TÉCNICO LISBOA

Bases de Dados

T23 - OLAP Parte I

Prof. Daniel Faria

Sumário

- Sistemas de Apoio à Decisão / Business Intelligence
- Data Warehouses
- Limitações do Modelo Relacional Normalizado (BD Operacionais)
- Esquema em Estrela
- Data Warehouses e SGBD

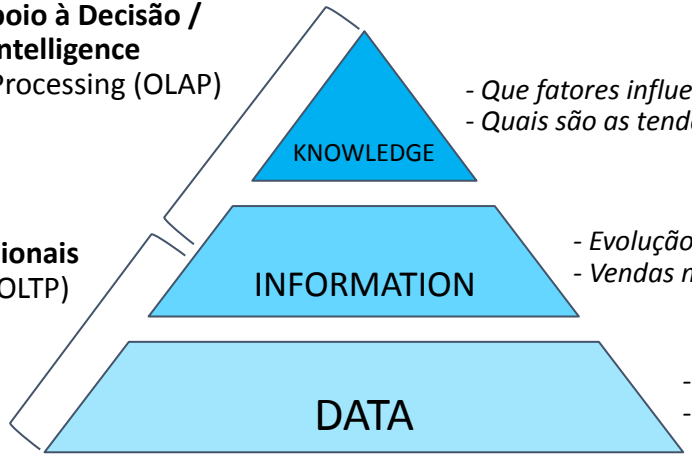
TÉCNICO LISBOA

Sistemas de Apoio à Decisão / Business Intelligence

Sistemas de Informação

Sistemas de Apoio à Decisão / Business Intelligence
Online Analytic Processing (OLAP)

Sistemas de Informação Operacionais
Online Transaction Processing (OLTP)



- Que fatores influenciam as vendas?
- Quais são as tendências das vendas?

- Evolução das vendas por cidade e mês
- Vendas mensais por produto em Lisboa

- Listar o preço de cada produto
- Registar vendas

OLTP vs. OLAP

- **Online Transaction Processing (OLTP)**
  - Dados dinâmicos
    - Operações de escrita frequentes
  - Transações de escrita ou leitura rápidas e simples
- **Online Analytic Processing (OLAP)**
  - Dados quase estáticos
    - Atualizações periódicas (e.g. mensais, anuais) em bulk
  - Transações complexas mas só de leitura

Sistema de Apoio à Decisão

- Sistema de informação que apoia os processos de tomada de decisão em organizações ou empresas
- Possibilita a análise de dados atuais e históricos com o objetivo de encontrar padrões e suportar a delineação de uma estratégia
- Análise complexa, interactiva, exploratória de grandes conjuntos de dados obtidos por integração das várias fontes internas e externas
- Tipicamente implica integração e agregação de dados (data warehousing)

Business Intelligence

- Sistema de apoio à decisão empresarial, que apoia decisões desde o nível operacional (e.g. preço dos produtos) até ao nível estratégico (e.g. desenvolvimento de novos produtos)
- Tipicamente engloba:
  - Engenharia e integração de dados: **data warehousing** (BI/DW)
  - Análise e exploração de dados: **data analytics**
  - Mineração de dados: **data mining**
  - Análise de processos
  - Avaliação de desempenho

Business Intelligence

1. Reunir dados de múltiplas fontes num **Data Warehouse**
   - Dados frequentemente requerem extração, transformação para um esquema comum, e carregamento
2. Gerar agregações e relatórios que sumarizam os dados
   - Dashboards com gráficos e relatórios
   - Sistemas **OLAP** para exploração interativa dos dados
   - Análise estatística
3. Construir modelos preditivos e utilizá-los para apoiar os processos de tomada de decisão

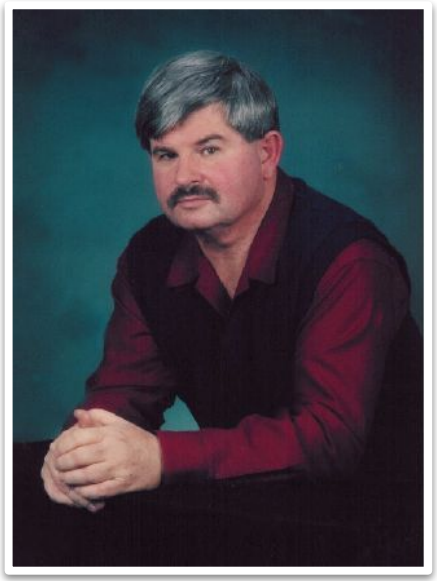
TÉCNICO LISBOA

Data Warehousing

Data Warehouse

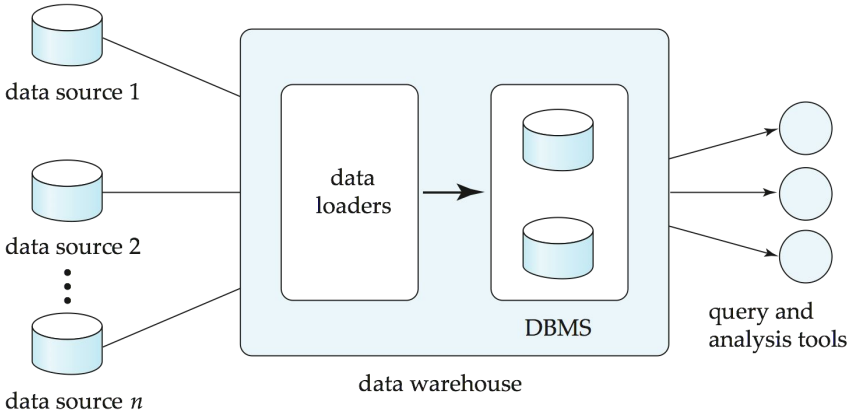
"A Data Warehouse [...] enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information which has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user"

