Estàndard de la base de dades de l'Observatori de Turisme Sostenible

Frank William Hammond Espinosa

Darrera actualització: 18 de juny de 2025

Índex

A aquest document es detallarà la metodologia a seguir a l'hora de crear i omplir taules a la base de dades d'aquest repositori.

Per qualsevol dubte relacionat amb aquest document, enviau un mail a frank.hammond@fueib.org o avazquez1@conselldemallorca.net.

1 Com llegir aquest document

Abans de guardar qualsevol taula a la base de dades de producció, és important tenir clars diversos conceptes relatius a aquest escrit.

Primordialment, és necessari llegir detingudament la secció 2 ja que és específica d'aquesta base de dades, i tenir clars els estàndards de normalització d'un model de dades "floc de neu".

En segon lloc, si les dades a introduir contenen molts valors "faltants" o ambigus, s'han de seguir les directrius detallades a la secció 5.

2 Nomenclatura de taules i columnes

2.1 Taules

2.1.1 Semàntica

La majoria de taules presents a la BDD a dia d'avui estan pensades per ser utilitzades a quadres de comandament (*dashboards*) fets amb PowerBI. Per aquest motiu es segueix una nomenclatura inspirada en la manera habitual de separar la informació en el món de l'analítica de negoci.

Les taules es separen en dos tipus: de **dimensió** (*dimension*) i de **fets** (*facts*).

Definició 1 (Taules de dimensió)**.** Les **taules de dimensió** són les que contenen els detalls dels conceptes individuals que conté la base de dades. Sempre tenen un identificador únic per fila (*primary key* o clau primària) i tota la informació emmagatzemada dels elements que representen.

A aquestes taules normalment s'inclouen els noms i/o descripcions dels objectes modelitzats a la base de dades, sovint en diferents idiomes (una columna per idioma). També s'anoten les relacions a altres objectes de la base de dades en forma de *foreign keys* o claus foràknies. Per exemple, per

representar la relació "cada ciutat pertany a un país", a la taula de ciutat es guarda una columna amb una clau forània a l'identificador de la taula de país.

Per conveni, a aquesta base de dades les claus primàries sempre estaran formades per una única columna de tipus `INT` (enter).

Definició 2 (Taules de fets)**.** Les **taules de fets** són les que contenen les relacions "reals" que s'han enregistrat a la base de dades. Normalment, aquestes dades que són les que s'empren a nivell últim per construir gràfiques, i no es solen creuar (operacions `JOIN`) entre elles.

Sempre que sigui possible, les columnes d'aquestes taules han de contenir identificadors numèrics de tots els elements que contenguin, és a dir, claus forànies a alguna taula de dimensió. En cap cas podran contenir descripcions textuals (*strings*) que apareguin a més d'una fila. Veure la secció 3 per més detalls sobre guardar identificadors numèrics o descripcions textuals i la secció 2.1.3 per exemples de taules.

2.1.2 Notació

Primerament, els noms de les taules han d'anar prefixats pels literals `DIM_` o `FAC_` segons siguin de dimensió o de fets (veure les definicions 1 i 2 per a una explicació detallada dels conceptes). En cas que en un futur fos necessari ampliar la classificació de taules (per exemple, afegir taules de metadades), s'afegiria un prefixe de tres lletres seguit per un guió baix, semblant als definits prèviament.

Després del prefixe, s'inclou el nom de l'objecte o relació que modelitza la taula. Això, evidentment, depèn molt del cas concret, però s'han de seguir una sèrie de directrius generals que s'inclouen a continuació. Aquestes directrius també apliquen pels noms de les columnes, com es detallarà a 2.2.2. Tot i així, ha de prevalèixer el sentit comú del programador, ja que és una qüestió a decidir cas per cas.

