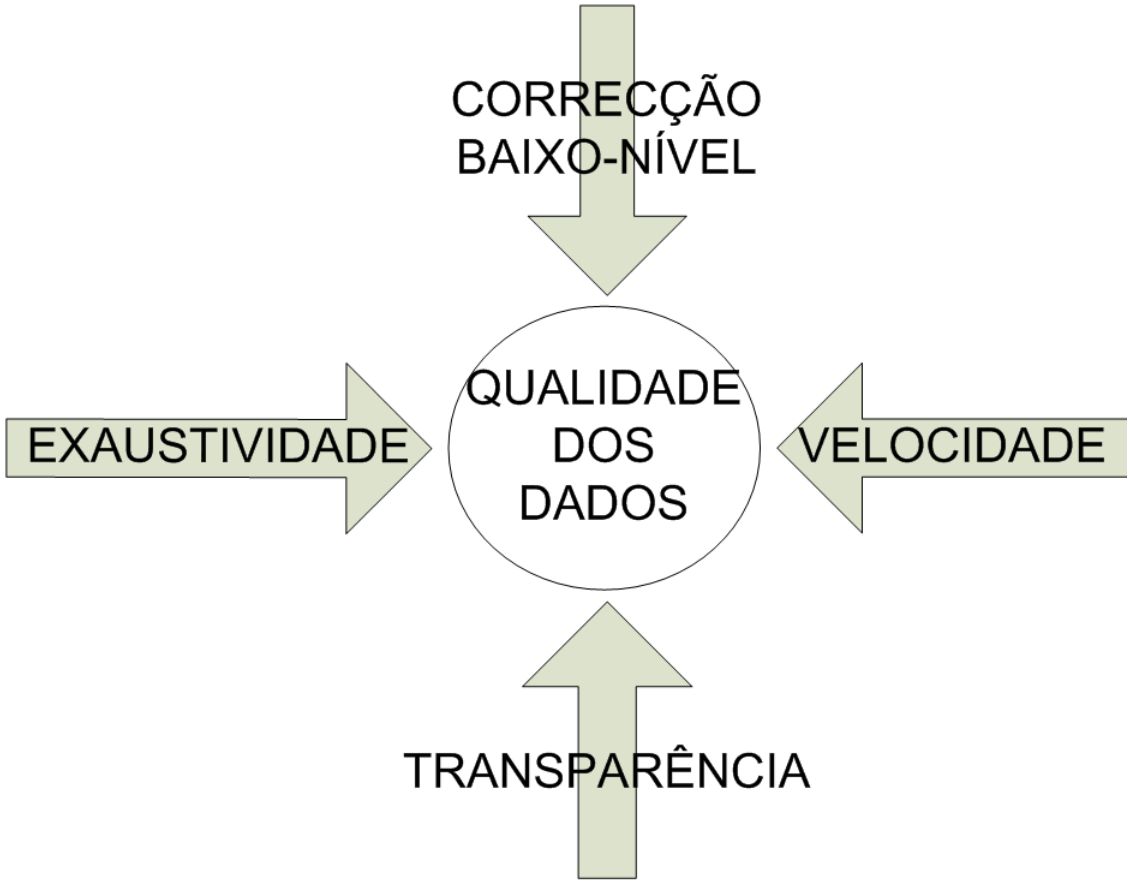
- Registo de Erros
  - Infraestrutura para registo de erros com um nível de detalhe adequado
    - Registo individual de cada erro
    - Registo de grupos de erros
  - Ficheiro de *Log*
    - Um ficheiro para cada iteração do processo ETL
  - *Transformation Error Logger* (TEL)
    - Histórico dos erros detetados
    - Sistema de Apoio à Decisão

# Qualidade dos Dados

- Análise de Erros
  - Análise estatística temporal e histórica
  - Resposta a questões do tipo
    - A qualidade dos dados tem melhorado?
    - Quais as fontes com mais problemas de qualidade?
    - Quais os *screens* que consomem mais tempo?
    - Há *screens* que já não sejam necessários por não serem registados erros?
    - …