—Bill Inmon



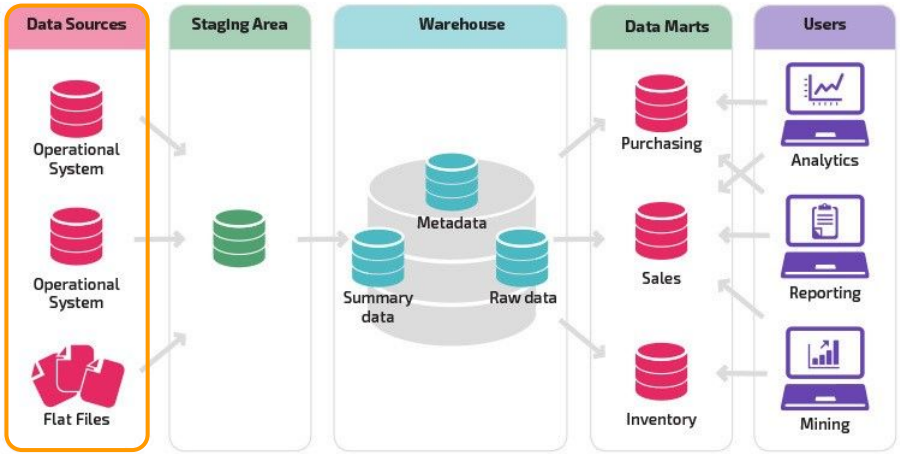
Data Warehouse

- Repositório central de grande volume de dados consolidados, históricos e agregados, complementados com sumários
- Permite simplificar queries complexas para análise de dados
- Base para reporte e análise de dados e componente chave de business intelligence



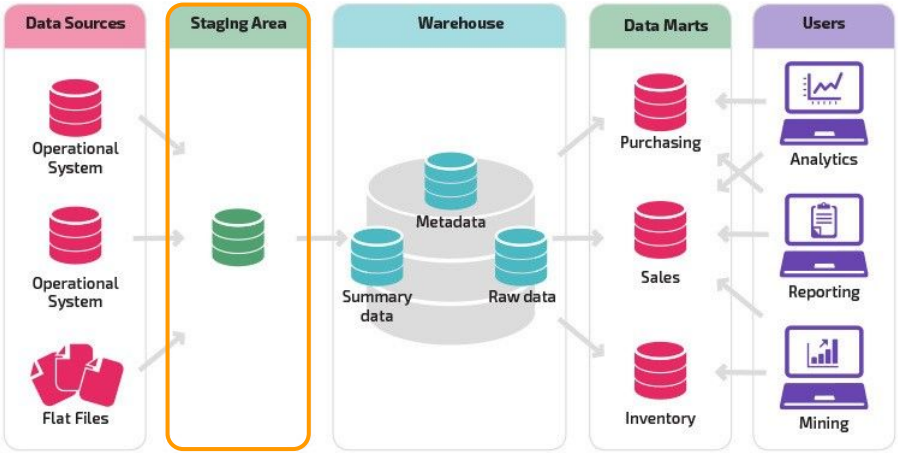
Data Warehousing

- O processo de construção de um data warehouse começa por identificar as **fontes de dados** a integrar
  - Tipicamente dados de vários sistemas operacionais da empresa (e.g., vendas, marketing)
  - Possivelmente dados externos (bases de dados ou em ficheiros)



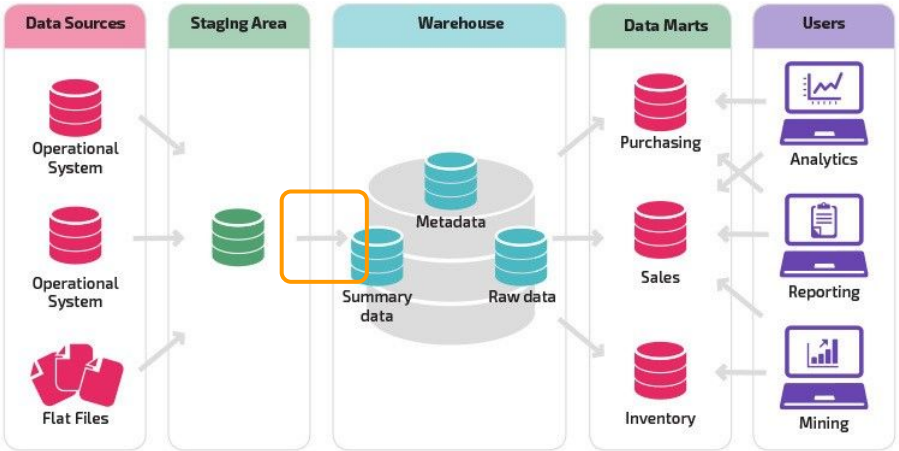
Data Warehousing

- A **área de staging** visa isolar o processo de carregamento do data warehouse das fontes de dados brutos
- Garante a disponibilidade dos dados
- Permite avaliar a qualidade dos dados
- Permite capturar alterações aos dados (i.e. determinar que dados são novos e precisam de ser carregados)



Data Warehousing

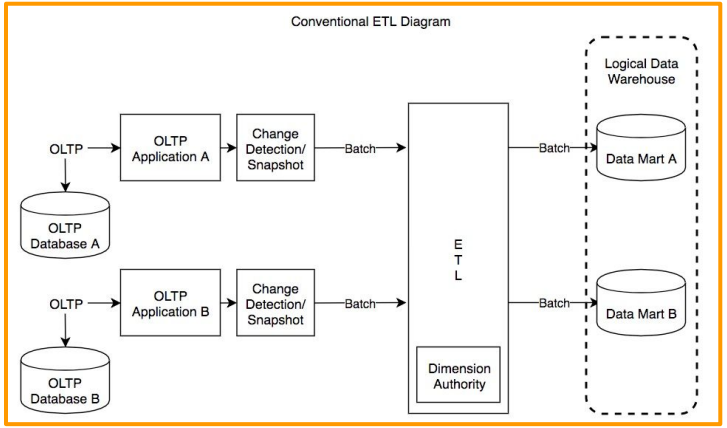
- Duas abordagens de integração de dados no data warehouse:
  - **Extract, transform, load (ETL)**
    - Dados são transformados e carregados no data warehouse no estado final
  - **Extract, load, transform (ELT)**
    - Dados são carregados no data warehouse e só aí transformados no estado final



