UNIVERSITAT DE BARCELONA

Tema 1: Magatzems de dades

Bases de dades avançades curs 17/18

Enric Biosca Trias ebiosca@maia.ub.es

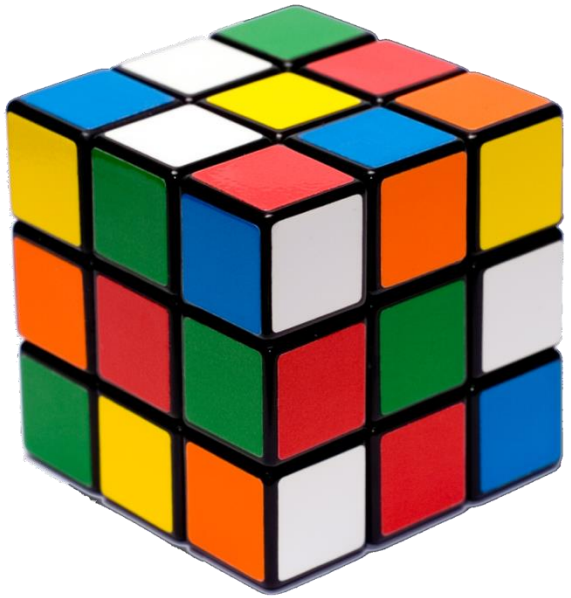
Dept. Matemàtica Aplicada i Anàlisi.

Universitat de Barcelona

UNIVERSITAT DE BARCELONA

Magatzems de dades

1. Business Intelligence
2. Arquitectura i disseny del datawarehouse
3. Disseny datamart exemple
4. Rendiment del datawarehouse



Magatzems de dades

Definició de BI

Entenem per business intelligence (BI) el conjunt d'estratègies i eines dirigides a l'administració i creació de coneixement a través de l'analisi de dades existents en una organització.

Característiques Principals d'un sistema de Business Intelligence:

- **Accesibilitat a la informació.** Les dades són la font principal de la BI. El més important que han de garantir aquestes eines es l'accés a les dades per part dels usuaris amb independència del seu origen.

- **Ajuda a la presa de decisions.** Es vol anar més enllà alhora de presentar la informació, de manera els usuaris tinguin accés a les eines d'analisi que els permetin seleccionar i navegar només aquelles dades que siguin del seu interés.

- **Orientació a l'usuari final.** Es busca la independència entre els coneixements tècnics dels usuaris i la seva capacitat per utilitzar aquestes eines. Els consumidors de la informació són la part superior de la jerarquia empresarial

Magatzems de dades

Definició de BI

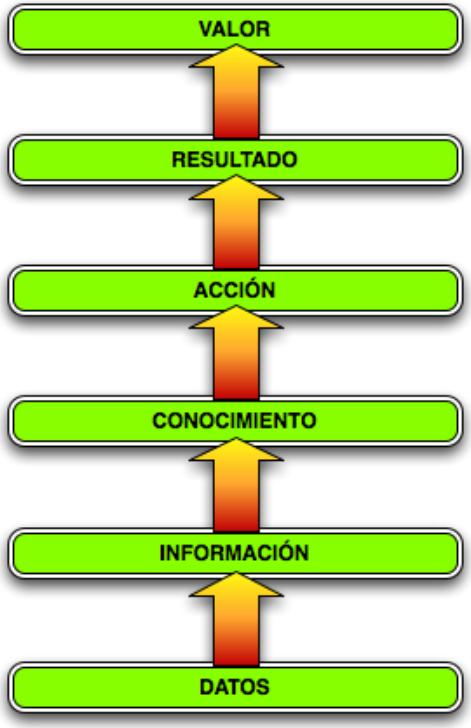
Actualment la informació empresarial es molt útil i dóna molt de poder a les organitzacions

L'objectiu principal de la Business Intelligence es convertir les DADES dels nostres sistemes en CONEIXEMENT per poder analitzar