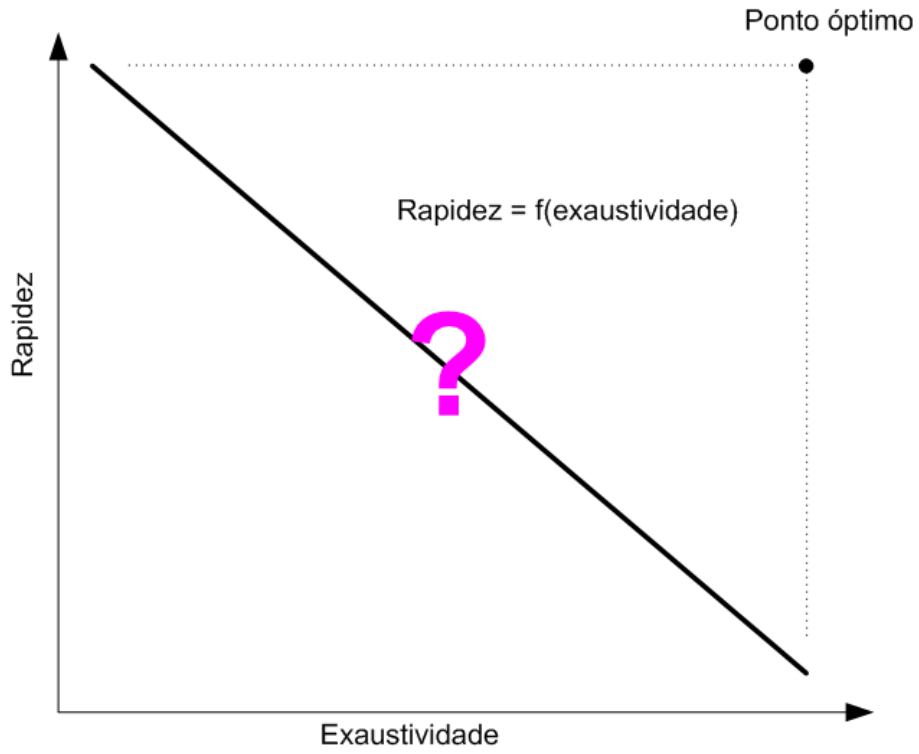
# Qualidade dos Dados

- Conflitos e Prioridades



# Qualidade dos Dados

- Conflitos e Prioridades
  - Exaustividade vs. Rapidez



# Qualidade dos Dados

- Conflitos e Prioridades
  - Correção vs. Transparência
    - Reforçar demasiado a transparência pode provocar demasiada reengenharia de processos e com isso parar o processo ETL
    - Demasiada correção faz com que os problemas não sejam realmente resolvidos, o que pode provocar mais problemas no futuro

# Transformação de Dados

- Sumário
  - Introdução
  - Processo de Transformação
  - Qualidade dos Dados
  - Limpeza de Dados
  - Transformações de Dados
  - Mapa Lógico 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)

# 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

# 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

# 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 |

**Sem chave única** **Anomalias** **Sem uniformização** **Ortografia** **Lixo**

# Limpeza 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 específicas

# Transformação de Dados

- Sumário
  - Introdução
  - Processo de Transformação
  - Qualidade dos Dados
  - Limpeza de Dados
  - Transformações de Dados
  - Mapa Lógico de Dados

# Transformações de Dados

- Integração de dados
  - Combinação de dados de múltiplos sistemas fonte
  - Obrigatória pois o modelo de dados do *Data Warehouse* é mais simples que o dos sistemas operacionais
  - Na junção de dados podem surgir dificuldades inesperadas
    - Ausência de chaves
    - Dados em falta, etc.

# Transformações 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

# Transformações 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

# Transformações de Dados

- Correção de erros
  - Deve ser feita a montante, nos sistemas fonte
  - Corrigir erros que devem ser corrigidos nas fontes é má política
    - Solução temporária e não definitiva
  - Necessidade de tabelas auxiliares
    - Representação correta para os valores errados
    - Tabelas T_LOOKUP

# Transformações de Dados

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

# Transformações de Dados

- Outras transformações
  - Exemplo 1
    - A modificação dos valores vai facilitar as análises de dados aos utilizadores finais

**Sistema Fonte**

T_CLIENTES

| Id | … | Sexo | Estado Civil |
|---|---|---|---|
| 17523 | | M | S |
| 17678 | | F | C |
| 18752 | | F | S |
| 19790 | | M | D |

**Data Warehouse**

T_DIM_CUSTOMER

| Id | … | Sexo | Estado Civil |
|---|---|---|---|
| 1 | | Masculino | Solteiro |
| 2 | | Feminino | Casado |
| 3 | | Feminino | Solteiro |
| 4 | | Masculino | Divorciado |

# Transformações de Dados

- Outras transformações
  - Exemplo 2
    - A criação de uma nova coluna vai facilitar as análises de dados aos utilizadores finais

**Sistema Fonte**

T_UNIDADES_CURRICULARES

| Código | … | Curso | Nome |
|---|---|---|---|
| 9119229 | | 9119 | Sistemas de Apoio à Decisão |
| 9885229 | | 9885 | Sistemas de Apoio à Decisão |