TÉCNICO LISBOA

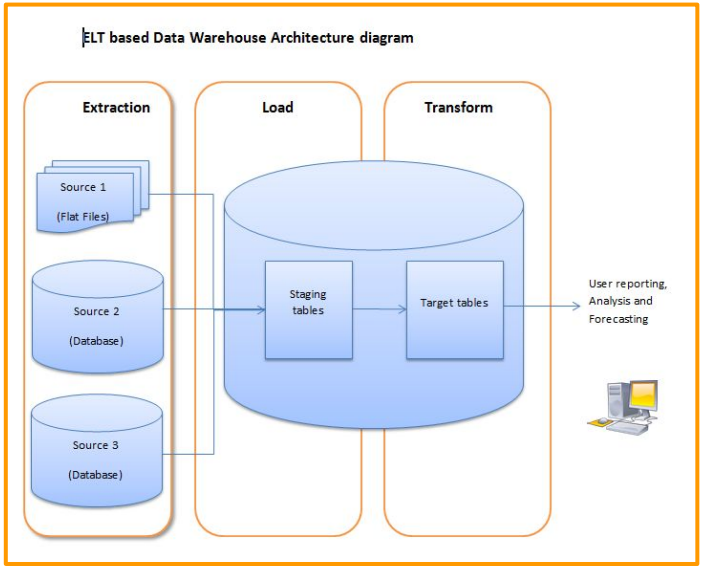
Data Warehousing

- **Extract, transform, load (ETL)**
  - Dados em bruto (área de staging) são transformados e integrados numa **área de integração**, frequentemente com recurso a uma base de dados *operational data store* (ODS)
  - Dados integrados são carregados no data warehouse e aí reorganizados (esquema em estrela)



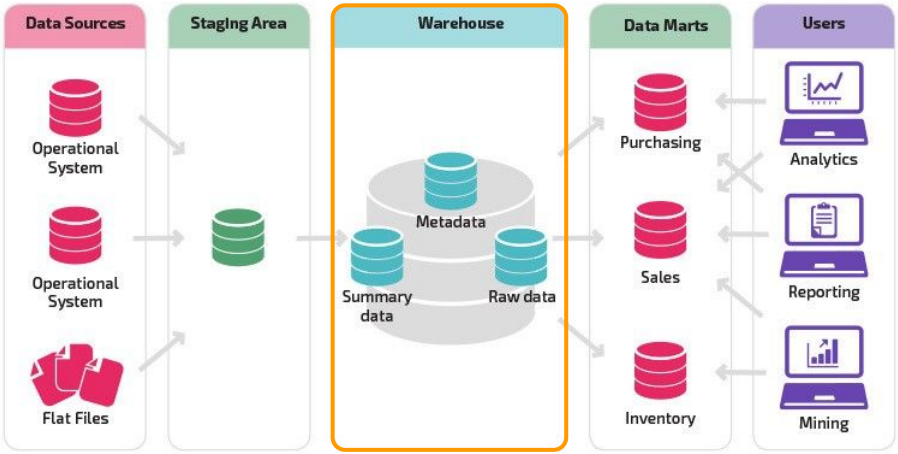
Data Warehousing

- **Extract, load, transform (ETL)**
  - A área de staging está contida no próprio data warehouse
  - Todas as transformações de dados são conduzidas dentro do data warehouse
  - Os dados transformados são introduzidos em tabelas finais (esquema em estrela)



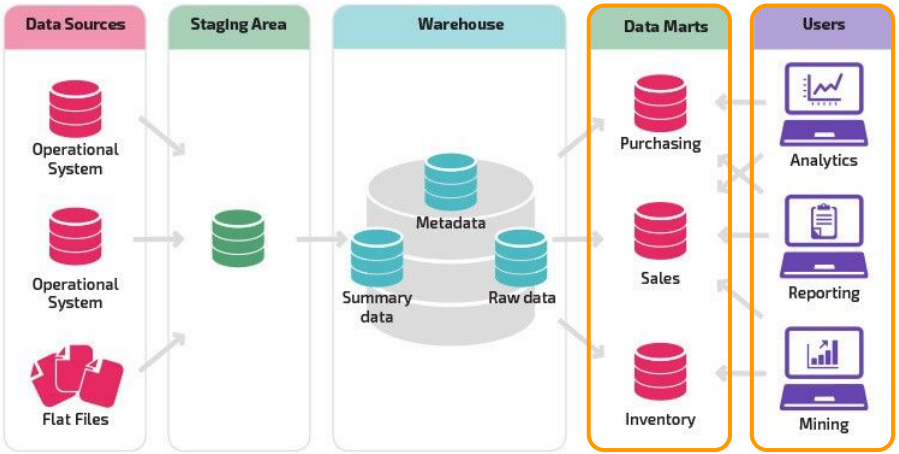
Data Warehousing

- O **warehouse** propriamente dito geralmente inclui
  - Dados em bruto, transformados mas não agregados
  - Dados agregados, com diferentes granularidades e perspectivas
  - Metadados que descrevem a proveniência e transformações dos dados