Definició 3 (Notació general per noms)**.** Les directrius a seguir a l'hora de posar noms a objectes dins la base de dades seran les següents:

1. Es segueix la convenció informàtica "*screaming snake case*". Això vol dir que els noms han d'estar enterament en majúscules, composts únicament per caràcters alfanumèrics ASCII (nombres i lletres de l'alfabet anglès, sense accents) i barres baixes (_), i les paraules han d'estar separades per barres baixes en comptes d'espais. En cap cas s'admeten signes de puntuació llevat de la barra baixa. Això inclou espais en blanc, guions (-), barres verticals o inclinades (|, \, /), cometes (', "), signes de puntuació estàndard (,, ., ;, :), etcètera.

2. Els noms han d'estar en anglès sempre que sigui possible. Noms o sigles d'organitzacions internacionals sempre han d'incloure's en anglès també (per exemple, *UN* en comptes d'*ONU*).

   Excepcions notables a aquesta regla són els topònims o noms de demarcacions territorials nacionals o comunitàries, o organismes propis. Aquests es mantindran en la llengua o sigles locals, en castellà si són entitats nacionals i en català si són balears. Per exemple, una

hipotètica taula que guardi informació respecte els diferents consells insulars de les balears s'hauria d'anomenar `DIM_CONSELL_INSULAR`; per les diferents comunitats autònomes d'Espanya, una taula `DIM_COMUNIDAD_AUTONOMA` o `DIM_INE_REGION` per una taula de regions de països segons l'Institut Nacional d'Estadísitca (noti's que no es tradueixen les sigles).

En cas que sigui necessari guardar un nom d'aquest caire que contengui accents o enyes ("Ñ"), s'intentarà "ASCIIficar" aquest el màxim possible, conservant raonablement la fonètica; és a dir, lletres accentuades o amb dièresi es guardaran com la mateixa lletra sense accent o dièresi, i els caràcters "Ñ" es substituiran per "NY".

3. És preferible utilitzar noms sencers a diminitius o sigles - encara que això resulti en noms verbosos - sempre que els diminutius no siguin part del llenguatge quotidià. Per exemple, no s'ha de posar `CA` en comptes de `COMUNIDAD_AUTONOMA`, perquè no és immediat inferir a què es refereixen aquelles sigles. En canvi, sí és recomanable emprar `UN`, `INE` o `FRONTUR` com a part d'un nom perquè són abreujacions ben establertes i que es podrien emprar com a part d'una frase oral en l'idioma corresponent sence necessitat d'aclarir què signifiquen.

Addicionalment, s'hauran de seguir indicacions que són específiques per a taules.

Definició 4 (Notació específica per noms de taules)**.** Si la taula modelitza algun concepte en mode que cada fila correspongui a un únic objecte d'aquell tipus, el nom ha d'anar en singular (e.g. `DIM_CITY` en comptes de `DIM_CITIES` per a una taula que guardi informació referent a ciutats).

2.1.3 Casos d'estudi

This is work in progress.

2.2 Columnes

2.2.1 Semàntica

Normalment, tota columna a la base de dades pertany a una, i només una[1], de les següents categories:

Definició 5 (Classificació de columnes)**.** Els tipus de columnes a identificar són els següents:

1. **D'identificador intern**. S'hi guarda la *primary key* (clau primària) de la taula en qüestió. Obligatòriament a les taules de dimensió s'ha de fer servir una única columna com a clau primària, i aquesta ha de tenir el tipus `INT`.
2. **D'identificador extern**. S'hi guarden *foreign keys* (claus forànies) a altres taules. Idealment, només hi haurà claus forànies a taules de dimensió.

[1] Una excepció notable a això és la taula de calendari i totes les claus forànies a ella. Això es discuteix detalladament a la secció 4.1.

3. **De codi**. Contenen codis únics però que no actuen com a clau primària. Per exemple, cada país té associat un codi únic per l'INE i tres codis, cadascun d'ells únics, per la ONU; a més a més, tenen un identificador únic, exclusiu d'aquesta base de dades. Idealment, estan anotades amb la restricció `UNIQUE`.