Magatzems de dades

objectiu d'un projecte de BI

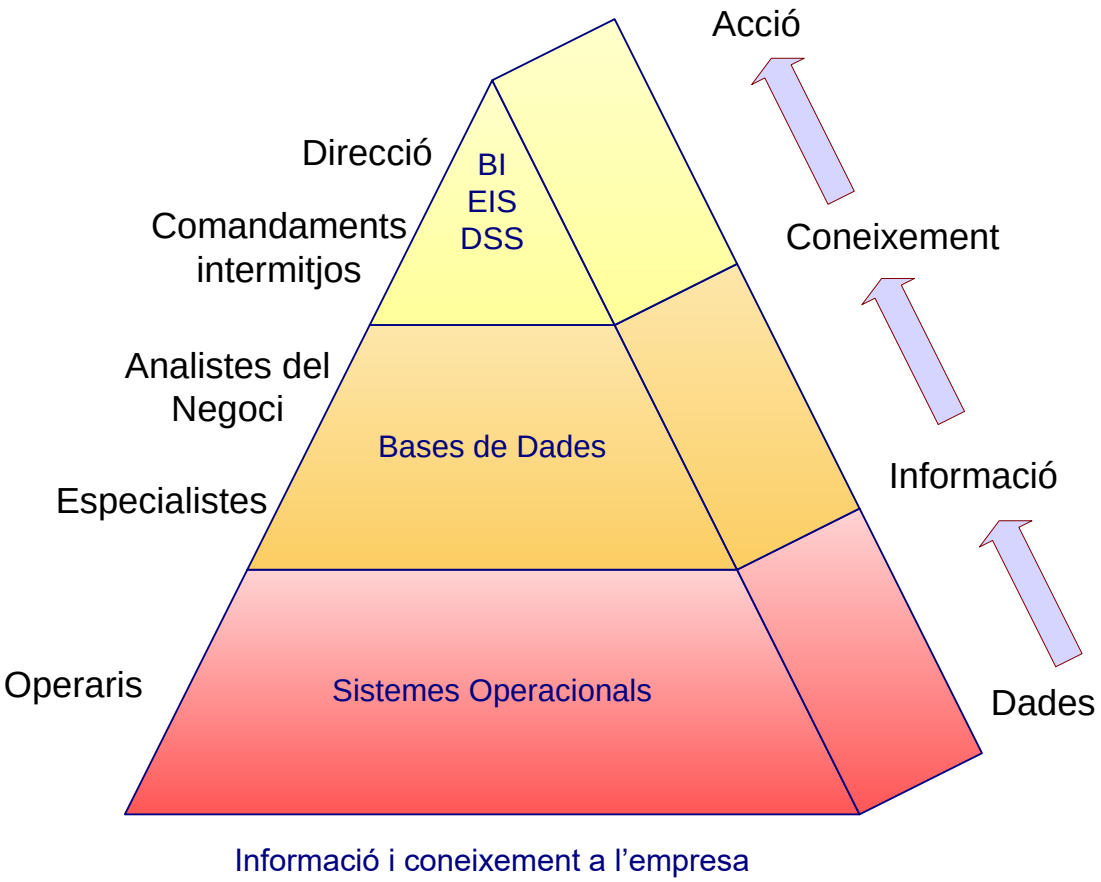
L'objectiu d'un projecte de BI és aportar valor afegit a la presa de decisions. Partint de les dades existents en l'organització.



