# Introducció a les bases de dades analítiques

PID_00270638

Alberto Abelló Gamazo
Josep Curto Díaz
Àngels Rius Gavidia
Montse Serra Vizern
José Samos Jiménez
Juan Vidal Gil
David Díaz Arias

Temps mínim de dedicació recomanat: 4 hores

Universitat Oberta de Catalunya

**Alberto Abelló Gamazo**

Doctor i enginyer en Informàtica per la Universitat Politècnica de Catalunya. Professor associat al Departament de Llenguatges i Sistemes Informàtics d'aquesta universitat. Coordina el programa de doctorat Erasmus Mundus IT4BI-DC a la UPC. Els seus interessos de recerca se centren en l'àrea de bases de dades, *Business Intelligence*, gestió de *Big Data*, fluxos de dades i gestió de metadades.

**Josep Curto Díaz**

Llicenciat en Matemàtiques per la Universitat Autònoma de Barcelona, màster en *Business Intelligence* i Direcció i Gestió de les Tecnologies de la Informació per la Universitat Oberta de Catalunya, i MBA per l'Institut d'Empresa Business School. Treballa en els àmbits de *Business Intelligence*, *Business Analytics* i *Big Data*. Des de l'any 2014 a Delfos Research, empresa de la qual és fundador, compagina aquesta activitat amb col·laboracions docents a IE Business School, UOC, EOI, U-TAD, IEB i Kschool.

**Àngels Rius Gavidia**

Enginyera en Informàtica per la UPC i doctora en Societat de la Informació i Coneixement per la UOC. Actualment professora dels Estudis d'Informàtica, Multimèdia i Telecomunicació de la UOC. Els seus interessos d'investigació se centren en la representació formal dels processos i serveis en entorns d'aprenentatge i la seva automatització.

**Montse Serra Vizern**

Enginyera en Informàtica i doctora enginyera en Informàtica per la UAB. Professora dels Estudis d'Informàtica, Multimèdia i Telecomunicació de la UOC i professora associada del Departament d'Arquitectures i Sistemes Operatius de la UAB. La seva investigació se centra en qüestions com: ètica professional i responsabilitat social de les TIC; gènere i equitat dins de les TIC; metodologies i eines d'aprenentatge a distància.

**José Samos Jiménez**

Doctor en Informàtica per la Universitat Politècnica de Catalunya. Professor titular del Departament de Llenguatges i Sistemes Informàtics de la Universitat de Granada, assignat a l'Escola Tècnica Superior d'Enginyeria Informàtica.

**Juan Vidal Gil**

Llicenciat en Físiques per la Universidad Complutense de Madrid. Experiència en solucions tecnològiques de *Business Intelligence* i *Data Warehouse*, com a cap de projectes en importants companyies i com a formador especialitzat en empreses del sector. Professor col·laborador de la UOC.

**David Díaz Arias**

Enginyer en Informàtica per la UOC. Enginyer Tècnic en Informàtica de Gestió per la UAB. Responsable tècnic i analista de dades de l'àrea de *Business Intelligence* en una empresa de l'àmbit de la salut. Professor col·laborador de la Universitat Oberta de Catalunya.


La revisió d'aquest recurs d'aprenentatge UOC ha estat coordinada per la professora: Àngels Rius Gavidia

Primera edició: febrer 2020

Av. Tibidabo, 39-43, 08035 Barcelona
Realització editorial: FUOC

# Índex

# Introducció

Generalment, l'estudi de les bases de dades s'inicia amb les bases de dades relacionals, que són les que, de manera majoritària, estan implantades des de fa unes dècades en la indústria. Aquest tipus de bases de dades permet emmagatzemar les dades i processar la informació generada amb l'operativa diària de l'organització. Per això es diu que les bases de dades ofereixen suport a l'activitat de negoci dins de les organitzacions. Així doncs, estan dissenyades per realitzar operacions de consulta i actualització de manera eficient per part de diferents usuaris. Alguns exemples d'operacions amb aquestes bases de dades poden ser la introducció de dades per emetre una factura, omplir un historial mèdic, gestionar una assegurança de vida, etc.

Aquesta assignatura presenta un altre tipus de bases de dades diferent als tradicionals, els que estan orientats a oferir suport a la presa de decisions en l'organització. Es tracta dels denominats magatzems de dades, coneguts també com a *Data Warehouse* (en aquest mòdul i en aquesta assignatura parlarem de tots dos indistintament).

L'objectiu principal del magatzem de dades o *Data Warehouse* és extreure rendiment de la informació emmagatzemada, i això vol dir extreure les dades per a una anàlisi posterior que ajudi a prendre decisions. Per tant, veiem que aquest tipus de bases de dades té un enfocament diferent respecte de les bases de dades convencionals.

Al llarg d'aquest mòdul, exposarem en què es basen els magatzems de dades i ho farem contraposant-los a les bases de dades operacionals perquè es vegin més clarament les diferències entre els dos tipus de bases de dades. Veurem com formen part d'un context més ampli, la factoria de la informació corporativa, de la qual són una peça clau.

Finalment, veurem quines són les tendències actuals dels magatzems de dades. No obstant això, no abordarem el tema de les bases de dades NoSQL (*Not only SQL*), que es tractaran en una altra assignatura. El terme bases de dades NoSQL s'utilitza per referir-se a una àmplia classe de sistemes de gestió de bases de dades que difereix del model clàssic del sistema de gestió de bases de dades relacionals en aspectes importants, el més destacat és que no utilitzen SQL com el principal llenguatge de consultes.

# Objectius

Els continguts inclosos en aquest mòdul s'orienten a aconseguir que l'estudiant arribi als objectius següents:

**1.** Conèixer l'orientació i els fonaments del magatzem de dades.

**2.** Conèixer quina ha estat l'evolució dels magatzems de dades i les seves característiques.

**3.** Saber distingir entre bases de dades operacionals i els magatzems de dades a diferents nivells.

**4.** Comprendre la importància de les dades i els processos en la presa de decisions, així com el paper que exerceix el magatzem de dades en la presa de decisions d'una organització dins d'un context més ampli i com a part del seu sistema d'informació.

**5.** Conèixer els elements principals que integren el context del magatzem de dades i la seva finalitat, així com la importància d'una explotació correcta de les dades emmagatzemades en ell.

**6.** Identificar els principals perfils professionals que apareixen vinculats al desenvolupament i l'administració dels sistemes BI, així com les seves responsabilitats principals.

**7.** Conèixer les tendències actuals sobre el magatzem de dades.

# 1. Què és un *Data Warehouse*. Característiques

El terme *Data Warehouse* o magatzem de dades ha estat concebut per Bill Inmon i R. D. Hackathorn.

La definició proporcionada per Bill Inmon és la següent: el **magatzem de dades** és una col·lecció de dades orientades al tema, integrades, no volàtils i historiades, organitzades per oferir suport als processos d'ajuda a la decisió.

D'aquesta definició, es desprèn el fet que es tracta d'un tipus de bases de dades, la importància de les quals rau en el suport que pot oferir a les organitzacions des d'un punt de vista estratègic i que, a primera vista, no sembla gaire difícil de construir. No obstant això, la dificultat principal en el moment de crear un magatzem de dades està a saber, *a priori,* quines dades es necessiten i de quina manera s'han d'organitzar.

Quantes empreses que volen dur a terme projectes d'aquest tipus coneixen exactament les dades que necessiten al magatzem de dades? L'experiència ens indica que pot haver-hi desconeixement pel que fa a les dades empresarials realment necessàries. Algunes d'aquestes empreses no saben que no tenen dades prou precises per introduir al magatzem de manera que després es puguin extreure resultats que serveixin per a la presa de decisions.

## 1.1. Evolució històrica

Abans de tractar les característiques d'un magatzem de dades, pot ser interessant veure com han evolucionat els sistemes d'informació pel que fa a l'emmagatzematge i explotació d'informació per a la seva anàlisi.

Breument, podem resumir l'evolució de la següent manera:

- **Dècada de 1960**. *Reporting manual*: la informació era difícil de trobar i analitzar. D'altra banda, els informes generats no presentaven cap flexibilitat a l'usuari, davant de cada nou requeriment era necessari reprogramar els informes.

- **Dècada de 1970**. Aparició dels sistemes de suport a la decisió i els sistemes d'informació executiva. Es tractava d'informació molt orientada a la direcció que intentava oferir suport a la presa de decisions.

La informació, tot i que es consolida per als informes, es trobava molt dispersa i cada nou requeriment implicava reprogramació.

- **Dècada de 1980**. Eines *desktop* d'anàlisis de dades. Es tracta d'aplicacions d'escriptori que utilitzaven una interfície d'usuari més amigable que les anteriors, però que a causa del creixement dels sistemes operacionals tenien informació d'origen molt dispers, difícil de trobar, creant sitges d'informació.

- **Dècada de 1990**. Creació dels primers *Data Warehouse* per centralitzar la informació provinent dels sistemes operacionals i per facilitar els processos d'anàlisis. Inmon va publicar el llibre *Building the Data Warehouse* el 1992, any en què es va començar a fer extensiu l'ús del terme.

- **Any 2000**. Emergència i desenvolupament de les plataformes d'Intel·ligència de Negoci entorn al *Data Warehouse*. Ampliació dels sistemes de planificació empresarial (ERP) amb un mòdul d'intel·ligència de negoci.

- **Anys 2003-2007**. Emergència i desenvolupament de les plataformes conegudes com a *Data Warehouse Appliances* (DWA) que optimitzen maquinari i programari per al treball analític. Els fabricants líders llancen al mercat diferents tipus de DWA basats en arquitectures de computació paral·lela de tipus MPP (*massively parallel processing*).

- **Anys 2008-2015**. Emergència i desenvolupament de tecnologies *Big Data*. Aparició i desenvolupament del *framework Hadoop*. Convivència entre aquestes tecnologies i els magatzems de dades tradicionals.

En l'actualitat, la intel·ligència de negoci i la gestió de la informació són activitats prioritàries en els departaments de Tecnologies de la Informació (TI) de les companyies.

### 1.2. Característiques d'un *Data Warehouse*

Com s'ha vist anteriorment, el magatzem de dades representa un canvi en el tractament de la informació. Per dur a terme un tractament adequat de la informació, el magatzem de dades ha de complir un conjunt de característiques: que estigui orientat al tema, que les dades estiguin integrades i que la informació sigui històrica i no volàtil.

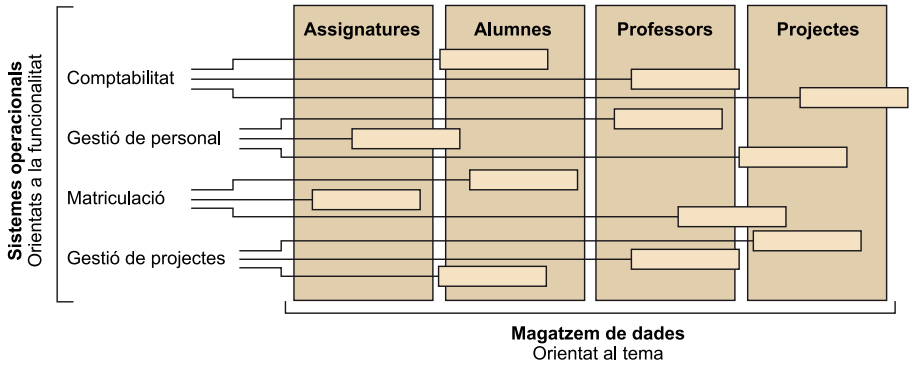
### 1.2.1. Orientat al tema

Aquesta primera característica fa referència a les directrius dels dissenyadors dels magatzems de dades. El disseny dels sistemes operacionals és causat per un conjunt de requeriments, ja que es construeixen per satisfer una necessitat concreta i molt coneguda. D'aquesta manera, parlem d'orientació a la funcionalitat.

