

# INTEGRACION DE DATOS – CURSO 23/24

## PENTAHO PDI (KETTLER)

---

### TAREFA A

A cada grupo de prácticas se lle asigna, a través do Moodle, un dominio de discurso sobre o que xirará o seu traballo.

Esta tarefa consta dos seguintes pasos:

#### 1) Deseño da base de datos operacional

Debes deseñar unha pequena base de datos para almacenar a información **que corresponda ao teu dominio de discurso**. Deseña o **modelo E/R** e o **modelo relacional** resultante (especificando **todas as claves primarias e/ou foráneas**).

Requisitos:

- O modelo E-R debe incluír, como mínimo, **cinco** tipos de entidade **ou** relacións M:N.
- Polo tanto, o modelo relacional debe incluír como mínimo **5 relacións**

#### 2) Deseño dun ficheiro csv con información adicional

Debes deseñar un ficheiro *csv* que almacene información non existente na BD relacional. A idea é que, para encher o *Datamart*, precisaremos complementar a información da BD coa información recollida no ficheiro.

#### 3) Deseño dun pequeno *Datamart*

Debes deseñar un *Datamart* baseado na información dispoñible na base de datos operacional e no ficheiro *csv*, que despois actualizaremos mediante un proceso ETL. Deseña o modelo relacional resultante, especificando todas as claves primarias e foráneas.

Requisitos:

- O *Datamart* debe incluír **tres** táboas de dimensións e **unha** táboa de feitos.
- Unha das tres dimensións deberá ser unha dimensión **temporal** (día, mes, trimestre, ano...) con contido **precargado**.
- Para as outras dúas dimensións consideraremos inicialmente unha actualización SCD-I (sobreescritura) sobre todas as dimensións. En consecuencia, as táboas dimensionais respectivas deben incluír:
  - Como clave primaria, un **surrogado**.
  - A **clave natural** asociada na base de datos operacional (ou no ficheiro *csv*).
- A información da táboa de feitos debe obterse realizando algún tipo de **agrupación**. Por exemplo, realizando unha consulta SELECT sobre a base de datos operacional que inclúa unha cláusula **GROUP BY**; ou utilizando as funcionalidades de agrupación de Pentaho PDI.

## 4) Implementación

Debes realizar os seguintes desenvolvementos:

1. BD Operacional: implementa no teu SXBD a base de datos operacional deseñada no punto 1. Introduce algúns datos que permitan despois probar suficientemente o proceso ETL.
2. Ficheiro csv: implementa o ficheiro csv deseñado no punto 2. Éncheo tamén con algúns datos de proba.
3. BDDatamart: implementa tamén no teu SXBD a base de datos do Datamart deseñado no punto 3.

**AVISO: A BD operacional e a BD do Datamart deben ser implementacións separadas (Bases de datos independentes, scripts de creación separados).**

## Entregables

A entrega final deberá incluír, xunto co material a entregar como resultado das outras tarefas, o seguinte:

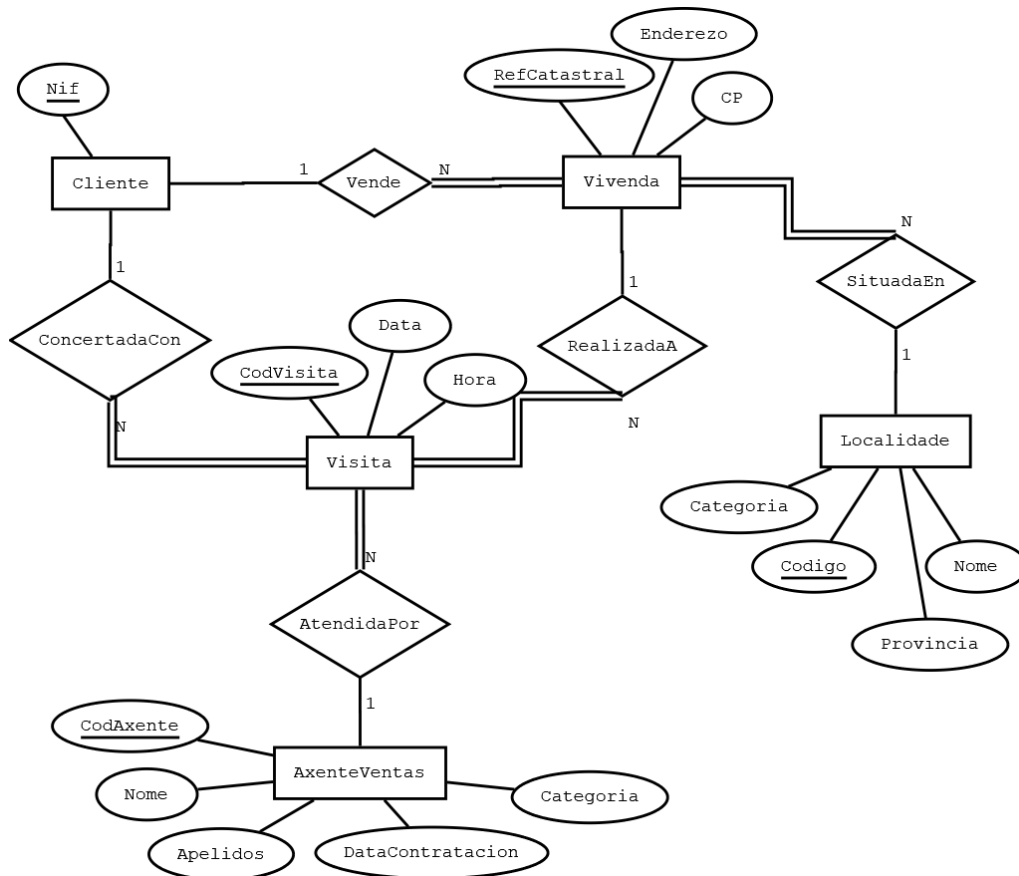
- **Modelo E/R e modelo relacional** DEFINITIVOS da BD operacional
- **Diagrama e modelo relacional** DEFINITIVOS do Datamart
- **Scripts** de creación das bases de datos (CREATE TABLE + INSERT)). Mínimo: un por cada BD
- Ficheiro **csv** con datos de proba rexistrados