Magatzems de dades

objectiu d'un projecte de BI

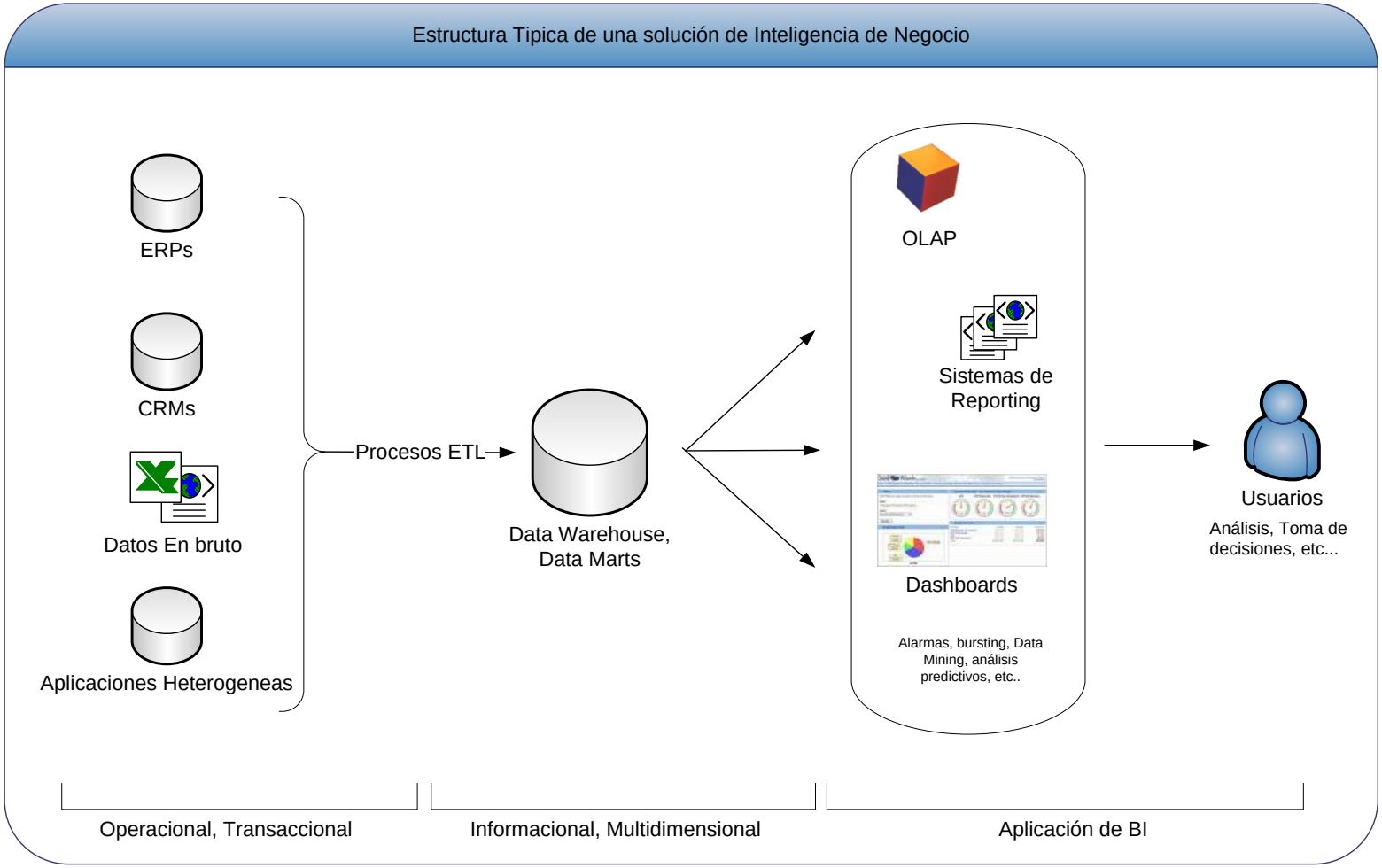
Una altra manera de veure els projectes de BI



UNIVERSITAT DE BARCELONA

Magatzems de dades

objectiu d'un projecte de BI



Definició i objectiu d'un Datawarehouse

Un magatzem de dades és una col·lecció de dades orientades al tema, integrades, no volàtils i historiades, organitzades per a donar suport a processos d'ajuda a la decisió.

Els principals objectius d'un DWH són:

- Ajudar en la presa de decisions.
- Segmentar les dades de negoci.
- Gestionar el coneixement de l'empresa.
- Depurar les dades.

Proporciona una visió global, comú i integrada de les dades de l'organització, independent de com s'han d'utilitzar posteriorment pels consumidors o usuaris, amb les propietats següents:

estable, coherent, fiable i amb informació històrica.

Conceptes Clau

Data warehousing

Data warehousing és el procés d'extreure i filtrar dades de les operacions comuns de l'organització, procedents dels diferents sistemes d'informació operacionals i / o sistemes externs, per transformar-los, integrar-los i emmagatzemar-los en un dipòsit o magatzem de dades (Data Warehouse, en anglès) per tal d'accedir-hi per donar suport al procés depresa de decisions d'una organització.