Per contra, quan dissenyem un magatzem de dades, no sabem quines seran les necessitats dels analistes. No podem saber quins són els requisits concrets que tenen, ni l'ús que es pot arribar a fer de les dades emmagatzemades (això es decidirà molt temps després, quan sorgeixi la necessitat de fer un estudi concret). Per consegüent, l'única cosa que el dissenyador pot considerar en aquest cas són les àrees o els possibles temes d'anàlisi.

Atès que no podem conèixer els requeriments dels usuaris en el moment en què es construeix el magatzem de dades, la informació no s'estructura segons la seva funcionalitat (l'ús que se li donarà), sinó dividida per temes d'interès.

Figura 1. Orientació al tema d'un magatzem de dades




A la figura 1 es considera el cas d'una universitat, en la qual cada sistema operacional accedeix exactament a les dades que necessita i se suposa que de la manera més eficient possible. Per exemple, l'aplicació de comptabilitat accedirà a dades tant d'alumnes, com de professors o de projectes amb empreses. No obstant això, probablement no accedirà a totes perquè no en requereix algunes, com per exemple les notes dels estudiants. En canvi, un magatzem de dades guarda les dades segons els possibles temes que es poden analitzar. Hem de tenir en compte que *a priori* no sabem quina utilitat concreta es donarà a les dades emmagatzemades. Simplement es guarden per a quan sigui necessari analitzar-les. A més, tampoc no es guardaran totes les dades dels sistemes operacionals ja que algunes no pertanyen a cap tema d'anàlisi que resulti d'interès, com per exemple podrien ser els números de telèfon dels estudiants.

### 1.2.2. Integració de dades

Sabem que els sistemes operacionals de les empreses són heterogenis: funcionen sobre maquinari i programari diferent, utilitzen models de dades diferents (unes orientades a l'objecte, d'altres de relacionals, etc.) i presenten el negoci des de diferents punts de vista (finances, vendes, gestió de personal, etc.). Per tant, el primer pas per oferir totes les dades als analistes ha de ser la integració de tots aquests sistemes, de manera que els analistes, encara que les dades provinguin de fonts diferents, les vegin com si provinguessin d'una única font. El sistema ha de facilitar la resolució d'heterogeneïtats tant de semàntica com de sistema.

Hem de tenir present que no es tracta d'usuaris informàtics, sinó d'usuaris no experts als quals s'ha de facilitar el treball. A més, la integració també ajudarà a trobar contradiccions entre les fonts de dades diferents.

La integració de les dades presenta múltiples problemes, que no sempre són fàcils de resoldre. Per esmentar-ne només alguns, podríem parlar d'unificar els tipus i les estructures de dades, definir claus primàries comunes, unificar nivells de granularitat, trobar una convenció en la terminologia i definicions o definir un esquema de dades comú (capaç de representar la informació de totes les fonts alhora), garantir la qualitat de les dades integrades i realitzar una gestió àgil de fonts amb grans volums de dades.

A més, és necessari esmentar que els magatzems de dades disposen d'un component que ajuda a integrar: les metadades, de les quals parlarem més endavant. Encara que cal tenir present que les metadades permeten simplificar i automatitzar l'obtenció de la informació des dels sistemes operacionals fins als sistemes informacionals i, per tant, són bàsics per al procés d'integració.

### 1.2.3. Informació històrica i no volàtil

Les dues últimes característiques dels magatzems de dades fan referència al temps. Com ja hem comentat abans, les dades temporals són especialment importants en tasques d'anàlisi.

> Cal distingir dos tipus d'informació temporal. El primer tipus ens indica quan es produeix un esdeveniment en el món real (la historicitat). El segon quan tenim constància del fet a la nostra base de dades (la no volatilitat).

La historicitat és important per analitzar com han evolucionat les coses, per veure una pel·lícula en lloc d'una fotografia. Qualsevol dada en el magatzem de dades ha d'anar acompanyada del seu període de validesa. En canvi, la no volatilitat ens mostra quan ens hem assabentat dels fets i ens serveix per saber

si un informe es va fer tenint en compte unes dades o uns altres. La no volatilitat implica que no existeixin les operacions de modificar i esborrar pròpiament dites. Les dades no s'esborren o es modifiquen, sinó que s'insereixen correccions i la data en la qual s'han registrat.

**Exemple d'historificació**

Un cas d'historificació pot ser l'entitat que guarda la informació relativa a les tarifes mòbil d'un operador de telecomunicacions. Les característiques d'aquestes tarifes varien en el temps. Per exemple, els minuts de trucades o els *megabytes* de navegació d'una tarifa poden patir variacions d'acord amb l'estratègia comercial de la companyia. En una base de dades operacional, ens interessa tenir l'última foto de cada tarifa, mentre que en un magatzem de dades guardarem un històric de canvis que ens permetran realitzar estudis al llarg del temps.

Figura 2. Exemple d'historicitat de magatzem de dades

Base de dades operacional

| tarifa | minuts | navegació |
|---|---|---|
| 100 | 200 | 1000 |

Magatzem de dades

| id | tarifa | minuts | navegació | data inici | data final |
|---|---|---|---|---|---|
| 10001 | 100 | 150 | 500 | 01/07/2014 | 31/01/2016 |
| 10002 | 100 | 200 | 1000 | 31/01/2016 | – |

La historicitat ens servirà per fer estudis sobre l'evolució del negoci, mentre que la no volatilitat garanteix que no perdem cap dada (ni tan sols les errònies).

# 2. Objectius d'un *Data Warehouse*

En aquest apartat enumerarem els objectius principals que hauria d'aconseguir o complir un magatzem de dades, tant en l'àmbit empresarial com en el tècnic: ser un repositori central i integrat d'informació empresarial i ser un repositori base per a processos d'anàlisi i *reporting*.

## 2.1. Repositori central i integrat d'informació empresarial

Com s'ha comentat en apartats anteriors, els sistemes operacionals contenen informació valuosa per al negoci, però dispersa en diferents sistemes i bases de dades. El *Data Warehouse* té com a un dels seus principals objectius ser un repositori central d'informació corporativa que pot provenir de diferents sistemes. Aquest repositori té diverses funcions:

- Integrar informació provinent dels diferents sistemes de la companyia.
- Consolidar i homogeneïtzar aquesta informació.
- Ser el punt central d'informació. Versió més fidel de la informació, evitant tenir diferents versions segons la font que es consulti.
- Depurar i netejar les dades, garantint-ne la qualitat.
- Facilitar processos de fusió empresarial, si s'utilitza amb aquesta finalitat.

## 2.2. Repositori base per a processos d'anàlisis i *reporting*

Els processos de *reporting* i anàlisi de la companyia necessiten nodrir-se d'una informació de base. Aquesta informació de negoci es recull de les bases de dades operacionals, però accedir-hi per realitzar un procés d'anàlisi o *reporting* que requereix informació de negoci diversa i recollida en diferents sistemes, pot ser un procés costós i complex a causa de les diferents ubicacions de les dades i la seva heterogeneïtat. Per això resulta més productiu accedir a un repositori centralitzat com és el *Data Warehouse*.

Partint de la informació integrada, consolidada i depurada del *Data Warehouse*, podem realitzar processos d'anàlisi i *reporting* de diferent naturalesa:

- Processos de *reporting* periòdic recurrent.
- Quadres de comandament.
- Processos de *reporting ad hoc* per a necessitats d'informació concretes.
- Processos d'analítica avançada (predicció d'esdeveniments de negoci, *forecasting* d'evolució temporal).

Aquests processos ens informaran de la situació i evolució de la companyia des de múltiples perspectives ajudant-nos a comprendre què està passant i per què. Així mateix, ens ajuden a intentar predir l'evolució en un futur. Aquestes anàlisis que donen suport a la presa de decisions serien menys àgils sense l'existència d'un *Data Warehouse*.

**Exemple de procés d'anàlisi partint del *Data Warehouse***

Si hem de fer una anàlisi dels nostres clients en el qual estimem la probabilitat d'abandonament de la companyia segons les diferents variables que caracteritzen el client, serà més senzill i òptim llançar aquests processos de propensió d'abandonament partint de la base de dades de clients del *Data Warehouse*, on tenim tota la informació dels nostres clients depurada i integrada, que haver d'accedir a cadascun dels sistemes que recullen informació dels nostres clients: CRM, ERP, facturador, vendes, etc.

# 3. Comparativa entre *Data Warehouse* i bases de dades operacionals

Una manera d'iniciar la comparativa entre els magatzems de dades i les bases de dades operacionals serà a partir dels exemples següents.

**Exemple 1**

Imaginem la base de dades que pot utilitzar un treballador de banca d'una sucursal quan treballa en l'atenció al públic per finestreta. És cert que el volum de dades global de la base de dades pot ser molt alt, però les dades que es manipulen en cadascuna de les transaccions són molt simples: l'operació d'un ingrés o d'un reintegrament en la base de dades probablement només involucra la inserció en una determinada taula d'una tupla que ho reflecteixi.

Per tant, en cadascuna de les operacions (de manera general) s'involucren molt poques dades, però és cert que el volum global resulta enorme i, atès que s'acumulen diàriament, tendeix a créixer molt ràpidament. A més, la disponibilitat de la base de dades ha de ser total: seria inacceptable que un client d'aquesta sucursal es veiés obligat a esperar quinze minuts a què el sistema gestor fes la transacció que reflecteixi un reintegrament per disposar de diners.

**Exemple 2**

Continuem amb la sucursal bancària. És evident que, si el director d'aquesta sucursal vol decidir si potenciar un determinat producte financer o no, i per això necessita analitzar l'evolució de l'índex de morositat del darrer any dels seus clients, no cal tenir en compte si un determinat client ha anat al matí a fer moviments al seu compte i si aquest fet ha variat la morositat (exceptuant casos significatius). Les necessitats del director són més globals: necessita conèixer l'evolució ascendent o descendent d'aquest índex sense entrar en detalls.

Com es pot comprovar, la funció que duu a terme cadascuna de les bases de dades en els exemples anteriors és molt diferent. En el primer cas es tracta d'una base de dades operacional i en el segon, d'un magatzem de dades.

Actualment, les bases de dades relacionals són operatives en un entorn molt concret que respon a les necessitats per a les quals es van crear. Aquestes necessitats solen involucrar entorns de gestió purs en els quals hi ha simplicitat de les estructures i dels tipus de dades, utilització de transaccions curtes, etc.

D'altra banda, les necessitats actuals d'informació de les organitzacions han variat. La disponibilitat de gran quantitat d'informació és de vital importància per als negocis, ja que les decisions de futur se solen prendre a partir d'aquesta informació.

**Continuem amb els exemples**

Està clar, per tant, que els fets que la base de dades operativa té no són els que el director necessita. De totes maneres, la globalització de les dades que busca el director es basa clarament en la informació reflectida en aquesta base de dades, però organitzada d'una altra manera (en aquest cas, resumida).

Aquest tipus de necessitats per reflectir tendències, evolucions, fets històrics en el negoci i possibilitats futures són factors que l'alta direcció de les institucions o empreses ha de manipular d'una manera habitual i que ha propiciat l'aparició al mercat d'eines d'ajuda en la presa de decisions.

## 3.1. Diferències en l'emmagatzematge, el disseny i l'estructura de les dades

**1) Temporalitat**

Les dades s'han de guardar el temps que sigui necessari. En les bases de dades operacionals aquest temps normalment oscil·la entre un i dos anys, i en el magatzem de dades s'amplia de cinc a deu anys. Més enllà d'aquests intervals de temps, les dades es deixen de considerar útils.

**2) Volum**

Evidentment, la característica de la temporalitat ens condiciona el volum. No és el mateix guardar les dades un any que deu. Per tant, en les bases de dades operacionals, el volum serà relativament petit i, en el magatzem de dades, serà molt més gran.