Data Warehousing

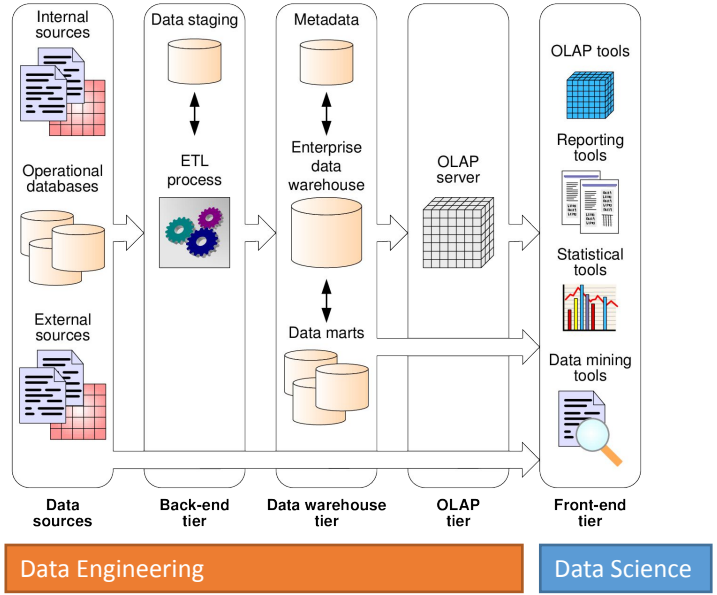
- **Data Marts** contêm um subconjunto dos dados no data warehouse orientado ao cliente
  - Oferecem uma vista dos dados adequada a um propósito específico (e.g. linha de negócio)
- **Utilizadores** acedem a um ou mais data marts através de interfaces configuradas para finalidades específicas



Data Warehousing

Desafios:

- **Heterogeneidade dos dados:** combinar e integrar dados de várias fontes e formatos
- **Integração semântica:** reconciliar esquemas de dados e representações de objetos
- **Atualização:** acrescentar novos dados, eliminar dados antigos
- **Gestão de metadados:** capturar proveniência e transformações



Data Warehousing

Atualização:

- **Source-driven architecture:** fontes de dados transmitem informação nova ao warehouse (de forma contínua ou periódica)
- **Destination-driven architecture:** warehouse periodicamente pede informação nova às fontes de dados
- **Replicação síncrona vs. assíncrona:**
  - Manter warehouse sincronizado com fontes de dados (e.g., two-phase commit) é geralmente demasiado dispendioso; é tipicamente aceitável ter dados ligeiramente desatualizados no warehouse

Data Warehousing

Transformação:

- Tipicamente requer **data cleansing**, e.g.:
  - Corrigir erros em moradas (typos, erros em códigos postais)
  - Fundir listas de moradas de várias fontes e eliminar duplicados
- Pode requerer **integração semântica**, e.g.:
  - Uniformizar diferentes formatos de morada (atributo único *vs.* vários atributos)
- Pode requerer **sumarização**: dados brutos podem ser demasiado volumosos para manter no warehouse e dados agregados podem ser suficientes

Data Warehouse *vs.* Data Lake

Data Lake:

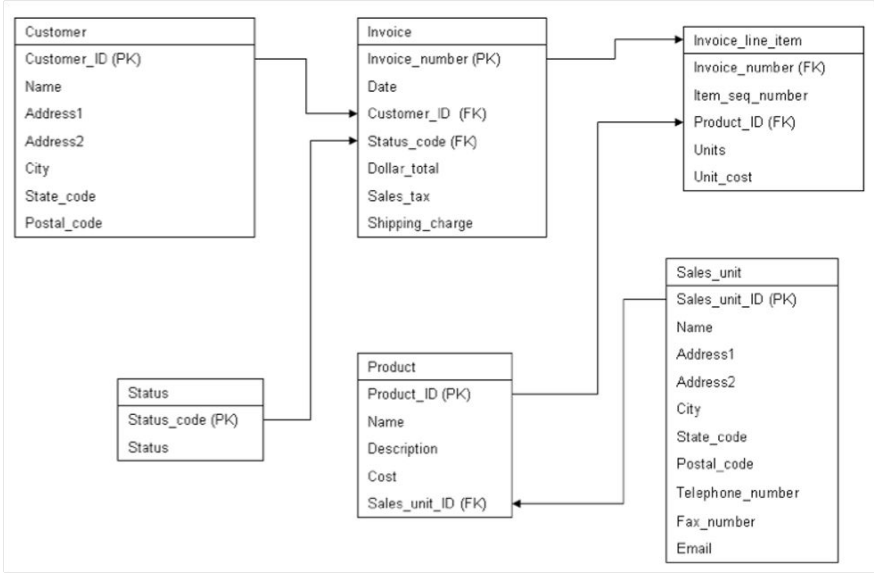
- Repositório que contém dados em múltiplos formatos sem integração de esquema
  - Para algumas aplicações não é necessária a transformação dos dados num esquema comum
- Trade-off:
  - **Data Warehouse:** maior esforço na organização e transformação dos dados
  - **Data Lake:** maior esforço na consulta de dados

TÉCNICO LISBOA

Limitações do Modelo Relacional Normalizado (BD Operacionais)