Datawarehouse

Proporciona una visió global, comú i integrada de les dades de l'organització, independent de com s'han d'utilitzar posteriorment pels consumidors o usuaris, amb les propietats següents:
estable, coherent, fiable i amb informació històrica.

Conceptes Clau

Datamart

Conté subconjunts de les dades del magatzem de dades, adaptats per analitzar un departament o àrea en concret.

Quina diferència hi ha llavors entre un datamart i un Datawarehouse?

El seu abast. El data mart està pensat per cobrir les necessitats d'un grup de treball o d'un determinat departament dins de l'organització.
En canvi, l'àmbit del Data Warehouse és l'organització de forma global.
El Datawarehouse és el magatzem natural per les dades corporatives comuns, el data mart proporciona el magatzem natural per les dades departamentals.

Principals diferències entre OLPT I DWH

	OLTP	DWH
Objectius	Operacionals	Informació per a la presa de decisions
Orientació	A l'aplicació	a la persona
Vigència de les dades	Actual	Actual + històric
Granularitat de dades	Detallada	Detallada + resumida
Organización	Organització normalitzada	Organització estructurada en funció de l'anàlisi a realitzar
Modificacions de dades	Constants	Estable

El model dimensional distingeix tres elements: fets, mètriques i dimensions.

Taula de fets: és la representació en el datawarehouse dels processos de negoci de la organització. pex: les vendes

Dimensió: és la representació en el datawarehouse d'una vista per a un determinat procés de negoci. Si reprenem l'exemple d'una venda, tenim com a dimensions el client que ha comprat, la data en la que s'ha realitzat la compra,…
Aquests conceptes poden ser considerats com a vistes.

Mètriques: son els indicadors de negoci d'un procés de negoci.
Els conceptes quantificables que permeten mesurar el procés de negoci.
Per exemple, en una vendra tindriem l'import d'aquesta.

Taules de fets

La taula de fets conté totes les dades que són rellevants per al negoci de l'entitat.

Cada registre de "fets" està compost per un conjunt de claus foranes (una clau per cada dimensió) i un o diversos valors numèrics.

Els valors vénen determinats pels objectius de l'organització i poden incloure mesures com nombre d'habitants, import de les subvencions, nombre d'inspeccions, edat mitjana dels càrrecs electes, etc ...

Generalment, els valors són emprats per calcular altres valors. Per exemple, dividint el nombre d'immigrants per la població total obtenim el percentatge d'immigrants de cada població.

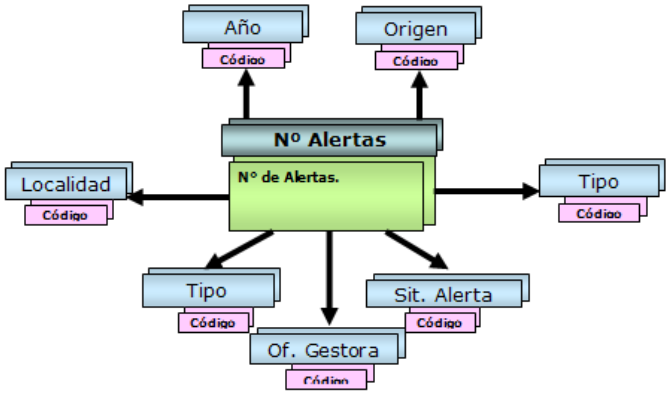
De forma habitual, aquest tipus de càlculs es duen a terme durant el processament de la consulta sol · licitada, per aquest motiu, aquest tipus de valors solen denominar "valors virtuals".

Taules de Dimensions

Les dimensions descriuen fets, les taules de dimensions contenen nombrosos atributs que permeten descriure un fet en major detall.

La construcció de consultes multidimensionals sobre un model en estrella (multidimensional) és molt senzill, ja que el format general d'aquestes consultes segueix el següent format:

Dona'm el <fet> de la <dimensió> de la <dimensió> .... de la <dimensió>.