4. **De data**. Contenen dates que corresponen a un dia o els límits del període de temps associat a la data que es registra. En cas que el registre vengui associat a un dia, es farà servir una única columna que contindrà el dia; i, en cas que vengui associat a un període de temps superior a un dia (mes, trimestre, any...), es faran servir dues columnes de data: una guardarà la data d'inici, i l'altra el dia immediatament posterior a la data de finalització. Per exemple, una fila corresponent a dades agregades a tot el mes de gener de 2024 tindrà enregistrades les dates d'1 de gener de 2024 i 1 de febrer de 2024. El tipus de dada a guardar a aquestes columnes serà `DATE`.

5. **De marca temporal**. Contenen marques temporals (data i hora). Si el registre ve associat a una marca temporal concreta, es guardarà només una columna; i si ve associat a un interval de temps se'n guardaran dues: una per l'inici i l'altra pel final. El tipus de dada a guardar a aquestes columnes serà `DATETIME2`.

6. **Numèriques**. Contenen valors numèrics que no siguin codis ni identificadors (per exemple, el valor d'una mesura o una quantitat de persones). Sovint, són mesures o quantitats que tenen associada una unitat. Sempre que sigui possible i això no introdueixi problemes computacionals o de precisió, s'empraran unitats "sense prefixe" (e.g. metres en comptes de kilòmetres) o, per contejos, el nombre sense agregar (per exemple, en comptes de tenir una columna de milers o milions de turistes, s'hauria de tenir una de nombre de turistes, i enregistrar 1000000 de turistes en comptes de 1000 milers o 1 milió de turistes). El tipus de dada a guardar serà `INT` o `FLOAT`[2], segons si els nombres involucrats són enters o no.

7. **Booleanes**. Contenen informació que només pot prendre dos valors. Es guardaran únicament dos nombres a aquesta columna: 1 i 0. El nom de la columna ha de ser prou descriptiu com per poder saber quin dels dos possibles valors correspon a 1 (vertader) i quin a 0 (fals). El tipus de dada a guardar a aquestes columnes serà `BIT`, i obligatòriament es posarà la restricció `NOT NULL`.

8. **De nom**. Aquestes columnes només estan presents a les taules de dimensió. Corresponen al nom que se li dona a l'objecte en un idioma en concret. En particular, a una taula donada hi haurà una columna de nom per a cadascun dels idiomes en què es registri la informació d'aquella taula. El tipus de dada a guardar a aquestes columnes sempre ha de ser `NVARCHAR`, idealment amb una longitud màxima sensible (s'ha d'evitar, en la mesura del possible, emprar `NVARCHAR(MAX)`[3]).

[2] És possible triar de manera explícita la precisió que es desitja a la representació de punt flotant que guardi la base de dades. Per defecte, SQLServer empra 8 bytes (https://learn.microsoft.com/en-us/sql/t-sql/data-types/float-and-real-transact-sql?view=sql-server-ver16), que és el màxim que es pot guardar a un FLOAT. Si és necessari tenir més precisió, s'ha d'emprar el tipus DECIMAL(p,r) amb valors de p i r adequats (cf. https://learn.microsoft.com/en-us/sql/t-sql/data-types/decimal-and-numeric-transact-sql?view=sql-server-ver16).

[3] És millor emprar NVARCHARs limitats perquè ocupen menys espai en memòria, però si es limita la longitud i s'intenta introduir text més llarg que aquella longitud, simplement es truncarà el text. Tingueu-ho en compte a l'hora de posar la fita superior.

9. **De descripció**. Aquestes columnes només estan presents a les taules de dimensió. Contenen descripcions més detallades dels objectes, a un idioma concret. A una taula donada, només pot haver una columna de descripció si ja hi ha una columna de nom a aquella taula i es necessita guardar una explicació més llarga, o si no té sentit posar-li un nom però sí una descripció a aquell objecte. El tipus de dada a guardar a aquestes columnes serà `NVARCHAR(MAX)`.