Modelo Relacional Normalizado

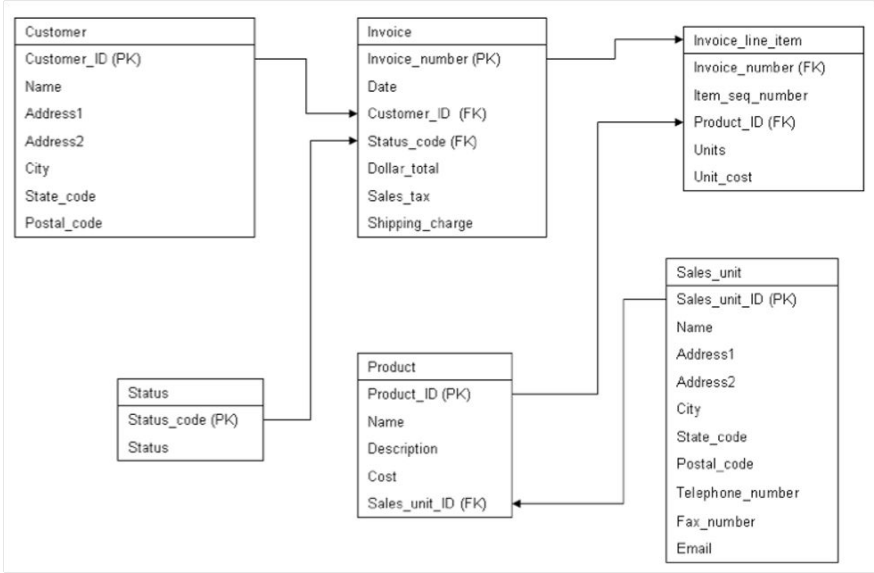
- Ideal para bases de dados operacionais (OLTP) em que há escrita frequente
  - Atomicidade dos dados minimiza custo de operações de escrita e evita inconsistências
- Pouco eficiente para processos analíticos devido à necessidade de atravessar várias tabelas (joins múltiplos)
  - Particularmente para agregações globais sobre os joins



Modelo Relacional Normalizado

Exemplos:

- Em que cidade(s) foram mais produtos vendidos no mês passado?
- Em que semanas, produtos e cidades observamos a maior variação de vendas para produtos em promoção?
- Quantos clientes compraram produtos da unidade A no primeiro mês deste ano?

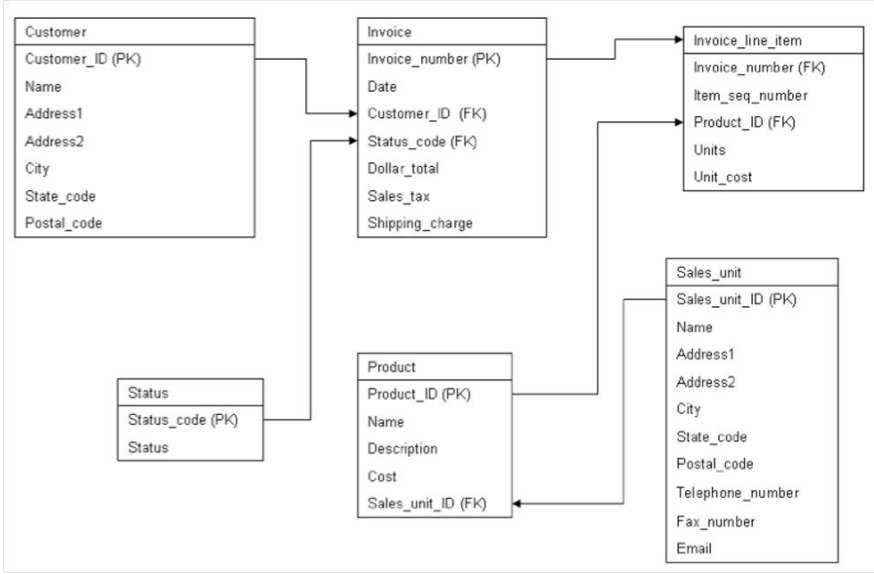


Modelo Relacional Normalizado

Exemplos:

- Quantos clientes compraram produtos da unidade A no primeiro mês deste ano?

```
SELECT COUNT(DISTINCT Customer_ID)
FROM Sales_unit SU
JOIN Product USING (Sales_unit_ID)
JOIN Invoice_line_item USING (Product_ID)
JOIN Invoice I USING (Invoice_number)
WHERE SU.Name = 'A'
AND EXTRACT(YEAR FROM I.Date) = 2023
AND EXTRACT(MONTH FROM I.Date) = 1;
```

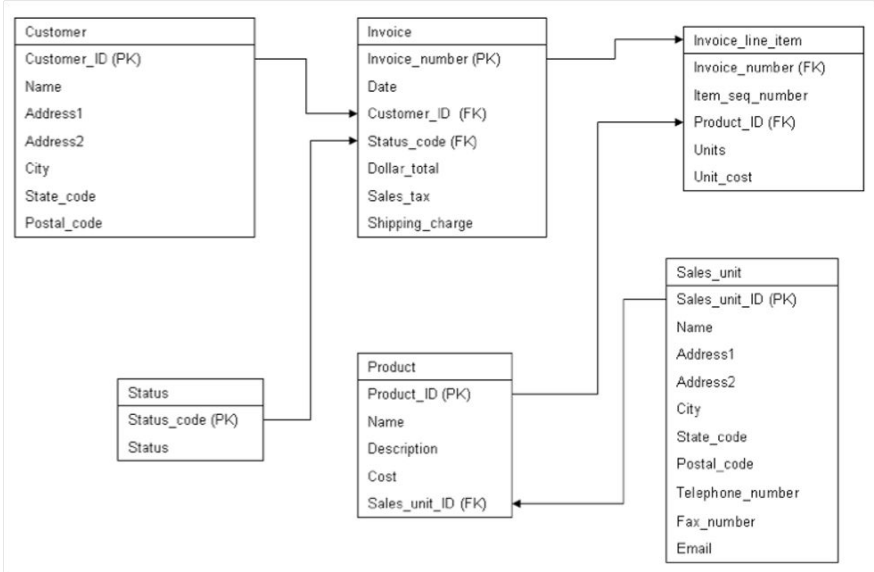


TÉCNICO LISBOA

Modelo Relacional Normalizado

Como tornar a análise de dados mais eficiente?

- Dados pré-agregados?
- Vistas (materializadas)?
- Outro modelo de dados?