**3) Nivell d'agregació**

El nivell d'agregació permet el cúmul de les dades. En un nivell 0, tindríem totes les dades de manera detallada. Aquest nivell d'agregació en les bases de dades operacionals sol ser únic i bastant baix. En canvi, en el magatzem de dades solen haver-hi diferents nivells. Aquest fet ens indica que algunes vegades tenim les dades duplicades de manera implícita.

**4) Actualització**

L'actualització de les dades en una base de dades operacional es fa constantment; per tant, la informació és molt canviant. Per contra, en el magatzem de dades es fa d'una manera periòdica i en intervals de temps definits. En la base de dades operacional les actualitzacions acostumen a ser atòmiques (registre a registre) i en el magatzem de dades per lots (conjunts de registres).

**5) Estructura**

El fet que les bases de dades operacionals i els magatzems de dades tinguin objectius diferents implica que necessitaran una estructuració de les dades diferent per aconseguir els objectius que tenen assignats.

En el cas de les bases de dades operacionals, tindran una estructura relacional, en la qual es dona molta importància a l'estabilitat. Aquest fet representa tenir bases de dades estàtiques, que no canvien amb freqüència la seva estructura.

En canvi, en els magatzems de dades hi haurà una visió multidimensional i alhora seran molt dinàmics: aquests s'han d'adaptar ràpidament a les necessitats del negoci per ser útils en els processos de presa de decisions.

En el disseny del magatzem de dades, cal tenir present el component temps, mentre que en les bases de dades operacionals no és necessari.

En el disseny de les bases de dades operacionals, cal que sigui més important que l'accés sigui immediat a una dada en concret, mentre que en els magatzems de dades solen predominar les consultes massives de dades.

Una altra diferència important és el fet que el disseny de les bases de dades convencionals ha de ser normalitzat, mentre que en els magatzems de dades és millor la desnormalització, ja que afavoreix la rapidesa de les consultes.

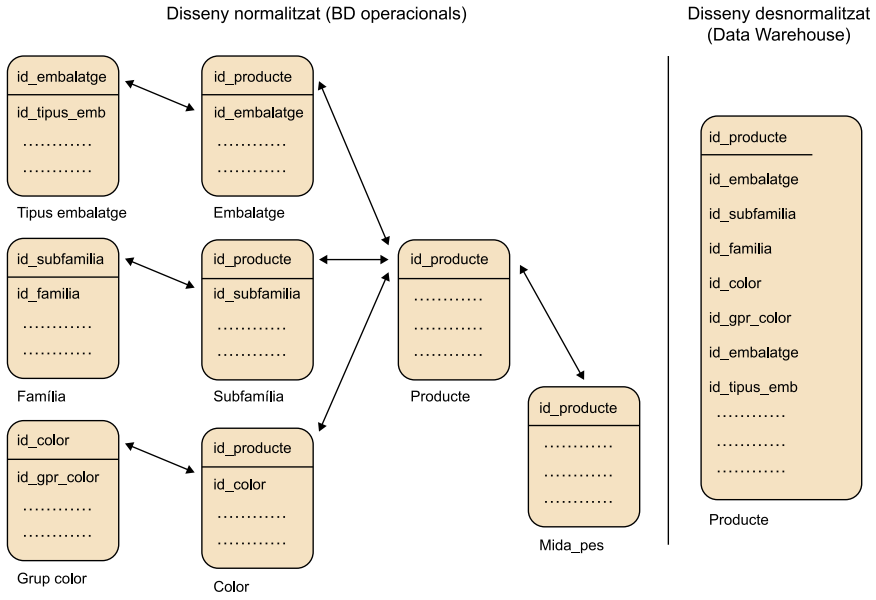
Pel que fa a la integritat de la informació veiem que les bases de dades operacionals normalment garanteixen la integritat definint restriccions en la base de dades (claus primàries i foranes), mentre que en els magatzems de dades, ens trobem dissenys en els quals la integritat es garanteix en el procés de càrrega (actualitzacions massives) i no es defineixen restriccions en la base de dades de destinació per millorar el rendiment de l'actualització.

**Exemple de diferència d'estructuració**

Una empresa que comercialitza un determinat grup de productes tindrà una base de dades operacional amb la informació dels seus productes. Entre la informació que guardem dels productes podem tenir característiques com la mida, el pes, la família, la subfamília, l'embalatge, el tipus d'embalatge, el color i el grup de color. Són característiques de naturalesa diferent i existeixen agrupacions entre elles (família-subfamília, embalatge-tipus embalatge, color-grup color). Hi ha la possibilitat de crear un disseny normalitzat amb una entitat de dades producte que tingui un identificador únic que permeti relacionar-lo amb altres entitats com mida-pes, subfamília, color i embalatge. Al seu torn existirà una relació entre color i grup de color, subfamília i família i embalatge i grup d'embalatge tal com mostra la figura 3. Aquest disseny respon a les necessitats d'un sistema operacional orientat a transaccions i pretén evitar redundàncies d'espai.

En canvi, si treballem en un disseny de base de dades per a un *Data Warehouse* el nostre disseny estarà orientat a les consultes, en molts casos massives. Per exemple, si en una consulta analitzem un indicador segons els productes i els seus atributs, és possible que necessitem realitzar moltes combinacions de taules en la consulta si s'ha dissenyat la base de dades d'acord amb el model de dades operacional. Això pot complicar i alentir la consulta. Per tant, si realitzem un disseny orientat a consultes, pot interessar-nos crear una única entitat producte amb tots els atributs que el descriuen, ja que això pot simplificar les consultes a una única entitat i millorar els seus temps de resposta. Cal considerar que aquest segon disseny és més adequat per a les consultes, però tenim redundància en les dades, especialment en els atributs d'agrupació (família, tipus d'embalatge i grup de color).

Figura 3. Disseny normalitzat enfront de disseny desnormalitzat




## 3.2. Diferències en el tractament de la informació

### 1) Explotació de la informació

En l'entorn de les bases de dades operacionals, amb freqüència els usuaris finals accedeixen a les dades mitjançant aplicacions predefinides.

En els magatzems de dades, les consultes acostumen a ser imprevistes. Pot haver-n'hi de predefinides, però la varietat de possibilitats que trobem fa impossible preveure quines seran les necessitats dels usuaris finals. A més, aquestes consultes estan orientades a àrees d'interès del negoci que amb freqüència són canviants. Dins d'aquesta varietat de possibilitats sí que és possible identificar entitats, agregacions o encreuaments d'ús freqüent d'acord amb els quals podem definir vistes o taules de bases de dades que continguin preagregats, índexs a les taules o un altre tipus d'estratègies d'optimització.

### 2) Temps de resposta

El temps de resposta de les operacions ha de ser instantani quan parlem de bases de dades operacionals, a causa de la freqüència amb la qual s'actualitzen les dades. Per contra, en el cas dels magatzems de dades, aquest temps ha de ser ràpid, però no necessàriament instantani. Les operacions en els magatzems de dades solen ser consultes massives que són impossibles d'obtenir de manera instantània, però sí que han d'estar en un temps raonable d'acord amb el treball de l'analista. Hi ha informes que realitzen un conjunt de consultes massives i que poden ser planificades en diferit perquè s'executin en *background* i puguin ser consultades posteriorment. El concepte d'execució de consulta en diferit és molt menys comú en les bases de dades operacionals.

### 3.3. Diferències en les funcionalitats

**1) Activitats**

L'activitat de les bases de dades operacionals es produeix diàriament amb les activitats del negoci, ja que s'utilitzen per a l'operativa o funcionament de l'empresa. Per tant, seran aplicacions fàcils de gestionar, en les que no s'haurà de pensar gaire en les opcions que ofereix i seran ràpides.

Al contrari, l'activitat dels magatzems de dades és d'anàlisi i decisió estratègica. Les aplicacions tindran unes funcionalitats diferents de les de l'entorn operacional, que es complementaran amb múltiples opcions i permetran moltes opcions de lliure aplicació.

**2) Importància de les dades**

Com ja hem dit anteriorment, la dada és molt important en els dos entorns. En el cas de la base de dades operacional, el què és important és la dada actual, mentre que en el cas del magatzem de dades la importància rau en les dades històriques.

**3) Usuaris**

En les bases de dades operacionals, sol haver-hi molts usuaris. Aquest fet es complementa amb el nivell d'usuari, ja que no tothom pot fer de tot. Els usuaris solen ser de l'estructura mitjana-baixa de l'empresa.

En l'entorn del magatzem de dades hi ha menys usuaris i solen definir-se diferents perfils segons la informació que es consultarà. Tradicionalment ha existit un usuari de perfil directiu (direcció, màrqueting, planificació estratègica, control de gestió, etc.) que accedeix a dades agrupades i/o acumulades; també existeixen perfils tàctics que accedeixen a dades agregades, però amb una visió més centrada en el seu departament o línia de negoci i, finalment, un perfil operatiu que accedeix a informació més relacionada amb l'operativa diària.

# 4. La factoria d'informació corporativa

William Inmon va presentar el 1998 el que s'anomena factoria d'informació corporativa. Es tracta d'un concepte per fer referència a un conjunt de components que interactuen per ajudar a gestionar tots els fluxos de dades, des dels sistemes operacionals de l'empresa fins als analistes. El seu objectiu és transformar les dades dels sistemes operacionals (matèries primeres) en informació útil per als analistes (producte elaborat) amb la finalitat d'utilitzar-la en els processos de presa de decisions en l'organització. En aquest mòdul veurem els diferents components d'aquesta factoria i com interactuen entre si.

Aquests components són els següents:

- Magatzems de dades departamental, corporatiu i operacional.
- Component de transformació i integració.
- Gestió de dades mestres.
- Metadades.
- Estructures multidimensionals.

## 4.1. Magatzem de dades departamental

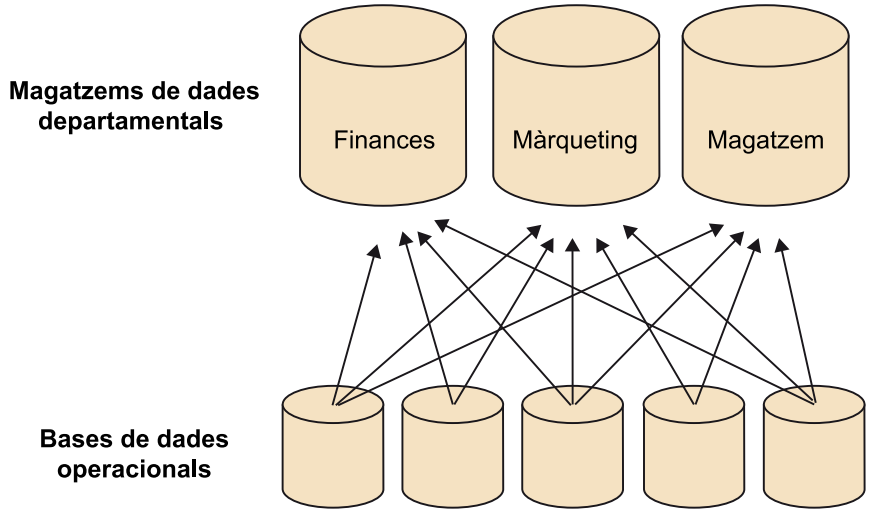
Construir un magatzem de dades és molt costós, a més de tenir uns requisits de rendiment difícils d'aconseguir. La solució per obtenir un temps de resposta baix és disposar de diferents magatzems només amb informació parcial del negoci (únicament la part que interessi a un departament o conjunt de persones).

Aquests magatzems de dades departamentals, també coneguts per la seva denominació en anglès *Data Mart*, normalment estaran dissenyats segons el model de dades multidimensional, la qual cosa facilita la millora en el rendiment mitjançant tècniques específiques d'emmagatzematge de les dades. A més, per no sobrecarregar els sistemes amb dades innecessàries, aquests només contenen dades històriques dins del període de temps que sigui estrictament necessari.