## Exemplo

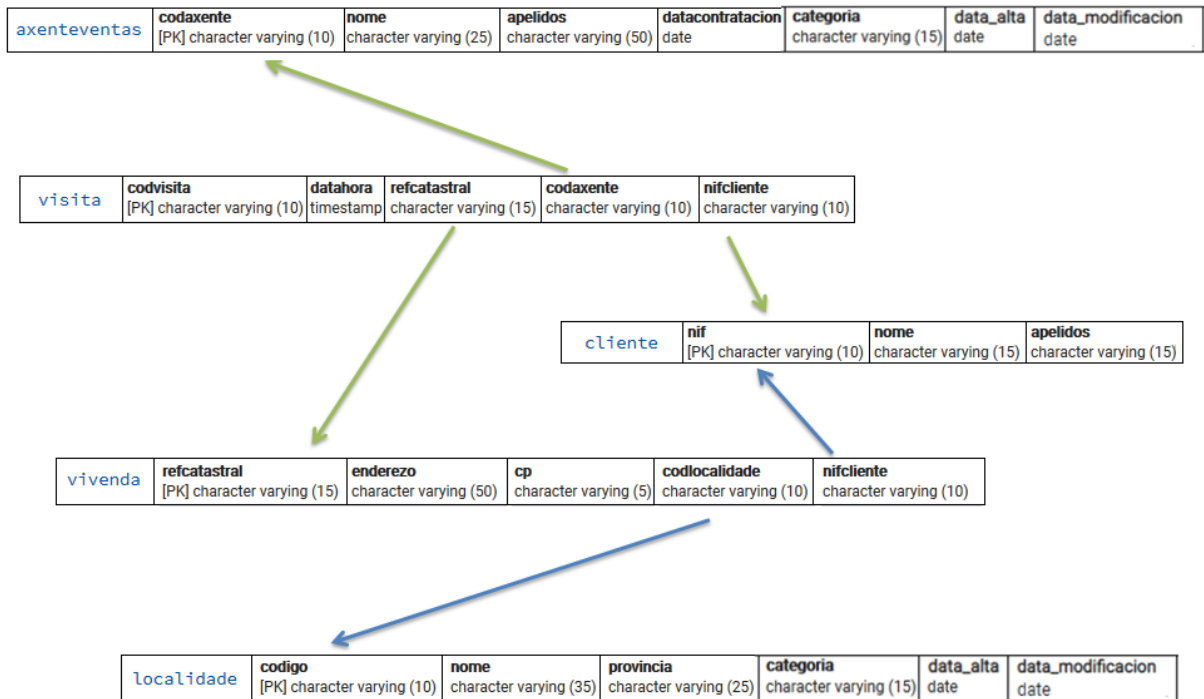
A seguir presentase un pequeno exemplo do traballo a realizar. Considerando como dominio unha axencia inmobiliaria, preséntanse:

### BD operacional: Modelo E/R

Este sería o modelo E/R proposto (que cumpre os mínimos esixidos). Gardamos información de vivendas en venta, para as que se organizan visitas de posibles compradores organizadas por axentes de vendas. [Aviso: Por claridade, omítense atributos non relevantes.](#)