2.2.2 Notació

Normalment, els noms de les columnes segueixen el següent format:

$$[\text{PREFIXE\_}]\text{NOM}[\text{\_SUFIX}]$$

On "`NOM`" correspon al nom de la columna i és obligatori, i "`PREFIXE`" i "`SUFIX`" són opcionals i la seva utilització ve detallada a la definició 6. Les barres baixes entre nom i afixes només s'han de posar si aquests estan presents, i els claudàtors no s'han d'incloure: només estan presents aquí per a indicar que els afixes són opcionals.

Per triar un nom de columna, s'han de seguir les directrius generals de la definició 3, a més d'una sèrie de consideracions específiques per columnes, que depenen directament del seu tipus (aquest, al seu torn, ha de ser determinat segons la definició 5). A continuació es recullen les regles a seguir per a cada tipus:

Definició 6 (Notació de columnes)**.** La notació a seguir pels noms de les columnes depèn del seu tipus segons la definició 5. A saber, per a cada tipus de columna, són:

1. **D'identificador intern**. Sempre que no sigui ambigu, el nom d'aquestes columnes serà senzillament "`ID`", i no tindran prefix ni sufix.
2. **D'identificador extern**. El nom d'aquestes columnes serà `ID_`, seguit pel nom de la taula l'ID de la qual es referencia, sense prefixes de taula. Per exemple, per a una referència a `DIM_COUNTRY.ID`, la columna de clau forània s'hauria de dir `ID_COUNTRY`.
3. **De codi**. El nom d'aquestes columnes anirà prefixat per `CO_`, seguit d'un nom prou descriptriu del que es codifica.
4. **De data**. El nom d'aquestes columnes dependrà del tipus de data que emmagatzemin i el context de la taula: si és una data única i no hi ha confusió possible respecte què representa aquella data al context de lataula, simplement serà "`DATE`". En cas que hi hagi confusió (per exemple, si hi ha més d'una data o rang de dates a una mateixa taula; o si el nom de la taula no permet inferir clarament què significa la data), s'emprarà el prefixe "`DA_`", seguit per un nom prou descriptiu. En cas que es guardin rangs de dates, s'empraran els sufixes "`_START`" i "`_END`" pels extrems inferior i superior, respectivament.
5. **De marca temporal**. La notació a seguir per a aquest tipus de columna serà igual que per les de data, llevat de les següents substitucions: en comptes de "`DATE`" s'emprarà "`TIMESTAMP`"

sota el mateix pretext d'unicitat i no ambiguïtat; i, en comptes de "`DA_`" com a prefixe, s'emprarà "`TS_`".

6. **Numèriques**. Quan aquestes columnes guardin qualque mesura física amb unitats associades, s'ha d'explicitar, com a sufixe, quina unitat de mesura té, en sense abreujar i en ortografia nordamericana. Per exemple, s'ha d'emprar "`_METERS`" o "`_LITERS`", i no "`_METRES`" ni "`_L`". Al cas de guardar valors monetaris, sí s'emprarà l'abreujació usual de tres lletres, segons l'estàndard ISO 4217. En són exemples "`_EUR`", "`_USD`", "`_JPY`"...

7. **Booleanes**. El nom d'aquestes columnes anirà prefixat sempre per "`SW_`". A més, el nom en sí ha d'estar sempre en afirmatiu; és a dir, els valors "TRUE" a aquella columna han de corresponde a entrades que satisfacin la condició. Per a aconseguir això, sol esser bastant útil emprar la paraula "is". Per exemple, per a una hipotètica columna que especifiqui si aquella dada s'ha extret d'una font oficial o no, en comptes d'emprar com a nom "`SW_OFFICIAL`" (no queda clar què vol dir "TRUE" i què vol dir "FALSE"), s'ha d'emprar "`SW_IS_OFFICIAL`", en mode que els valors "TRUE" corresponguin a registres que, efectivament, són oficials.