Els principals components d'un model multidimensional són:

• Facts o "fets": variables numèriques que mesuren el negoci.
• Dimensions: principals eixos d'anàlisi de la informació.
• Nivells o atributs: categories dins d'una dimensió.
• Jerarquies: ordenacions dels atributs en la dimensió.

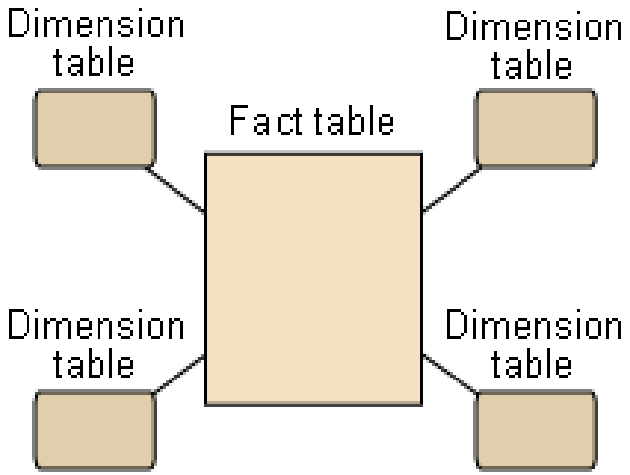
Magatzems de dades

Arquitectura i disseny del datawarehouse

Els conceptes bàsics-fets, mètriques i dimensions
(facts, measures i dimensions) - es representen en el model com relacions (taules) dins d'un esquema dimensional. Segons les tècniques de modelatge utilitzades, aquest esquema dimensional pot adoptar forma d'estrella o de floc de neu, el que dóna lloc als dos principals esquemes de representació:

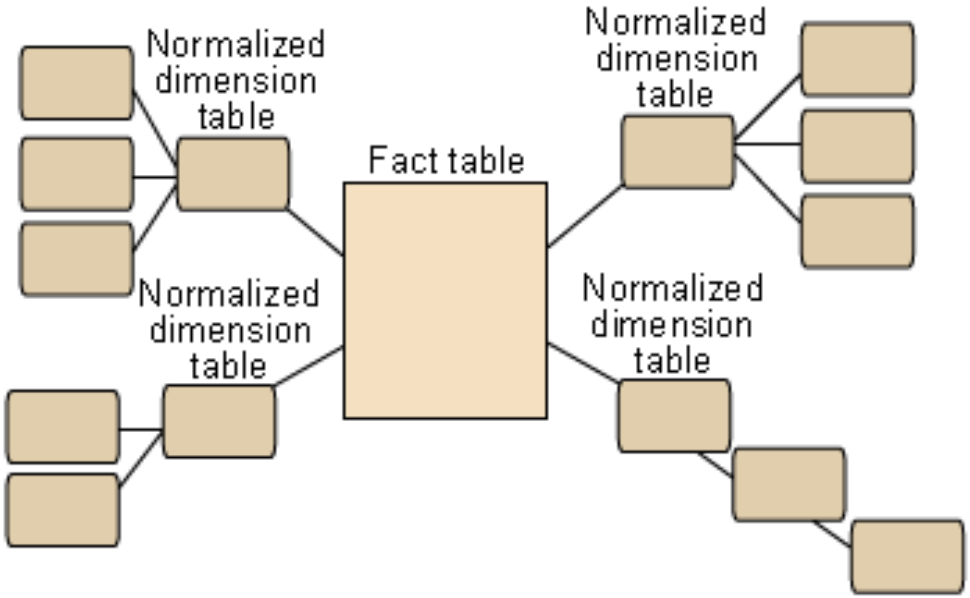
- Esquemes en estrella (star esquema)
- Esquemes en floc de neu (Snowflake esquema)

Esquema en Estrella (Star)



En l'esquema en estrella la taula de fets és l'única taula de l'esquema que témúltiples joins que la connecten amb altres.
Les taules de dimensió es troben a més totalment denormalitzades,és a dir, tota la informació referent a una dimensió s'emmagatzema en la mateixa taula.