## BD operacional: Modelo relacional



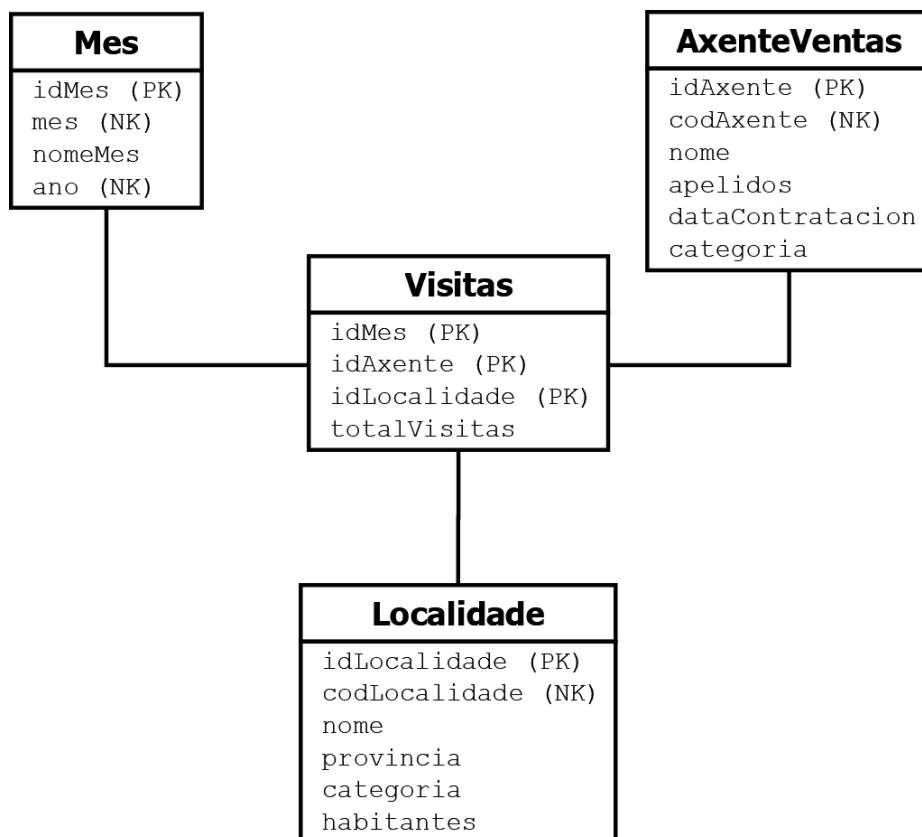
## Ficheiro CSV

Na BD recóllense o CP e localidade das vivendas. Pero a inmobiliaria está interesada en que o Datamart recolla o número de habitantes de cada localidade. Polo tanto, creamos un ficheiro csv para complementar a información da BD onde indicamos, para cada localidade rexistrada na BD operacional, o número que lle corresponde.

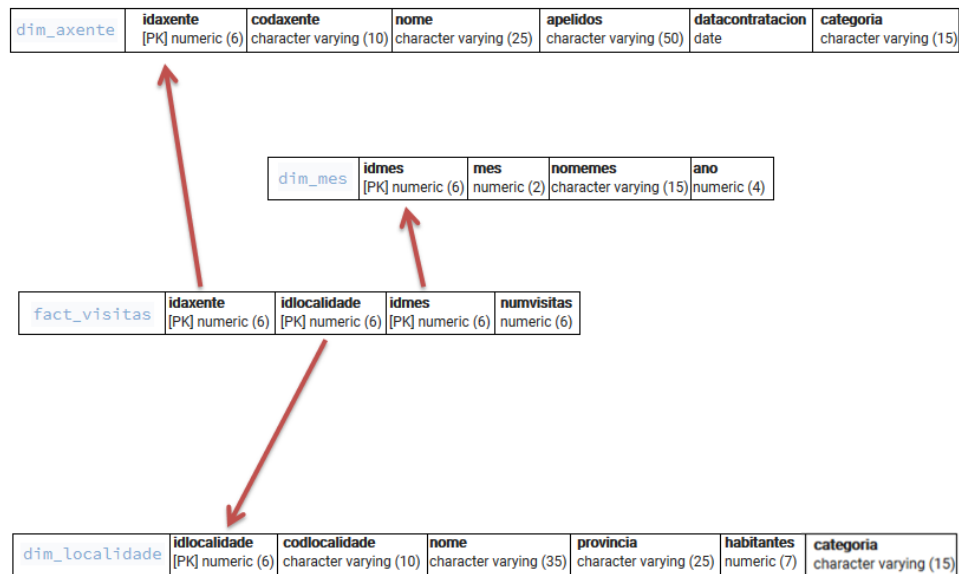
```
habitantes.csv X
localidade;habitantes
L001;244099
L002;95207
L003;98025
```

## Datamart: Diagrama

No Datamart queremos recoller, por cada combinación Mes / Axente / Localidade, o total de visitas realizadas. Este sería o diagrama resultante:



## Datamart: Modelo relacional



## Anexo: ¿Cómo exportar as bases de datos?

Se utilizas PostgreSQL como SXBD, podes xerar os scripts de creación das BDs do seguinte xeito:

BD operacional:

```
pg_dump -U postgres --no-owner --column-inserts --no-tablespaces  
-foperacional.sql <nome da vosa BD>
```

BD Datamart:

```
pg_dump -U postgres --no-owner --column-inserts --no-tablespaces  
-fdatamart.sql <nome da vosa BD>
```