**Exemple de tècnica d'emmagatzematge per millorar el temps de resposta**

Una tècnica per millorar el temps de resposta és la preagregació. Aquesta tècnica consisteix a guardar els resultats de les funcions d'agregació (suma, mitjana, mínim, etc.) ja calculats per a quan l'usuari els demani. Això vol dir que hem de conèixer (o imaginar) quines consultes voldran fer-se per calcular prèviament els resultats, de manera que el càlcul no s'hagi de fer en el moment concret en què se sol·licita.

Figura 4. Magatzems de dades departamentals




Tal com es pot veure a la figura 4 per a cada departament o grup d'usuaris es construeix un magatzem. Aquest només conté les dades necessàries per satisfer les necessitats concretes del departament o grup d'usuaris i les integra amb independència de la font de dades de procedència. Aquestes dades es modelen seguint la visió de la realitat que tingui el departament corresponent i no fa falta que es consensuï amb tota l'empresa.

Un aspecte fonamental en el disseny dels magatzems de dades departamentals és la gestió de les entitats comunes entre magatzems. Per exemple, els departaments de Màrqueting i Finances utilitzaran entitats amb dades de clients o productes. És important que s'utilitzi la mateixa entitat de cara a la integritat de les dades entre magatzems de diferents departaments. Aquestes entitats comunes es denominen dimensions conformades en els models dimensionals i són entitats del tipus clients, productes, proveïdors, comptes que com que són crítics per al negoci s'utilitzen en molts departaments.

Un altre avantatge dels magatzems de dades departamentals és que no necessiten tenir les dades amb el màxim nivell de detall. Per exemple, si els analistes només volen veure les dades mensuals, no és necessari emmagatzemar les dades diàries. D'aquesta manera, no caldria emmagatzemar les vendes diàries de l'empresa, sinó només el total que s'ha venut durant un mes, la qual cosa representa un estalvi d'espai clar.

Tenir molts magatzems de dades petites permet abaratir costos, ja que són més econòmics que un de gran que satisfaci les necessitats de tothom alhora. A més, fent-ho així, facilitem la configurabilitat. Finalment, també és més fàcil controlar tant els costos (que s'imputaran al departament corresponent) com els accessos, processos i configuració del sistema (que correspondran a un conjunt d'usuaris molt restringit).

D'altra banda, un magatzem de dades departamental des d'un punt de gestió de projectes té un abast més limitat i definit que un magatzem de dades corporatiu i el nivell de risc és menor. Facilita un plantejament de projecte per fases. Des d'un punt de vista de negoci, cal assenyalar que no tots els departaments evolucionen al mateix ritme pel que fa a les necessitats analítiques,

hi ha departaments molt demandants d'analítica i dades com poden ser Màrqueting o Finances i d'altres la demanda dels quals pot ser menor com el Legal o el de Recursos Humans.

Els magatzems de dades departamentals guarden una història parcial de les dades que interessen a un departament. Estan dissenyats per obtenir un bon temps de resposta davant de les consultes d'un conjunt d'analistes.

### 4.2. Magatzem de dades corporatiu

Tenir múltiples magatzems de dades departamentals independents genera problemes a llarg termini, tot i que són més econòmics i fàcils de construir a curt termini. El primer problema és que, com podeu veure a la figura 4, tenim processos independents d'integració i transformació per a cada magatzem de dades departamental. A més, on guardem la informació que actualment no interessa a cap departament? No tenim cap lloc on la puguem guardar i no la podem menysprear. Cal tenir un magatzem de dades corporatiu que guardi tota la història de totes les dades i sempre amb el màxim nivell de detall possible. No obstant això, els magatzems de dades departamentals encara són necessaris.

Convé conèixer la terminologia anglosaxona generalment utilitzada per referir-se a cadascun dels magatzems.

- Magatzem de dades departamental: *Data Mart*.
- Magatzem de dades corporatiu: *Enterprise Data Warehouse*.

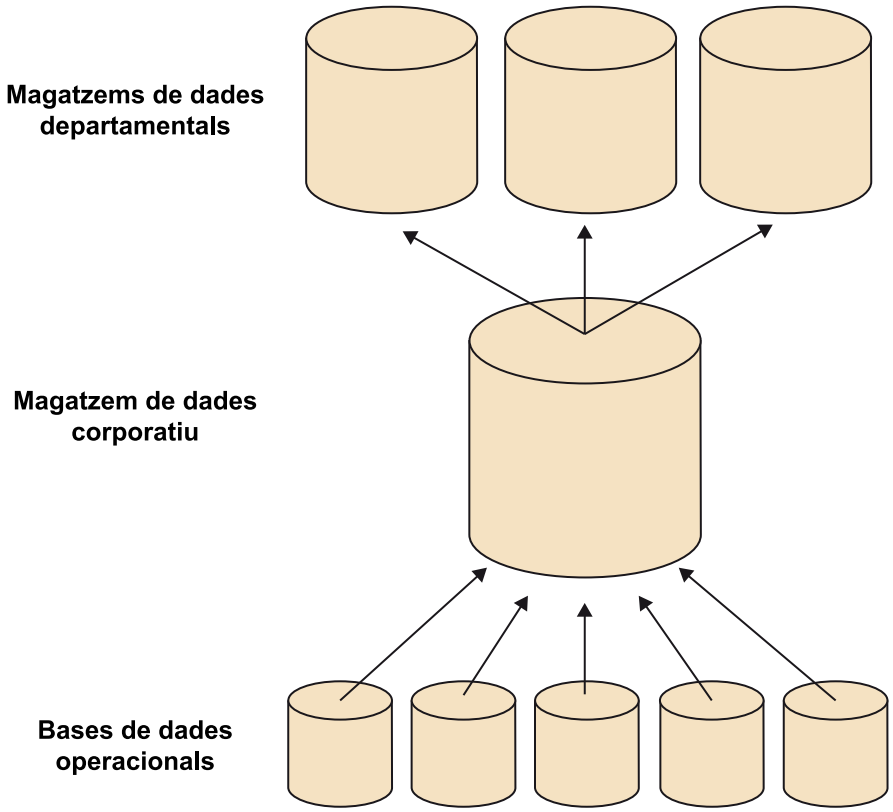
Normalment quan es parla de magatzem de dades en general, sense especificar el tipus, sol utilitzar-se el terme anglosaxó. Usualment per referir-nos a un magatzem de dades corporatiu utilitzem també *Data Warehouse*, encara que de vegades i per remarcar el seu caràcter corporatiu en el terme emprem *Enterprise Data Warehouse*. A vegades sentim a parlar de *Data Warehouse* de màrqueting per referir-nos a un magatzem de dades del departament de Màrqueting, però seria més apropiat denominar-lo *Data Mart* de Màrqueting.

El magatzem de dades corporatiu no és adequat per als usuaris finals, perquè està dissenyat per gestionar i integrar grans quantitats de dades que, juntament amb l'excés d'usuaris, degraden el temps de resposta. No es pot dissenyar per afavorir a un grup d'usuaris concret, sinó que ha de servir a tothom alhora de la millor manera possible.

D'aquesta manera, com es pot veure a la figura 5, el magatzem de dades corporatiu és el resultat d'un procés d'integració i transformació de totes les fonts de dades únic i complex, que estudiarem detalladament en l'apartat correspo-

nent d'aquest mòdul. Els magatzems de dades departamentals ara s'obtenen simplement com a resultat d'un procés de transformació a partir del magatzem corporatiu.

Figura 5. Magatzem de dades corporatiu




El magatzem de dades corporatiu guarda tota la història de totes les dades de l'empresa integrades. Està dissenyat per emmagatzemar-les de manera eficient.

La taula 1 resumeix les diferents característiques dels dos tipus de magatzem de dades que hem vist fins ara:

Taula 1

| **Característica** | **Magatzem de dades** | |
|---|---|---|
| | Departamental | Corporatiu |
| Temàtica | Específica | Genèrica |
| Fonts de dades | Poques | Moltes |
| Mida | Gigabytes | Terabytes |
| Temps de desenvolupament | Mesos | Anys |
| Model de dades | Multidimensional | Multidimensional, relacional i orientat a objectes |

En primer lloc, el magatzem de dades corporatiu ha de ser genèric i ha de guardar dades de tota l'empresa seguint una visió consensuada del negoci. Per contra, els magatzems de dades departamentals són absolutament específics. Només contenen les dades que demana un conjunt d'usuaris, les guarden segons la concepció que aquests tenen del negoci i estan optimitzats per obtenir un bon rendiment davant de les tasques d'anàlisi que es desitgen realitzar.

### 4.3. Magatzem de dades operacional

Desgraciadament, és possible que amb els magatzems de dades departamentals i el corporatiu encara no tinguem cobertes totes les necessitats d'informació de l'empresa. Atès el seu volum de dades i les tècniques d'implementació que s'utilitzen, el magatzem de dades corporatiu (i, per tant, els departamentals que s'actualitzen a partir d'aquest) no es pot tenir constantment actualitzat (només se sol actualitzar durant les nits o els caps de setmana). D'altra banda, els seus usuaris tampoc no ho requereixen, perquè estan més interessats en les dades històriques que en les actuals. No obstant això, pot haver-hi altres usuaris que també demanin dades integrades i que les vulguin completament actualitzades. Encara necessitem un altre tipus de repositori d'informació.

L'aparició d'aquest repositori, també conegut com a ODS (*Operational Data Store*), és causada per la típica ponderació entre volum de dades i velocitat del sistema. Fins ara, en els altres magatzems, el que volíem era tenir absolutament qualsevol dada que poguéssim arribar a necessitar per prendre una decisió. Com a conseqüència d'aquest requeriment, el temps de resposta pot arribar a degradar-se i, en qualsevol cas, ens veiem obligats a renunciar a tenir les dades constantment actualitzades. En aquest cas, valorem més el fet que les dades sempre estiguin actualitzades, que no que les tinguem totes. Per tant, renunciem a tenir dades històriques i disposem d'un repositori volàtil.

Aquest és el preu que s'ha de pagar per reduir el volum de dades i poder-les mantenir constantment actualitzades. D'aquesta manera, el magatzem de dades operacional i el corporatiu es complementen: el corporatiu guarda totes les dades històriques, però no està actualitzat sempre, i l'operacional sempre està actualitzat, però no conté dades històriques.

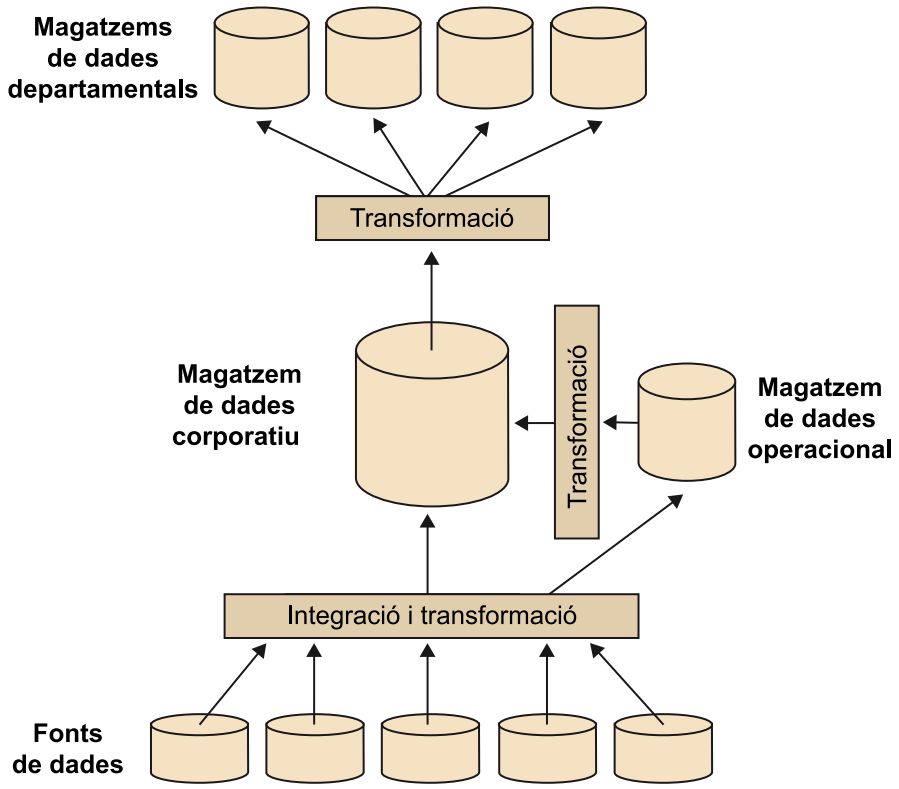
El magatzem de dades operacional és una estructura a cavall entre el món operacional i el de la presa de decisions. Està orientat al tema i integrat com qualsevol magatzem de dades, però en aquest cas no conté cap mena d'informació temporal.