**Data Warehouse**

T_DIM_UNIDADE_CURRICULAR

| UC_Key | UC_Natural_Key | … | UC_Nome | UC_Código_Nome |
|---|---|---|---|---|
| 11554 | 9119229 | … | Sistemas de Apoio à Decisão | 9119229::Sistemas de Apoio à Decisão |
| 11832 | 9885229 | … | Sistemas de Apoio à Decisão | 9885229::Sistemas de Apoio à Decisão |

# Transformação de Dados

- Sumário
  - Introdução
  - Processo de Transformação
  - Qualidade dos Dados
  - Limpeza de Dados
  - Integração de Dados
  - Transformações de Dados
  - Mapa Lógico de Dados

# Mapa Lógico de Dados

- Definição do Mapa
  - 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
    - Fluxos de dados (*pipelines*)

# Mapa Lógico de Dados

- Definição do Mapa
  - 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:
    - Origem
    - Transformação
    - Destino
  - Para cada um dos componentes principais são definidas várias colunas

# Mapa Lógico de Dados

- Estrutura do Mapa: Origem
  - Base de Dados
    - Nome da base de dados origem
    - Nome da tabela origem
    - Nome da coluna origem
    - Tipo de dados da coluna origem
  - Ficheiro
    - Nome do 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
    - …

# 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, semiaditivo ou não aditivo)

# Mapa Lógico de Dados

- Exemplo
  - Plataforma ETL^2

| EXTRACT | | | | | TRANSFORM | | | | LOAD | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| source | | | target | | target | | | operation | target | | | | |
| - | - | - | - | - | - | - | - | - | T_DIM_CUSTOMER | customer_key | NUMBER(12) | PK | SCD 1 |
| VIEW_CLIENTES | src_id | integer | t_data_customers | id | T_CLEAN_CUSTOMERS | id | NUMBER | - | | customer_natural_Key | NUMBER(10) | - | SCD 1 |
| | src_card_number | varchar2(20) | t_data_customers | card_number | | card_number | VARCHAR2 | - | | customer_card_number | VARCHAR2(20) | - | SCD 1 |
| | src_name | varchar2(40) | t_data_customers | name | | name | VARCHAR2 | to UPPERCASE | | customer_name | VARCHAR2(40) | - | SCD 1 |
| | src_address | varchar2(60) | t_data_customers | address | | address | VARCHAR2 | to UPPERCASE | | customer_address | VARCHAR2(60) | - | SCD 1 |
| | src_location | varchar2(60) | t_data_customers | location | | location | VARCHAR2 | to UPPERCASE | | customer_location | VARCHAR2(60) | - | SCD 2 |
| | src_district | varchar2(40) | t_data_customers | district | | district | VARCHAR2 | to UPPERCASE | | customer_district | VARCHAR2(40) | - | SCD 2 |
| | src_zip_code | varchar2(8) | t_data_customers | zip_code | | zip_code | VARCHAR2 | | | customer_zip_code | VARCHAR2(8) | - | SCD 2 |
| | src_phone_nr | number(9) | t_data_customers | phone_nr | | phone_nr | NUMBER | | | customer_phone_nr | NUMBER(9) | - | SCD 1 |
| | src_gender | char(1) | t_data_customers | gender | | gender | VARCHAR2 | IF (gender='M') THEN 'MALE'<br>ELSE IF (gender='F') THEN 'FEMALE'<br>ELSE 'OTHER'; | | customer_gender | VARCHAR2(15) | - | SCD3 |
| | src_age | number(3) | t_data_customers | age | | age | NUMBER | | | customer_age | NUMBER(3) | - | SCD 2 |
| | src_marital_status | char(1) | t_data_customers | marital_status | | marital_status | VARCHAR2 | IF (marital_status='C') THEN 'MARRIED''<br>ELSE IF (marital_status='S') THEN 'SINGLE'<br>ELSE IF (marital_status='V') THEN 'WIDOW'<br>ELSE IF (marital_status='D') THEN 'DIVORCED'<br>ELSE 'OTHER'; | | customer_marital_staus | VARCHAR2(15) | - | SCD2 |
| | src_card_number | varchar2(20) | t_data_customers | card_number | | - | VARCHAR2 | (JOIN OPERATION)<br>t_data_customers JOIN t_data_customers_reg | | customer_type | VARCHAR2(20) | - | SCD2 |
| VIEW_REGISTOS | src_card_number | varchar2(20) | t_data_customers_reg | card_number | | - | | | | | | | |
| | src_customer_type | varchar2(10) | t_data_customers_reg | customer_type | | customer_type | | | | | | | |

# Transformação de Dados

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