8. **De nom**. El nom d'aquestes columnes anirà prefixat sempre per "`TX_`" (de "TeXt"), i sufixat sempre per un codi de dues lletres corresponent a l'idioma en què està escrit el nom: "`ES`", "`CA`" i "`EN`" pel castellà, català i anglès, respectivament. Si l'objecte que modelitzen ve explicitat pel nom de la taula (hauria de ser així a la majoria de casos), s'ha d'evitar la repetició i simplement posar "`TX_ES`" (o semblant, segons l'idioma). Per exemple, a la taula "`DIM_COMUNIDAD_AUTONOMA`", basta posar "`TX_ES`" per guardar el nom de les comunitats pertinents, i no s'ha d'emprar "`TX_COMUNIDAD_AUTONOMA_ES`".

9. **De descripció**. La notació a seguir per a aquestes taules serà igual que per les de nom, llevat que en comptes del prefixe "`TX_`", s'haurà d'emprar "`DS_`" (de "DeScription").

2.2.3 Casos d'estudi

This is work in progress.

3 Model de dades i normalització

A decidir.

4 Taules especials

Algunes taules a la base de dades no segueixen totes les regles detallades a aquest document per motius específics.

A continuació s'inclou una llista completa de tals taules, així com una descripció detallada de quines convencions violen i una explicació dels motius pel qual ho fan.

En cas d'introduir una taula així a la base de dades, serà necessari afegir una subsecció a aquesta secció del document.

4.1 Calendari

5 Emmagatzematge de valors faltants, imprecisos o provisionals.

5.1 Valors faltants

Definició 7. Una situació comuna a l'hora de tractar amb dades és trobar valors "faltants". Normalment, aquests es representen a programes informàtics com a `NULL` (SQL), `None` (Python), `NA` (R), `N/A` (Excel), `NaN` (pandas, numpy, Matlab) o semblants, o 0 a alguns casos, si les dades són numèriques. A aquest text, anomenarem **vàcua** a qualsevol dada d'aquest caire.

Direm que una dada vàcua és **numèrica** si les dades no vàcues de la seva agrupació (columnes a SQL o pandas, variables a R...) són numèriques (cf. la definició 5); altrament, direm que és **no numèrica**.

Si bé la manera ideal de tractar dades vàcues depèn molt de la semàntica específica de cada conjunt de dades, proposam una sèrie d'heurístiques generals a seguir dins aquesta base de dades. Com sempre, ha de prevalèixer el criteri del programador.

Definició 8. Per representar conjunts de dades amb entrades vàcues a aquesta base de dades, s'han de seguir, en ordre, les següents indiciacions:

1. Si la presència d'una dada vàcua no numèrica té un signficat clar dins el context del conjunt de dades en què se troba, s'haurà de guardar un registre a la taula corresponent en base de dades que contengui `NULL` a la columna pertinent. Per exemple, si es guarden dades sobre el guanyador de una competició anual i es sap amb certesa que un any no va haver guanyador per causes sobrevingudes, a la columna corresponent al guanyador haurà d'haver un valor `NULL`.
2. Si la dada vàcua és numèrica, la seva presència té un significat clar dins el context del conjunt de dades en què se troba, i aquest ha de ser emprat amb finalitats estadístiques, queda a criteri del programador si substituir-la per algun estadístic de la resta de dades (mitjana, mediana, mínim, màxim...) o per 0.
3. En qualsevol cas, si existeix dubte raonable respecte el significat de la dada vàcua i la necessitat de guardar-la, **no s'haurà de guardar** cap registre per a aquella dada. És a dir, es prefereixen representacions denses de les dades a representacions *sparse* no justificables.

5.2 Valors provisionals.

A decidir.

5.3 Valors imprecisos.

A decidir.