TÉCNICO LISBOA

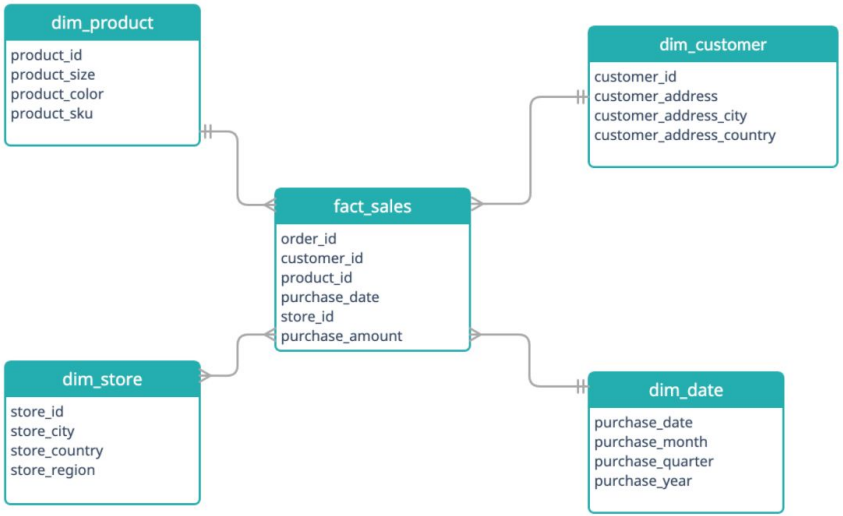
TÉCNICO LISBOA

Esquema em Estrela

Esquema em Estrela

Dois tipos de tabelas:

- **Tabela(s) de factos**
  - Grande dimensão
  - Frequentemente normalizada(s)
  - Objeto primário de análise de dados
- **Tabelas de dimensões**
  - Relativamente pequenas
  - Geralmente não normalizadas
  - Contém informação adicional sobre os elementos (ou dimensões) da tabela de factos



TÉCNICO LISBOA

Esquema em Estrela

Tabela(s) de factos:

- **Atributos-medida**: quantificam os factos e (geralmente) podem ser agregados
  - E.g. purchase_amount
- **Atributos-dimensão**: correspondem a dimensões sobre as quais os atributos-medida podem ser analisados
  - Geralmente índices numéricos que são chaves estrangeiras para as tabelas de dimensões
- A **chave** da tabela de factos é a combinação de chaves estrangeiras das várias tabelas de dimensões

Esquema em Estrela

Atributos-Medida:

- **Não-Aditivos:** não podem ser agregados em nenhuma dimensão
  - [Date, Product, Store, Margin] (‘margins’ não podem ser agregadas)
- **Aditivos:** podem ser agregados em todas as dimensões
  - [Date, Product, Store, Quantity]
- **Semi-Aditivos:** podem ser agregados nalgumas dimensões mas não em todas
  - [Date, Account, Balance] (‘balance’ não deve ser agregado em ‘Date’)

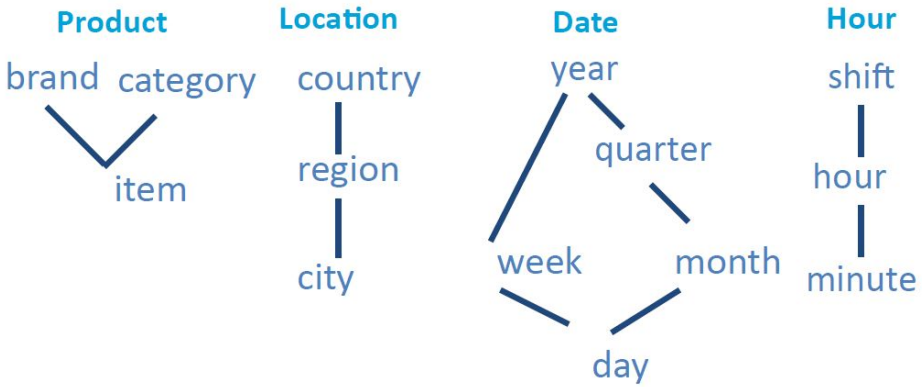
Tabelas sem atributos-medida são "factless":

- Factos podem ainda ser contados
  - [Date, Product]

Esquema em Estrela

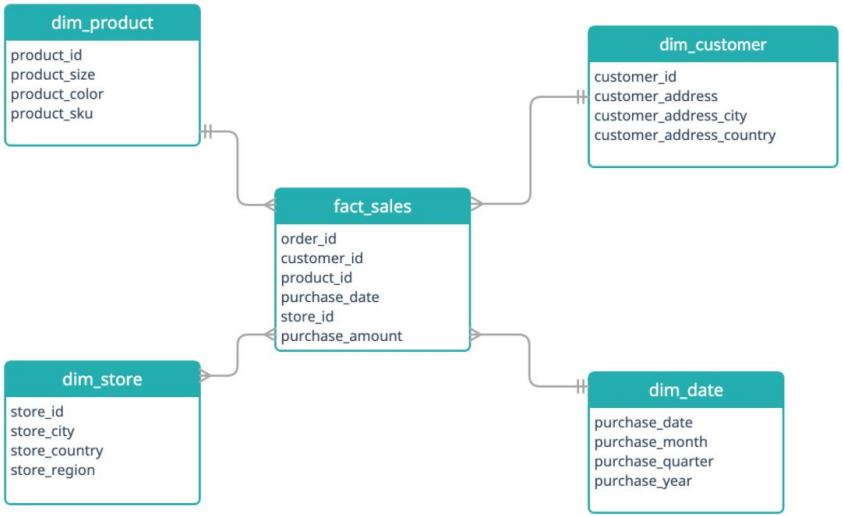
Tabelas de dimensões:

- Explicações dos factos: quem, onde, quando, o quê, …
- Contêm informação frequentemente redundante e hierárquica
  - Redundância é menos importante do que eficiência de acesso
  - Operações de escrita são raras



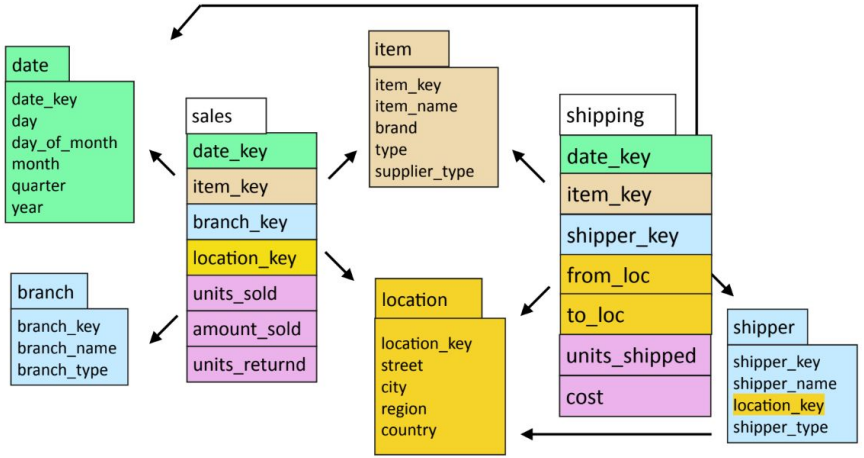
Esquema em Estrela

- **Query típica:**
  - Join entre a tabela de factos com uma ou mais tabelas de dimensões
  - Agrupamento em um ou mais atributos das tabelas de dimensões
  - Agregação sobre um ou mais dos atributos-medida da tabela de factos



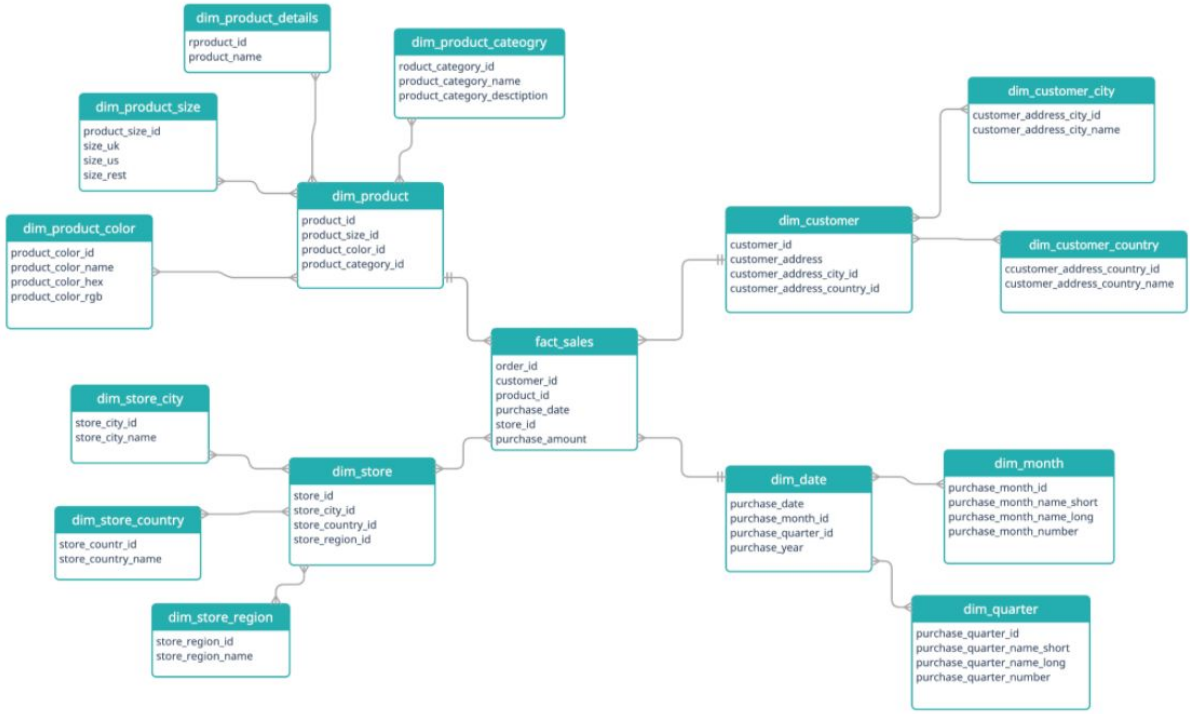
Variantes ao Esquema em Estrela

- **Constelação / Galáxia:**
  - Múltiplas tabelas de factos ligadas às mesmas tabelas de dimensões