### 4.4. El component d'integració i transformació

Com hem vist a l'apartat 3.2 «Diferències en el tractament de la informació», els sistemes operacionals dels quals disposen les organitzacions generalment no compleixen els requeriments dels analistes. Com a solució, s'ha definit el concepte de magatzem de dades, tant en l'àmbit departamental com en el corporatiu, segons les característiques de les seves dades, que el diferencien dels sistemes operacionals.

Així i tot, les dades dels magatzems de dades s'obtenen a partir dels sistemes operacionals de l'empresa, així com de fonts externes. Per les seves característiques diferents pel que fa a estructura i organització, les dades obtingudes de les fonts no es poden utilitzar directament en el magatzem de dades, sinó que s'han d'adaptar als seus requeriments en aquests aspectes. És per això que és necessari el component d'integració i transformació de dades, l'arquitectura lògica de les quals es pot veure a la figura 6.

Figura 6. Component d'integració i transformació




La missió del component d'integració i transformació consisteix a obtenir les dades per als diferents magatzems de dades de l'organització. Aquest component també es coneix com a ETL, per les seves sigles en anglès: *Extract, Transform and Load*.

Originalment, les dades s'obtenen a partir dels sistemes operacionals i altres fonts de dades, i s'han de transformar, depurar i integrar i, segons l'estructura dels esquemes dels magatzems de dades, també s'han de transportar i carregar perquè es puguin utilitzar en els diferents magatzems de dades de l'organització.

A diferència dels magatzems de dades, l'element principal de les quals és la base de dades, l'element principal del component d'integració i transformació és el programari encarregat de dur a terme la missió descrita.

Tant les fonts de dades com els diferents magatzems de dades es poden trobar en plataformes diferents, i, per tant, el component d'integració i transformació tindrà elements en les diferents plataformes en les quals estiguin la resta dels components de la FIC.

El component d'integració i transformació està format per programari que s'executa en les diferents plataformes en les quals funciona la resta dels components de la FIC.

### 4.5. Gestió de dades mestres

Generalment tots els magatzems de dades tenen una sèrie d'entitats crítiques pel que fa a la informació que contenen, com poden ser: clients, productes, proveïdors o comptes. Aquestes entitats intervenen en moltes consultes i la seva correcta actualització és fonamental per realitzar una anàlisi precisa. És comú que aquestes entitats estiguin compartides per diversos magatzems de dades departamentals i de vegades per sistemes no informacionals que accedeixen a aquestes entitats pel fet de ser el magatzem de dades corporatiu la seva imatge més fidel. Així mateix, aquestes entitats del magatzem de dades acaben essent entitats mestres que requereixen una gestió especial de cara a la realització d'activitats com poden ser: consolidar tota la informació rellevant de l'entitat que pot procedir de diferents sistemes, assegurar la qualitat d'aquesta informació, el seu refresc i la sincronització amb altres sistemes, entre altres activitats.

Aquestes activitats se solen denominar gestió de dades mestres o *Master Data Management* (MDM) i s'enmarquen dins de les activitats de govern de dades o *Data Governance* (DG).

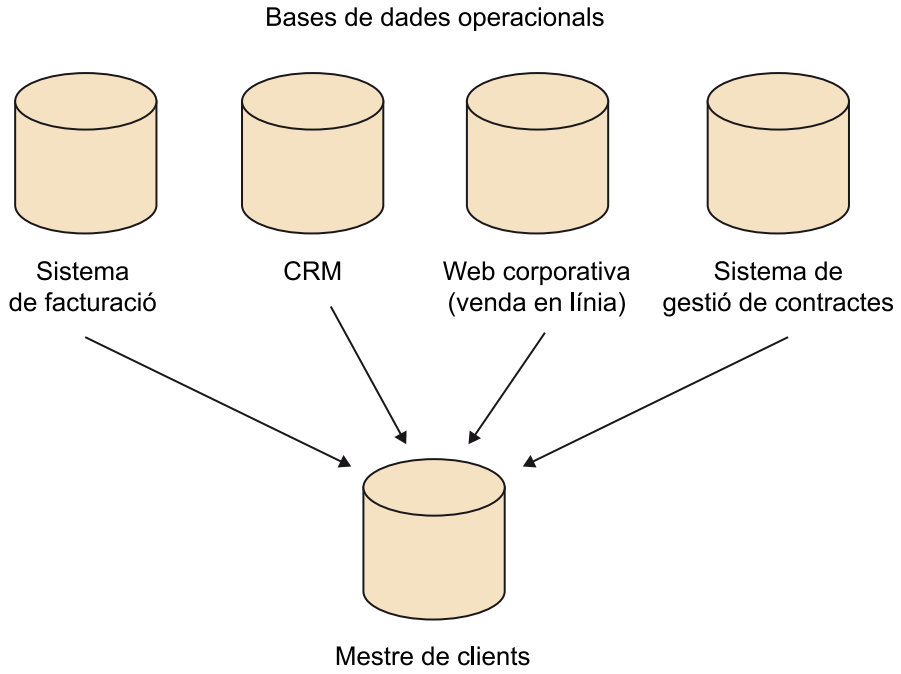
La gestió de dades mestres té com a objectius principals:

- Identificar les fonts d'origen de les dades mestres.
- Identificar als productors i consumidors de dades mestres.
- Recopilar i analitzar metadades sobre les dades mestres recopilades.
- Actualitzar i mantenir de manera centralitzada les dades mestres. Processos de consolidació i enriquiment. Generació de registres mestres.
- Determinar els responsables (administradors) de les dades mestres.

- Assegurar la qualitat de la dada d'aquests mestres.

A continuació, a la figura 7, s'il·lustra gràficament la creació d'un fitxer mestre de Clients a partir de diferents bases de dades operacionals. Hi veiem com les diferents entitats mestres s'integren dins dels magatzems de dades.

Figura 7. Mestre de clients generat des de BD operacionals




Directament relacionats amb les activitats de MDM hi ha els processos de seguiment de la qualitat de la dada que s'implementaran sobre les dades mestres i que permetran monitoritzar la seva qualitat, revisant aspectes com ara l'exactitud, integritat, consistència i completesa.

### 4.6. Les metadades

Les metadades no són un element específic de la FIC: apareixen en molts contextos del món del programari. La definició més freqüent que hi ha del concepte de metadada està basada en la seva etimologia: «Les metadades són dades sobre dades». Les dades generalment representen característiques de les entitats que modelen; en el cas de les metadades, representen característiques d'altres dades que en faciliten l'administració i ús. És a dir, el que diferència una dada d'una metadada, més que la seva estructura o contingut, és el seu propòsit i ús.

Les metadades descriuen les seves característiques (per exemple, format, origen, ús, etc.) sobre un conjunt de dades. Aquestes metadades són dades i al seu torn podem tenir altres metadades que descriguin les seves característiques (metadades sobre metadades), i així de manera successiva.

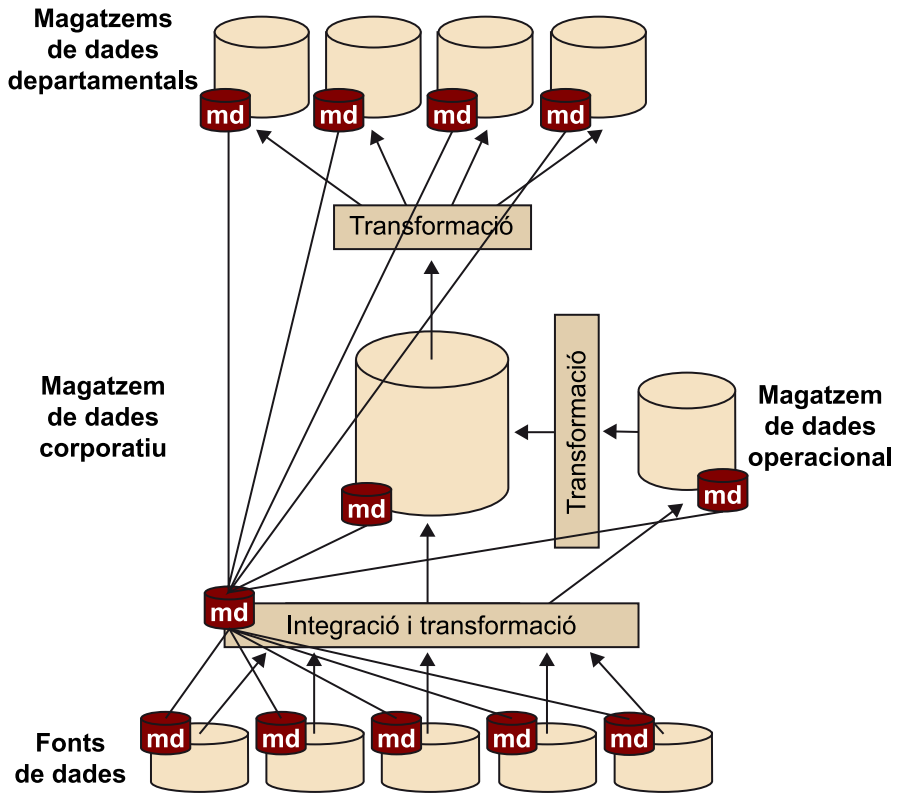
En aquest apartat, comencem revisant l'ús de les metadades a la FIC, a continuació es presenten diferents tipus de metadades segons el seu ús. També s'analitza la manera com es creen les metadades, així com els estàndards definits per permetre compartir-les entre diferents components. Finalment, es comenta la necessitat d'utilitzar diferents versions de metadades a la FIC.

### 4.6.1. Metadades i components de la FIC

A la FIC, es produeix un flux de dades des de les seves fonts fins als analistes. Aquest flux està format per les dades pròpiament dites, que representen característiques d'entitats del món real, i per les metadades, dades que ofereixen informació sobre les altres dades transferides o emmagatzemades.

Les metadades estan associades a tots els components de la FIC tal com s'il·lustra a la figura 8. No obstant això, poden considerar-se com un component per si mateixes. Inmon les defineix dins de la FIC com la «cola» que manté unida la resta dels components, i per aquest motiu les considera com el component més important de la FIC.

Figura 8. Metadades a la FIC




#### 1) Metadades de les fonts de dades

Les bases de dades dels sistemes operacionals o les fonts de dades en general, des del punt de vista de la FIC, tenen com a component fonamental les dades. No obstant això, a més d'aquestes, hi ha metadades que estan generades per eines CASE (*Computer Aided Software Engineering*) si aquestes s'han utilitzat

en la seva construcció; en el cas d'estar construïdes sobre un SGBD, tindrem aquelles que defineixen les bases de dades que intervenen i les relacions entre els seus elements.