Esquema en Floc de neu (Snow Flake)



En l'esquema de floc de neu de normalitzen les taules de dimensions, resultant múltiples taules per a una sola dimensió. Hi ha esquemes Snowflake complets o Snowflake parcial (no totes les dimensions estan normalitzades)

Star vs SnowFlake

Un esquema en estrella (star) permet realitzar les consultes de forma més ràpida ja que la quantitat de joins que hem de fer és molt menor que en un esquema de floc de neu (Snowflake).

Un esquema en Snowflake requereix en general menys temps de càlcul i càrrega que un esquema en estrella. No obstant això les consultes solen trigar més ja que cal fer joins per navegar per les dimensions.

Hi ha certs escenaris que són molt difícils de representar amb un esquema en estrella, normalment quan una dimensió es divideix en subcategories amb atributs diferents.

La taula de fets és la que conté el major volum de dades en un datawarehouse, complicar amb molts camps de claus externes (De les diferents dimensions i nivells, en un estel) pot alentir també les consultes del nostre sistema.

Staging Area

Àrea que replica els orígens de dades amb l'objectiu de treballar de forma asíncrona amb l'entorn operacional. Permet afegir noves fonts i alguns camps extres necessaris. Permet descarregar-vos l'entorn productiu de les extraccions.

Delivery Area

Conté una implementació del datawarehouse per a la precàrrega de les dades. S'utilitza per fer validacions abans de sobreescriure dades al datawarehouse productiu

Datawarehouse Area

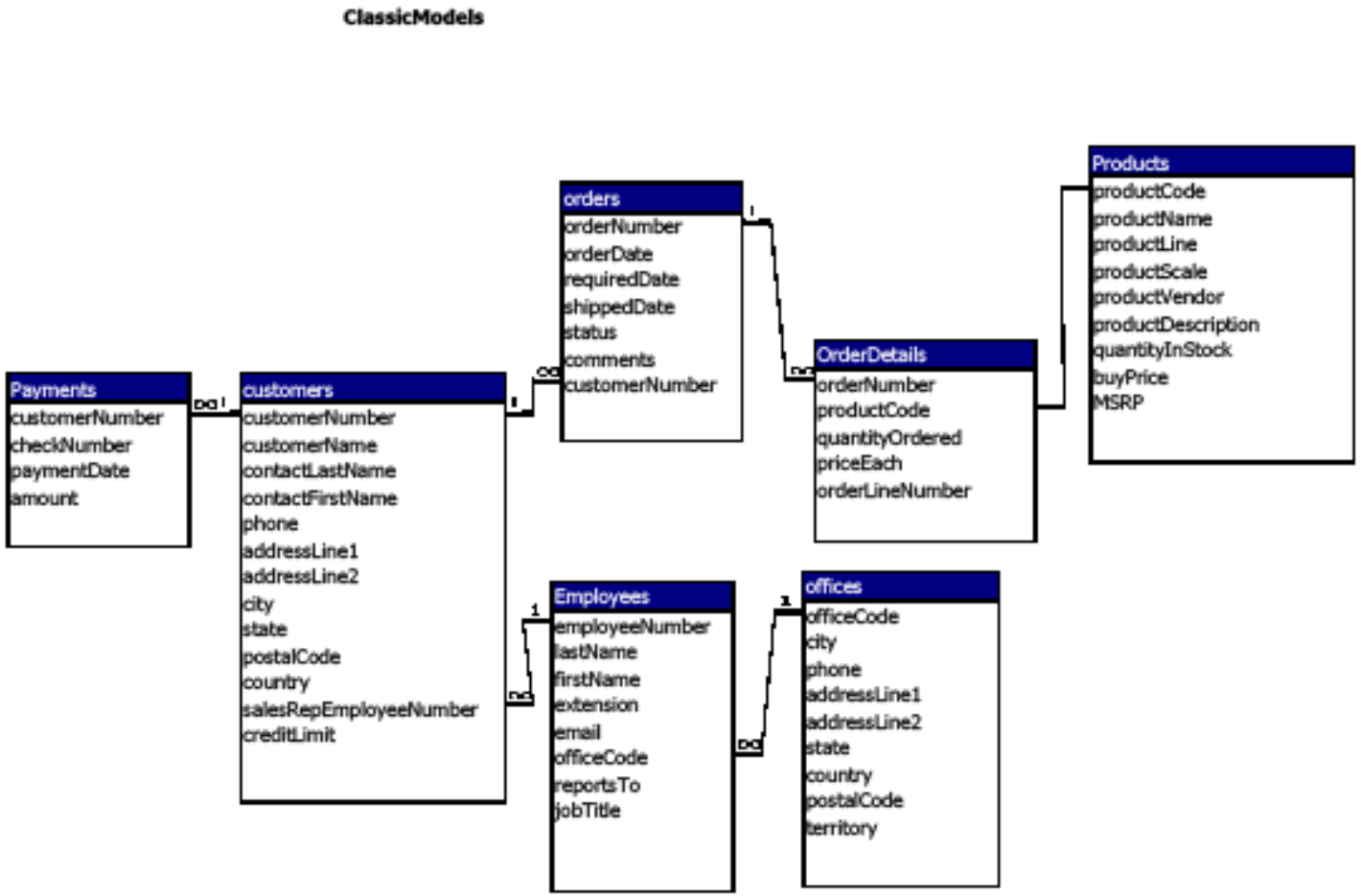
Àrea destinada a l'emmagatzematge no volàtil de les dades per explotar mitjançant els serveis destinats a aquest efecte.