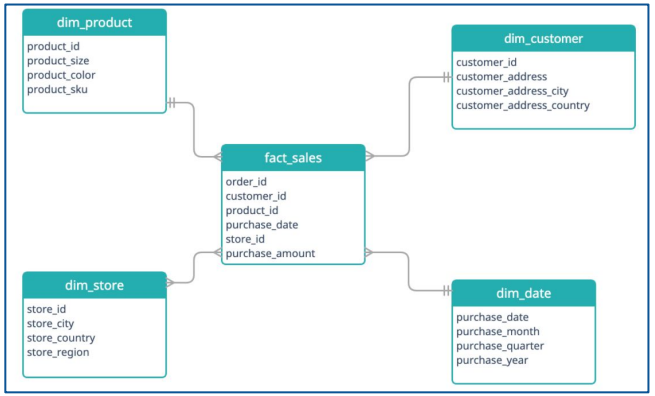
Variantes ao Esquema em Estrela

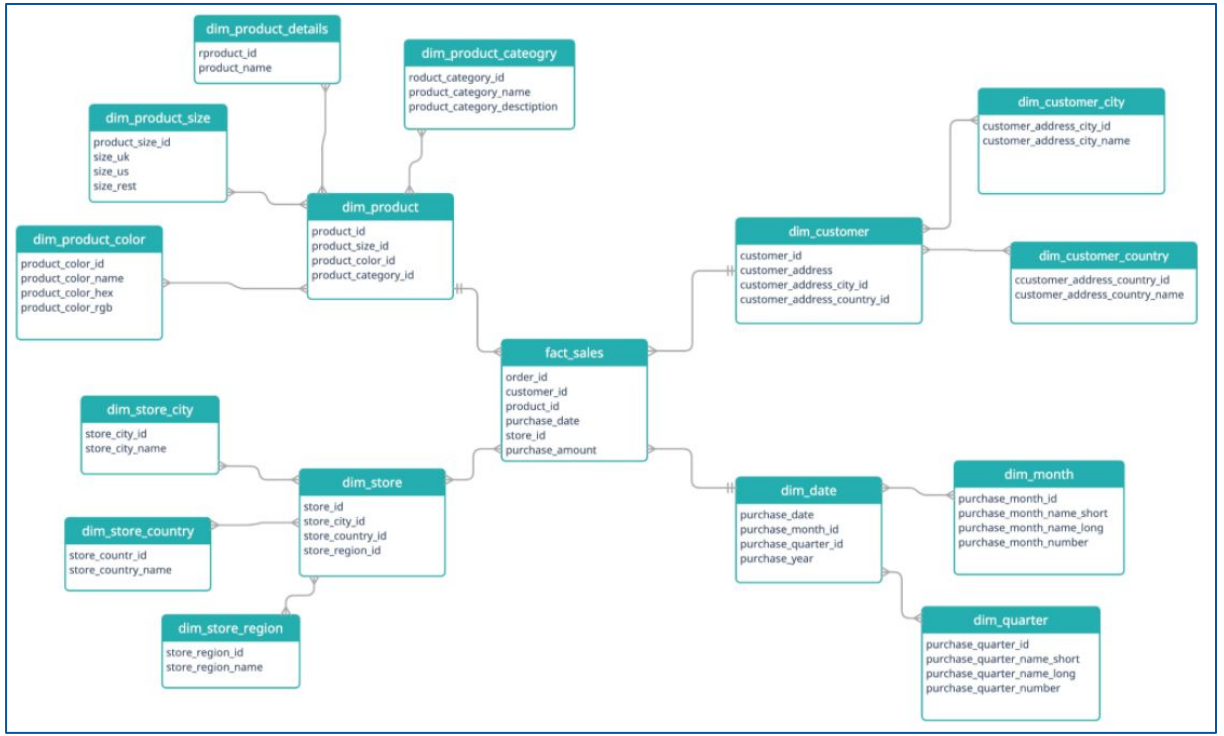
- **Floco de Neve (Snowflake):**
  - Tabelas de dimensões normalizadas dividindo em mais tabelas dimensionais (lookup tables)



TÉCNICO LISBOA

Esquema em Estrela vs. Snowflake





TÉCNICO LISBOA

Esquema em Estrela vs. Snowflake

Esquema em Estrela:

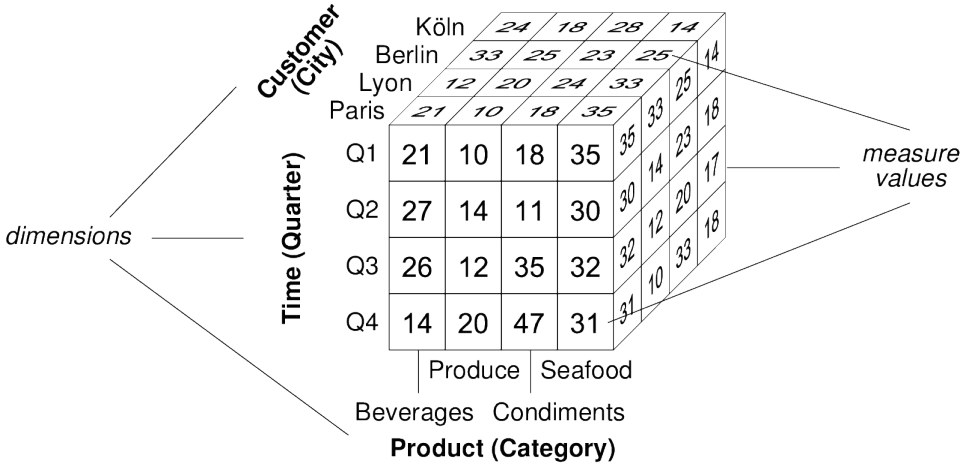
- Consulta mais eficiente (modelo de dados mais simples, menos joins)
- Armazenamento menos eficiente (redundância de dados)
- Potencial para problemas de integridade

Esquema Snowflake:

- Armazenamento mais eficiente (não há redundância)
- Consulta menos eficiente (modelo de dados mais complexo, exige mais joins)

OLAP

- O esquema em estrela pode ser visto como um hipercubo
- Essa é a lógica dos sistemas OLAP



A. Vaisman, E. Zimányi, *"Data Warehouse Systems: Design and Implementation"*, Springer, 2014

TÉCNICO LISBOA

TÉCNICO LISBOA

Data Warehouses e SGBD

Data Warehouses e SGBD

MOLAP

- Armazenamento baseado em colunas: arrays persistentes em disco
  - Arrays podem ser comprimidos, reduzindo custos de armazenamento, I/O e memória substancialmente
  - Queries apenas precisam de localizar os atributos relevantes, reduzindo custos de I/O e memória

ROLAP

- Implementação de esquema em estrela (ou snowflake) em SGBD relacionais
  - Menos eficiente (OK para data warehouses pequenos)

TÉCNICO LISBOA