Generalment, a les fonts de dades, les metadades descriuran, entre altres característiques, les estructures segons les quals s'emmagatzemen les dades, la quantitat de registres emmagatzemats, la seva forma d'emmagatzematge i les condicions sota les quals es produeixen les dades.

**2) Metadades dels magatzems de dades**

En els magatzems de dades, tindrem les metadades associades als SGBD sobre els quals que estan construïdes, i trobem algunes semblances a les descrites per a les fonts de dades. A més, és possible trobar informació sobre l'ús de les dades per part dels usuaris: estadístiques d'ús, informació sobre seguretat (qui està autoritzat a fer quines operacions), etc.

**3) Metadades en el component d'integració i transformació**

El component d'integració i transformació utilitza les metadades de la resta dels components però, a més, pot definir com a metadades l'origen de les dades, la seva destinació, les transformacions que es fan a les dades de les fonts per obtenir les dels magatzems i la freqüència o el resultat d'aquestes transformacions.

Una vegada definides totes les metadades, a partir d'aquestes es pot generar de manera automàtica el programari que faci la funció d'aquest component. És més fàcil i ràpid mantenir les metadades que mantenir un programari desenvolupat manualment.

Les metadades són el component més important de la FIC, ja que cohesionen la resta dels components dels quals també formen part.

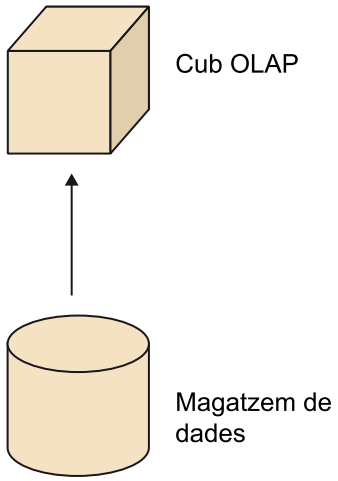
## 4.7. Estructures multidimensionals

La informació incorporada als magatzems de dades de la FIC és explotada i visualitzada des de la capa de presentació. A vegades, de cara a millorar els temps d'accés als magatzems es crea una estructura addicional amb informació agregada. Tenint en compte l'alt nombre de mètriques i dimensions d'anàlisi que podem tenir en un magatzem, les agregacions possibles són molt nombroses. Existeixen estructures d'emmagatzematge que contemplen totes aquestes possibilitats d'agregació com poden ser els cubs OLAP. S'entén per OLAP, o procés analític en línia, el mètode per organitzar i consultar dades sobre una estruc-

tura multidimensional. A diferència de les bases de dades relacionals, totes les consultes potencials estan calculades per endavant, la qual cosa proporciona més agilitat i flexibilitat a l'usuari de negoci.

Un cub OLAP, tal com es pot veure a la figura 9, és un conjunt de cel·les de dades organitzades segons diferents dimensions. Es tracta d'una forma de representació d'una base de dades multidimensional, en la qual l'emmagatzematge físic de les dades es realitza mitjançant una estructura multidimensional. Els cubs es poden considerar com una ampliació de les dues dimensions d'una taula convencional. Aquesta disposició de dades permet una anàlisi ràpida, pel fet de trobar-se gran quantitat de la informació precalculada.

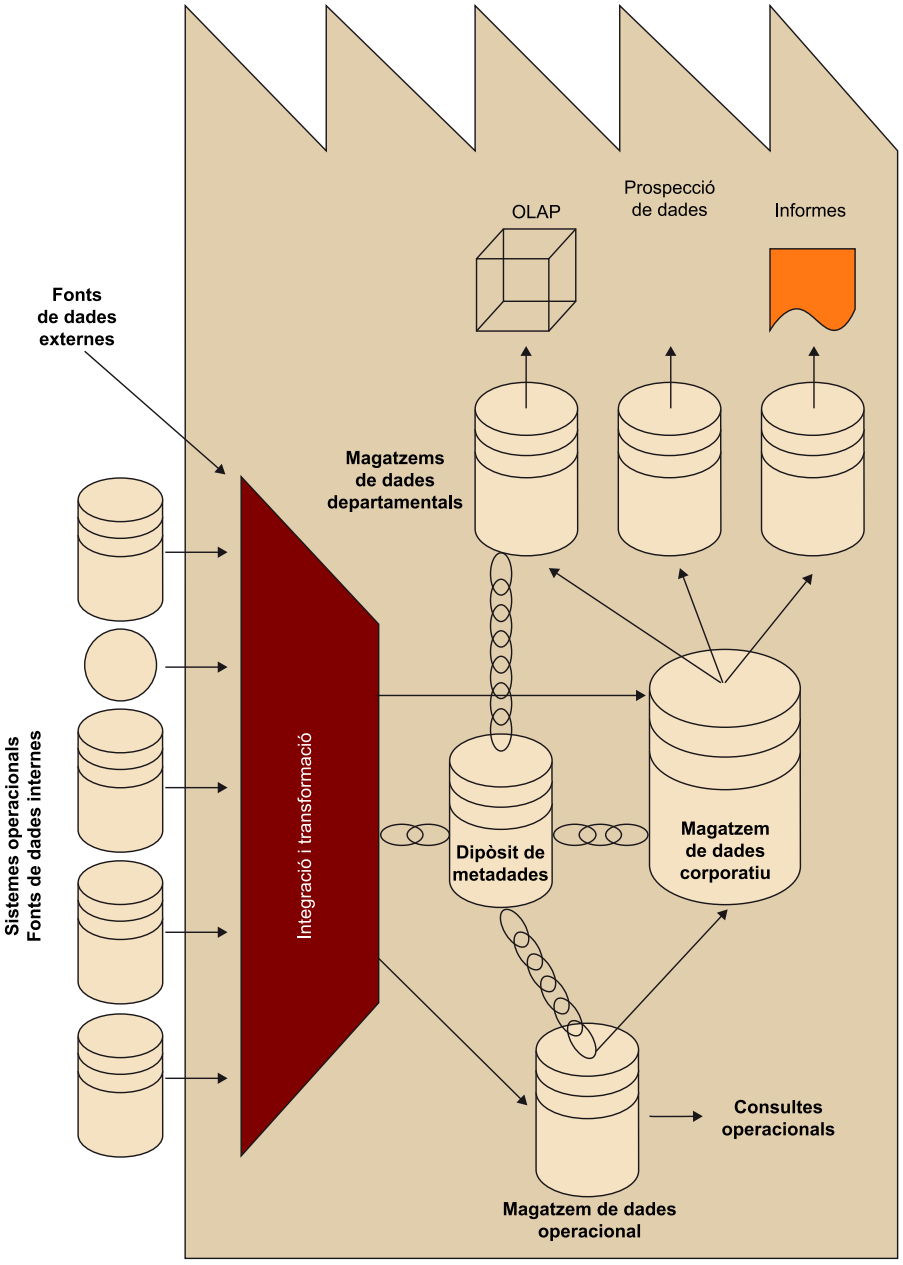
Figura 9. Cub OLAP




## 4.8. Integració components de la FIC

Arribats en aquest punt i coneixent cadascun dels components que formen la factoria d'informació corporativa, en aquest apartat veurem com formen un tot. La figura 10 esquematitza tots els components de la factoria d'informació.

Figura 10. FIC completa




Les dades entren, provinents dels sistemes operacionals de la mateixa empresa o altres fonts de dades externes, directament al component d'integració i transformació. Aquest component de programari les prepara per guardar-les al magatzem de dades operacionalment o directament al magatzem de dades corporatiu. També és aquest component de transformació el que genera una part de les metadades que utilitzaran la resta dels components en el seu funcionament. Les dades del magatzem de dades operacional serviran tant per ser consultades, com per alimentar el magatzem de dades corporatiu. Finalment, segons la utilitat que es donarà a les dades, aquestes es dipositen en petits magatzems de dades departamentals que estan a punt per ser consultats o tractats.

Un error de terminologia també bastant comú és denominar el tot (la factoria d'informació corporativa) com si fos només una part (el magatzem de dades). Es parla d'un component en lloc de parlar del procés que utilitza aquest

component. Stephen R. Gardner defineix l'emmagatzematge de dades com un procés, no un producte, per reunir i governar dades de diferents procedències amb la finalitat d'obtenir una visió única i detallada, total o parcial, d'un negoci. Aquesta idea no sembla tan diferent de la factoria d'informació presentada per William Inmon. Més aviat només és un altre punt de vista, que en certa manera inclou el primer. El fet de parlar d'un procés implica que hi hagi elements que ho facin possible o, com a mínim, que ajudin a fer-ho possible.

Podem considerar la factoria d'informació com el conjunt d'elements que fan possible el procés d'emmagatzematge d'informació. El magatzem de dades simplement seria un component més, com també ho són el repositori de metadades, el component d'integració i transformació, etc.

En aquest punt, encara ens podríem plantejar la necessitat d'aquesta factoria d'informació. Per què cal afegir tota aquesta complexitat als sistemes d'informació de l'empresa? Si ja tenim les dades en els sistemes operacionals, per què els repliquem en la factoria d'informació? Per què els analistes no consulten les dades directament en els sistemes operacionals? No estem malbaratant recursos? Podeu trobar les respostes a aquestes preguntes més o menys implícites en els apartats anteriors d'aquest mateix mòdul, però ara desmentirem de manera explícita aquesta suposada duplicitat de dades:

**a)** Els sistemes operacionals contenen les dades que l'empresa utilitza en el seu dia a dia en l'execució del negoci. En canvi, la factoria d'informació conté dades d'anàlisis, generalment extretes d'aquests sistemes operacionals, però no necessàriament coincidents. Pot haver-hi dades operacionals (per exemple, el número de telèfon dels clients) que no interessin per la presa de decisions i dades molt importants per prendre decisions (com el benefici) que no s'utilitzin en el funcionament diari de l'empresa.

**b)** Generalment, els sistemes operacionals no contenen dades històriques per no retardar de manera innecessària el seu funcionament. En canvi, aquestes dades històriques són imprescindibles a l'hora de prendre decisions.

**c)** Els sistemes operacionals sempre guarden les dades detallades (per exemple, els articles venuts a cada client). En els sistemes decisionals, a vegades, no interessa entrar en tant detall. Només es vol l'import total de la venda, la despesa mensual del client o, simplement, el total venut durant el mes a tots els clients.

**d)** Finalment, una altra diferència entre les bases de dades dels sistemes operacionals i les de la factoria d'informació és que aquestes darreres contenen dades netes. Durant la fase d'entrada de dades a la factoria d'informació, aquestes es netegen, se substitueixen o s'eliminen els valors nuls, es detecten inconsistències, possibles contradiccions entre diferents fonts de dades, etc. En els sistemes operacionals, amb una entrada contínua de dades, no es pot garantir aquesta netedat.

La factoria d'informació no conté les mateixes dades que els sistemes operacionals, malgrat que la intersecció no és buida.

# 5. El magatzem de dades dins d'un sistema de *Data Warehouse*

El magatzem de dades que creem gràcies als processos i elements de la factoria d'informació corporativa és un actiu crític per a la presa de decisions de la companyia i quedarà integrat en els sistemes de tipus informacional o d'intel·ligència de negoci. Sistemes destinats a consultar i analitzar informació, que faciliten la presa de decisions a la companyia dotant de capacitats d'intel·ligència les activitats de gestió i direcció de la companyia.

Un sistema d'intel·ligència de negoci té diferents subsistemes que es recolzen en el magatzem de dades per generar informació. Aquests subsistemes consulten o analitzen la informació de diferents maneres:

- Informes estàndard: informes predefinits que s'executen periòdicament.
- Informes *ad hoc*: informes sota petició per a una consulta de negoci determinada.
- Anàlisi OLAP: anàlisi multidimensional que es recolza en els cubs OLAP.
- Quadres de comandament: quadre resum amb els principals KPI i informes de control.
- Processos d'analítica avançada: processos d'analítica predictiva, sèries temporals, etc.