Magatzems de dades

Disseny datamart exemple

Utilitzant la base de dades open source Classic Models. Construirem un datamart de vendes a partir de la informació de les comandes.

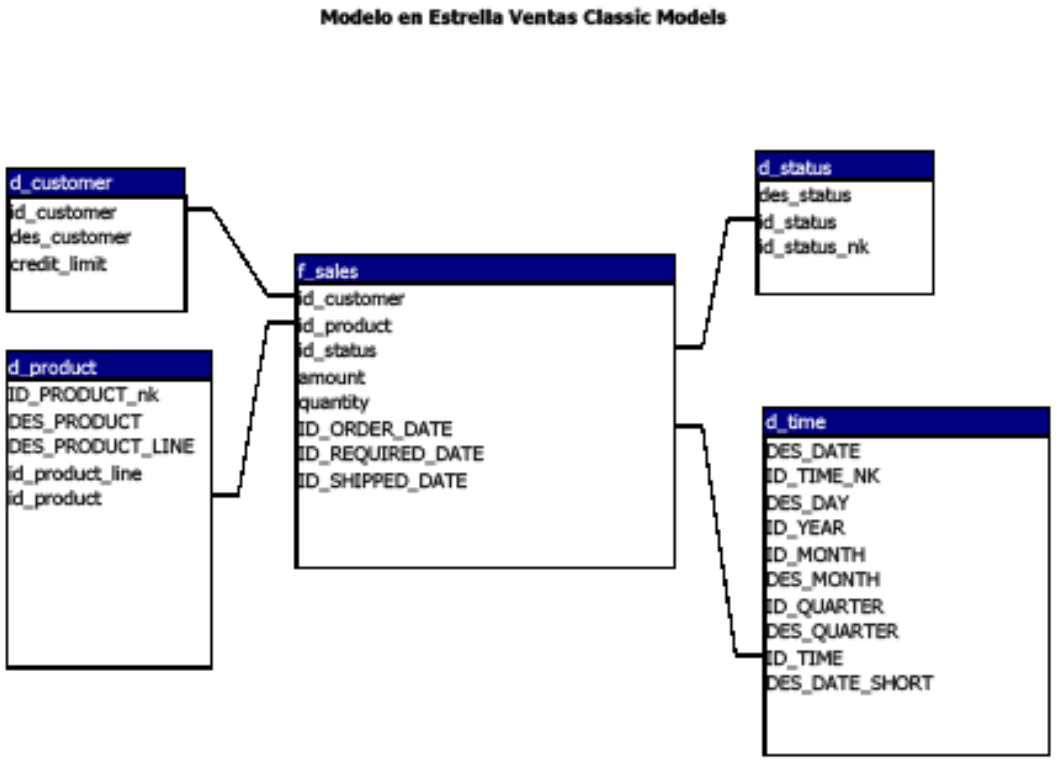
Diagrama Taules classic Models



Discusió del modelat

Magatzems de dades

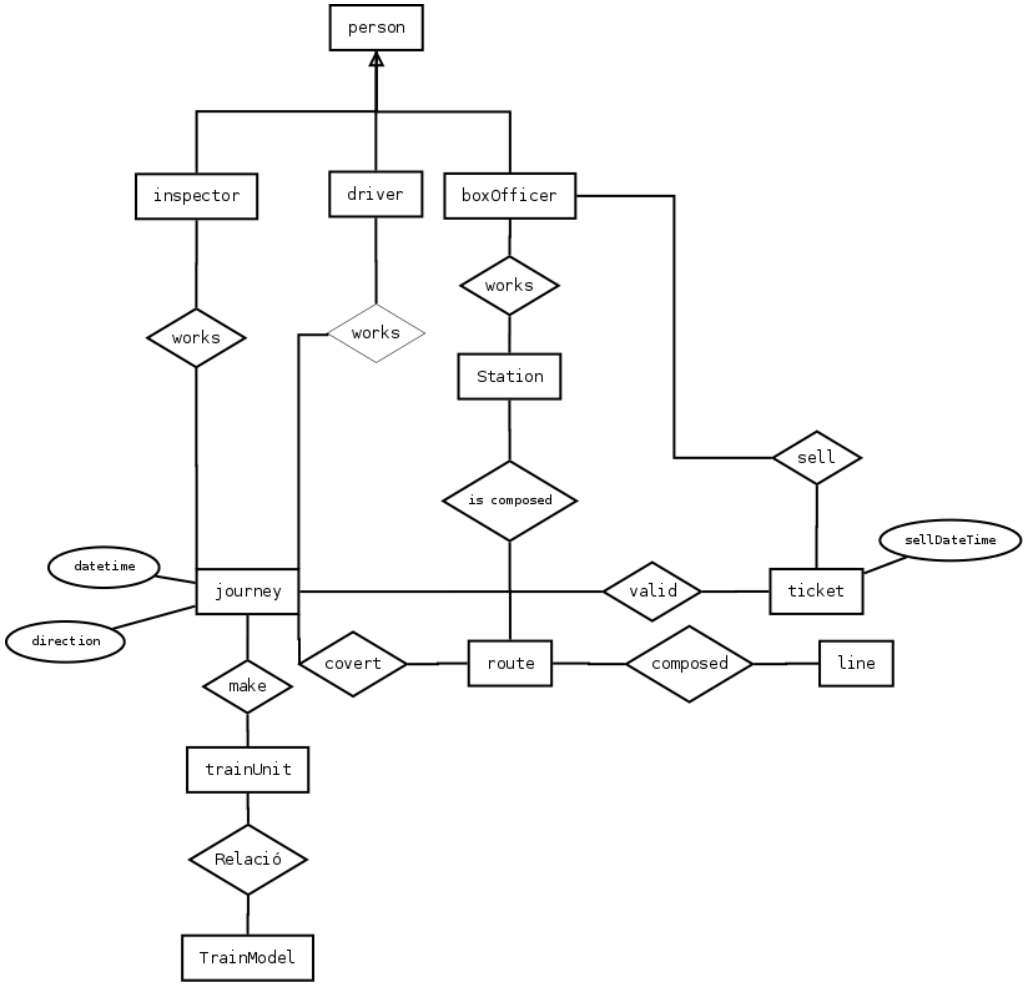
Disseny datamart exemple



Magatzems de dades

Disseny datamart exemple

Un exemple més complet



Magatzems de dades

Rendiment del datawarehouse

El volum de dades en un datawarehouse és molt elevat. La qüestió del rendiment és crítica. Hi ha diferents estratègies per augmentar el rendiment del datawarehouse:

- Taules Agregades
- Índexs
- Escollir el disseny adequat
- Jugar amb la caché de dades (MOLAP/ROLAP/HOLAP)