Tots aquests subsistemes es nodreixen del magatzem de dades, i és crítica la seva correcta actualització en temps i forma.

# 6. Magatzems de dades locals i al núvol

A causa del gran volum de dades que gestionen els *Data Warehouses* i al trànsit de xarxa generat per les operacions d'extracció, transformació i càrrega necessàries per alimentar-los, tradicionalment els sistemes de *Data Warehouse* s'han implementat en local (*on-premise*) en els centres de processament de dades (CPD) de les organitzacions. En l'actualitat, l'evolució de les tecnologies al núvol (*on-cloud*) i les millores dels canals de comunicació permeten plantejar el desplegament de solucions de *Data Warehouse* al núvol.

Cada vegada més organitzacions es plantegen, en les seves estratègies corporatives de dades, la migració al núvol dels seus sistemes de *Data Warehouse*. L'objectiu final d'aquesta transformació digital és la reducció de costos i l'augment de la productivitat.

Aquesta migració al núvol consisteix a substituir els magatzems de dades locals, tradicionalment amb un alt cost de manteniment a causa de la renovació periòdica de llicències de programari i substitució de maquinari, per serveis oferts per centres de dades externs a la mateixa organització. Aquest model conegut com a DWaaS (*Data Warehouse as a Service* ) permet a les organitzacions contractar els serveis analítics necessaris en cada moment, segons la càrrega de treball, la demanda de dades o l'execució de processos, sense necessitat de proveir grans inversions en maquinari.

El gran avenç tecnològic que estan experimentant, en l'actualitat, els sistemes de suport a l'analítica de dades fa que el maquinari dels *Data Warehouses* locals pugui quedar obsolet en un termini de temps que fa difícil rendibilitzar la inversió realitzada per les empreses. Amb el model *on-cloud*, les organitzacions passen de tenir un cost fix amb importants inversions inicials a un cost variable, en funció dels serveis contractats en cada moment.

Aquest canvi de model no és fàcil ni ràpid, per la qual cosa existeixen models híbrids que combinen infraestructura local amb serveis al núvol, facilitant així el procés de migració. Aquest model mixt també és una bona solució final per a organitzacions que, per diferents circumstàncies, requereixen tenir les dades en local, sense renunciar al potencial dels serveis al núvol.

És important destacar que la metodologia de disseny i construcció de la factoria d'informació corporativa és independent de la localització dels magatzems de dades. Tant si el *Data Warehouse* és local com si està al núvol, l'objectiu és comú i tant els components de la FIC com la interacció entre ells han estat descrits amb anterioritat a l'apartat de la factoria de la informació corporativa.

# 7. Explotació i visualització de dades

La visualització de dades és un element clau en les organitzacions. Determina el valor real que s'obté de la informació i aporta més nivell de coneixement sobre les dades disponibles.

El descobriment de dades visual i interactiu que permeten les noves eines d'autoservei de BI ha afavorit l'evolució de les eines tradicionals. Amb les dades d'un sistema BI, és possible generar informes globals o departamentals; crear escenaris respecte a una decisió; fer pronòstics i anàlisis multidimensionals; generar i processar dades; identificar patrons de comportament, etc. Disposar d'informació útil genera coneixement a l'organització i permet millorar el procés de suport a la presa de decisions.

Sens dubte, la cultura organitzativa i el nivell de maduresa en la gestió de dades en el qual es trobi una organització determinaran quins instruments s'implantaran i quan. Les organitzacions són conscients dels problemes que ocasiona la baixa qualitat de les dades registrades, però no sempre es contemplen els inconvenients d'una presentació de les dades inadequada.

Les tècniques de visualització de dades han d'il·lustrar les tendències i les relacions de manera ràpida i senzilla. Són una forma eficient de transmetre la informació des de la base de dades als usuaris finals, perquè puguin prendre decisions basades en dades a escala operativa, tàctica i estratègica.

No obstant això, cal anar amb compte: una mala representació de la informació pot resultar enganyosa. Hi ha moltes maneres de proporcionar informació enganyosa, ja sigui deliberadament o, com passa de manera més freqüent, inintencionadament.

Es poden mostrar les dades de moltes maneres diferents, des de senzills gràfics de barres a diagrames de dispersió més complexos, mapes temàtics i piràmides de població animades. Atès que la interpretació de gràfics pot ser complicada, hi ha d'haver un equilibri entre disseny i funció. Existeixen algunes metodologies formals, que ens ajudaran en el procés de visualització de la informació.

# 8. Administració dels sistemes de *Data Warehouse*

De la mateixa manera que el magatzem de dades és un factor crític per a l'èxit d'un projecte de *Business Intelligence*, també ho són la seva creació i posterior administració.

És necessari gestionar el cicle de vida del *Data Warehouse* des de les primeres etapes de planificació i disseny, passant per la seva posada en marxa i creixement, i arribant fins a les fases de monitoratge, manteniment i optimització, que assegurin un bon funcionament del sistema al llarg del temps.

En la fase de disseny, a més de les dades, cal incloure-hi la tecnologia i les aplicacions que les suportaran. El monitoratge i supervisió és una etapa d'alt valor, si tenim en compte l'elevat grau d'automatització que es pot aconseguir. A més, com que els sistemes d'intel·ligència de negoci evolucionen i creixen en complexitat i requeriments, es fa necessari planificar correctament l'abast del manteniment del sistema. Finalment, l'etapa d'optimització del rendiment del sistema permet que la usabilitat i agilitat que percep l'usuari final siguin notablement millors amb l'evolució del sistema.

Per dur a terme les diferents tasques, els responsables del magatzem de dades disposen de múltiples eines que els seran de gran utilitat i que aportaran flexibilitat i robustesa de manera independent a la plataforma on resideix i al volum de dades que gestionem.

Com en qualsevol projecte de TI, podem gestionar la seva administració fent servir guies i normes existents que recopilen bones pràctiques en l'àmbit de la gestió de serveis i projectes TI (ITIL, ISO, COBIT, PMBOK…). També existeixen experiències d'ús de metodologies àgils (SCRUM i Kanban), que puguin ser aplicades a l'administració dels sistemes de magatzems de dades.

Actualment, les organitzacions prenen consciència de què implementar un magatzem de dades requereix determinar qui hauria de fer les accions descrites, segons sigui el seu perfil, formació, responsabilitat o experiència. L'especialització en les tasques identificades requereix establir rols com el de director de BI, el d'analista de sistemes, el d'administrador de la base de dades, el de desenvolupador d'integracions o aplicacions i els serveis de tercers.

Segons la mida de l'organització i el seu departament de TI, no trobarem tantes persones com rols. Amb freqüència, una persona de l'equip assumeix un o diversos rols.

L'usuari final serà qui rendibilitzi el magatzem de dades. En aquest sentit, podem dir que l'usuari és la peça clau sobre la qual ha de girar l'estratègia d'implantació i serà qui faci de prescriptor d'aquesta eina.

# 9. Tendències actuals

Des de la concepció del magatzem de dades, les tecnologies i tècniques d'implementació han evolucionat per adaptar-se a les necessitats de les organitzacions. En l'actualitat, hi ha diversos factors que condicionen l'evolució dels magatzems de dades:

**1) Creixement exponencial de l'univers digital**. Els usuaris i les xarxes de sensors dupliquen anualment les dades de les organitzacions, i amb freqüència aquestes no estan estructurades. Aquest creixement no només planteja un repte pel que fa a l'emmagatzematge, sinó també en la gestió i manipulació de les dades. Ens referim, doncs, a un problema que té tres dimensions: velocitat de generació de les dades, volumetria de les dades i variabilitat de les dades, que són les tres dimensions que caracteritzen les tecnologies denominades *Big Data*.

Bona part del creixement de dades prové d'informació no estructurada. L'emmagatzematge i processament d'aquest tipus d'informació suposa un repte i canvis en els magatzems de dades tradicionals, que es basaven en una estructura i esquema incompatible amb la naturalesa d'aquest tipus d'informació no estructurada.

Les dades estructurades tenen un esquema identificable i típicament es poden emmagatzemar en taules de bases de dades relacionals. Les dades no estructurades són aquelles que no tenen ni un format ni un esquema definits que permetin emmagatzemar-les en aquestes bases de dades. Estaríem parlant de dades com: imatges, àudios, vídeos, dades de xarxes socials, correus electrònics, dades generades per dispositius, sensors, etc.

Existeix una tercera categoria de dades coneguda com a dades semiestructurades. Es caracteritza per tenir certa organització de les metadades, sense arribar a tenir un esquema clarament definit. En aquesta categoria estaríem parlant, per exemple, de documents XML, HTML o JSON.

**2) Noves tècniques de modelització**. Daniel Linstedt va publicar l'any 2000 una nova tècnica denominada *data vault*. El seu objectiu era la creació de magatzems de dades flexibles i auditables en temps real. Es tracta d'una tècnica de modelització basada en tres tipus d'entitats:

- *Hubs*: conté els indicadors clau de negoci.
- *Links*: conté les relacions.
- *Satellites*: conté les descripcions.

Planteja una situació intermèdia entre la modelització en tercera forma normal i l'esquema en estrella. En aquesta tècnica preval la flexibilitat i l'escalabilitat, i permet que el model pugui adaptar-se de manera senzilla als canvis en el negoci i en l'organització.

**Vegeu també**

L'esquema en estrella que es desenvolupa detalladament en el mòdul «Disseny i implementació multidimensional d'un *Data Warehouse*» d'aquesta assignatura.

El desenvolupament d'un *data vault* es realitza en una sèrie d'etapes:

- Identificar els *hubs*.
- Establir les relacions (*links*).
- Establir les descripcions (*satellites*).
- Afegir components independents com calendaris o taules de relació.
- Afegir taules necessàries per millorar el rendiment: taules pont, estructures *point-in-time*, etc.

**3) Maduresa de tecnologies de manipulació de dades**. Les organitzacions actuals necessiten suport en la presa de decisions, i aquesta es fonamenta en dades de negoci que sovint requereixen temps. Aquest fet ha motivat l'aparició de tecnologies de complement del magatzem de dades tradicional. A continuació, s'esmenten les següents.

**a) Anàlisi contínua de dades**: mitjançant fluxos continus de dades, es poden analitzar dades en temps real de manera contínua. Un possible cas d'ús podria contextualitzar-se en el monitoratge del trànsit d'una ciutat. Suposem que cal identificar els punts on es produeixen incidències, habilitar en temps real una alerta basada en patrons i, a continuació, automatitzar algunes accions que cal prendre per reduir el nombre d'incidències. Aquestes accions podrien consistir en avisar al personal de manteniment o canviar el comportament dels elements de la xarxa.

**b) Processament d'esdeveniments complexos**: permet identificar patrons dins dels processos de negoci i automatitzar algunes accions que es repeteixen. Per exemple, si s'identifiquen clients que compleixen certes característiques, es podrien automatitzar ofertes dirigides a clients que segueixen un mateix patró.

**c) BI d'autoservei**: en l'actualitat existeixen eines de *Business Intelligence* que escurcen els cicles de creació de quadres de comandament i faciliten una interfície àgil i amb importants capacitats de connexió i integració. Aquestes solucions poden treballar connectant-se al magatzem de dades o directament a les fonts de dades origen. El paper que té en aquest tipus de solucions el magatzem de dades no és tan crític com en les solucions de *Business Intelligence* tradicionals.

**d) Bases de dades en memòria**: mitjançant la memòria d'un servidor que utilitza tècniques OLAP, aquestes bases de dades permeten analitzar dades de gran volumetria en temps real. Amb freqüència, aquesta tecnologia dona suport a les tecnologies anteriors.

**e) L'ús d'infraestructura de maquinari que dona suport als *Data Warehouse*** ha canviat i han aparegut nous tipus d'emmagatzematge físic com els discos d'estat sòlid que milloren notablement els processos de lectura i escriptura. D'altra banda, l'opció de processar i emmagatzemar en entorns de computació al núvol cobra cada vegada més rellevància en les organitzacions.

**f)** Els **canvis en les infraestructures** i la necessitat creixent de **disposarde servidors dedicats per a tasques analítiques** ha donat lloc a l'aparició dels *Data Warehouse Appliances* que consisteixen en una plataforma de maquinari i programari orientada al *datawarehousing* i processos analítics. Molts fabricants tenen aquest tipus de plataformes que ofereixen: maquinari, sistema operatiu i base de dades optimitzats per a *dawarehousing*.

**g) *Data Warehouse* al núvol**. Els principals fabricants ofereixen solucions per treballar amb magatzems de dades al núvol, bé centrant la solució en el *Data Warehouse*, o ben integrat dins d'una solució *Business Intelligence* completa. Existeix l'opció de la infraestructura híbrida que combina solucions al núvol amb solucions en servidors i infraestructura pròpia.

**h) *Virtual Data Marts***: aparició de *Data Marts* virtuals basats en la federació de dades. Aquests *Data Marts* virtuals no existeixen físicament i es creen generant una capa virtual que surt del magatzem de dades corporatiu. Redueixen el moviment de dades, tot i que presenten les limitacions habituals pel que fa a temps de resposta i suport de regles de negoci complexes que té la federació de dades.

**i) Hadoop, MapReduce, Spark i altres tecnologies equivalents**: empreses com Google, Amazon o Facebook gestionen diàriament una gran quantitat de dades que han de ser introduïdes al sistema i consultades en temps real. Amb aquesta finalitat, amb freqüència es treballa amb xarxes de servidors que es consulten en paral·lel i amb bases de dades en columnes o altres SGBD no relacionals. Aquest enfocament es coneix com a NoSQL (ja que no només utilitza el llenguatge SQL).

**4) Analítica de negoci**. Utilitza tècniques estadístiques i de mineria de dades en processos operatius de negoci. L'objectiu és facilitar les decisions relatives a l'operativa i proposar tàctiques de negoci basades en prediccions. Alguns fabricants especialitzats en magatzems de dades inclouen algorismes per facilitar la creació d'aquest tipus d'avantatges competitius. Els magatzems de dades són emprats, amb freqüència, com a origen de dades d'aquests processos.

**5) Convivència entre els *Data Warehouse* i entorns *Big Data* com Hadoop**. Tal com s'ha assenyalat, les tecnologies *Big Data* han experimentat un important creixement en els últims anys. Actualment, a les companyies hi conviuen tecnologies *Big Data*, com Hadoop, amb els magatzems de dades. L'ecosistema Hadoop ha anat adquirint aquests darrers anys una funció cada vegada més important en la gestió de la informació i els processos d'anàlisi de les companyies. Actualment hi ha diferents situacions de convivència entre els magatzems de dades i l'ecosistema Hadoop d'acord amb el paper exercit per aquest últim.

Aquesta convivència entre els *Data Warehouses* i els *Data Lakes* permet enriquir i potenciar l'entorn analític de les companyies. Un *Data Lake* és una arquitectura d'emmagatzematge de dades massives en brut, capaç d'emmagatzemar dades no estructurades en el seu format original. Hadoop és el principal component tecnològic d'aquesta arquitectura orientada a entorns *Big Data*.

Existeixen diversos models on els *Data Warehouses* i les tecnologies com Hadoop es complementen per adaptar la solució final a les necessitats de cada organització.

**a) Hadoop com a ODS (*Operational Data Store*)**: en una primera etapa d'implantació, Hadoop rep tota la informació no estructurada o generada en temps real, atès que aquest tipus d'informació és molt costosa d'emmagatzemar i gestionar a les bases de dades relacionals. La infraestructura d'emmagatzematge de Hadoop és econòmica i molt escalable, la qual cosa produeix que Hadoop també recopili informació estructurada i pugui emmagatzemar fonts de dades que alimenten els magatzems de dades, dins d'aquesta configuració, el creixement de Hadoop es dirigeix a convertir-se en l'ODS dels magatzems de dades, recopilant tota la informació en brut que posteriorment es consolida als magatzems. Es tracta d'un ODS amb una gran capacitat d'emmagatzematge, de creixement i molt eficaç per recopilar dades generades en temps real. En aquesta configuració la interacció entre Hadoop i els magatzems de dades és la necessària per comunicar els magatzems amb l'ODS.

**b) Hadoop com a origen de dades del *Data Warehouse***: avui dia, hi ha moltes organitzacions que centren el BI en els magatzems de dades que, al seu torn, reben dades dels *Data Lakes*. Aquest tipus d'integració permet a les organitzacions rendibilitzar la seva infraestructura analítica, potenciant-la amb les noves característiques dels *Data Lakes*. Aquesta arquitectura dona suport tant a les dades estructurades en els *Data Warehouses*, com a les no estructurades en els *Data Lakes*, incrementant així la capacitat de gestió de grans volums de dades amb esquemes flexibles, dissenyats per adaptar-se a canvis freqüents. Prepara l'entorn analític de les companyies per afrontar projectes d'intel·ligència artificial i *Machine Learning*, amb un cost limitat gràcies a l'escalabilitat horitzontal del model.

**c) Processos d'anàlisi i *Business Intelligence* (BI) sobre Hadoop**: el paper que pugui exercir Hadoop dins dels processos d'anàlisi i BI també determina la interacció entre Hadoop i els magatzems de dades. Cada vegada hi ha més eines d'anàlisi o visualització de dades que s'integren amb Hadoop i permeten realitzar determinats processos de BI en aquest entorn, de manera que serà necessari portar informació dels magatzems de dades a Hadoop per enriquir aquests processos d'anàlisi.

# Resum

En aquest mòdul hem introduït el concepte de magatzem de dades per disposar dels fonaments suficients per a la resta de l'assignatura.

Primer, hem explicat què és un magatzem de dades i vist que en realitat no és un concepte nou, ja que de manera implícita s'estava utilitzant amb altres eines. Hem vist que els centres d'informació han estat els precursors del magatzem de dades. A continuació, hem definit el magatzem de dades segons Inmon i hem repassat les seves característiques principals: orientació al tema, integració, no volatilitat i dades històriques.

A més, hem vist que els magatzems de dades no són un altre tipus d'organització de bases de dades, sinó que atorguen un valor afegit molt important a l'organització pel fet d'aportar més coneixement a l'empresa i ajudar-la en la presa de decisions. S'han comparat les bases de dades operacionals amb els magatzems de dades i s'ha vist que les diferències són realment molt importants.

També s'ha introduït el concepte de la factoria de la informació corporativa, detallant els seus principals components: magatzem de dades departamental, corporatiu, operacional, el component d'integració i transformació, les estructures multidimensionals i les metadades. Així mateix, s'ha analitzat el paper del magatzem de dades dins d'un sistema informacional.

Finalment, s'han repassat les tendències actuals en els magatzems de dades, posant l'accent en l'evolució de tecnologies com ara el *Big Data*, l'analítica avançada, l'anàlisi contínua de dades, els canvis en la infraestructura (*cloud*, bases de dades en memòria...) o el creixement del *Business Intelligence* d'autoservei, entre d'altres.

## Activitats

**1.** Proposeu al fòrum quin projecte de magatzem de dades voleu desenvolupar, que correspongui, si és possible, amb la vostra àrea d'activitat professional.

**a)** Expliqueu quins són els objectius d'aquest projecte.

**b)** Quines dades creieu que són rellevants per aconseguir-ho?

**c)** Quina diferència veieu amb el projecte de base de dades operacional en el cas que n'hi hagi algun?

**2.** Busqueu per la xarxa els cinc projectes de magatzem de dades que estan desenvolupats i que cregueu que són més interessants.

**a)** Quins són els objectius que té cada projecte?

**b)** Us en sorprèn algun? Per què?

**c)** Compartiu aquestes experiències al fòrum.

## Exercicis d'autoavaluació

**1.** En quines característiques es basen els magatzems de dades?

**2.** Tenim una base de dades operacional que està perfectament normalitzada i els processos que treballen sobre aquesta base de dades són molt ràpids.
**a)** Ens serviria aquesta estructura per fer processos per prendre decisions?
**b)** Si és que no, quines diferències caldria implementar per construir un magatzem de dades?

**3.** Quina és la diferència principal entre el magatzem de dades corporatiu i el departamental?

**4.** Com justificaríeu la necessitat del magatzem de dades operacional?

**5.** Per què els magatzems de dades departamentals no s'alimenten directament dels sistemes operacionals, en lloc de fer-ho del magatzem de dades corporatiu?

**6.** Quines operacions duu a terme el component d'integració i transformació?

**7.** Quin és l'element principal del component d'integració i transformació?

**8.** Quin paper tenen les metadades a la FIC?

**9.** Hi ha redundància entre les dades de les bases de dades operacionals i les de la factoria d'informació corporativa?

**10.** Ompliu la taula següent indicant les principals diferències que hi ha entre els sistemes operacionals i decisionals:

| Característica | Sistemes operacionals | Sistemes decisionals |
|---|---|---|
| Usuaris típics | | |
| Nombre d'usuaris | | |
| Tuples a les quals s'ha accedit | | |
| Objectiu del sistema | | |
| Funcions principals | | |
| Disseny | | |
| Característiques de les dades | | |
| Ús | | |

| Característica | Sistemes operacionals | Sistemes decisionals |
|---|---|---|
| Accés | | |
| Unitat de treball | | |
| Requeriments | | |
| Mida | | |

# Solucionari

## Exercicis d'autoavaluació

**1.** Les característiques principals d'un magatzem de dades són l'orientació a temes, la integració, la no volatilitat i les dades històriques. Aquestes característiques es basen en la filosofia que Inmon va descriure.

**2.**
**a)** No serveix la mateixa estructura.

**b)** Des del punt de vista del disseny, hi ha diferències en la temporalització, el volum de dades, el nivell d'agregació, l'actualització i l'estructuració. Des del punt de vista del tractament de la informació, les diferències són d'explotació de la informació i de temps de resposta. Per acabar, des del punt de vista de funcionalitats, hi ha diferències en les activitats, en la importància de les dades i en els usuaris finals.

**3.** La diferència principal és la mida. Mentre el magatzem de dades corporatiu conté totes les dades que interessen o poden arribar a interessar a qualsevol persona de l'empresa, un magatzem departamental només conté aquelles que en un moment determinat interessen a un conjunt d'analistes.

**4.** El magatzem de dades operacional serveix per satisfer de manera eficient i sense interferir en els sistemes operacionals les necessitats d'accés integrat a dades no històriques.

**5.** Si carreguem les dades dels magatzems de dades departamentals directament de les bases de dades operacionals, multipliquem els processos necessaris d'integració i transformació de les dades.

**6.** El component d'integració i transformació obté les dades de les fonts de dades, les depura, transforma i integra, les transporta als magatzems de dades i les hi carrega. També obté dades del magatzem de dades operacional i les transforma, transporta i carrega al magatzem de dades corporatiu. A més, fa la mateixa operació entre el magatzem de dades corporatiu i els magatzems de dades departamentals.

**7.** A diferència d'altres components de la FIC, l'element principal dels quals és la base de dades, l'element principal del component d'integració i transformació és el programari que implementa la seva missió.

**8.** Generalment, les metadades són dades que ens donen informació sobre altres dades. En la FIC, són el component que s'encarrega de cohesionar la resta dels components.

**9.** Sí, hi ha algunes dades que estan en els dos sistemes. No obstant això, aquesta redundància és mínima i necessària, ja que els sistemes operacionals no guarden dades històriques, ni agregades, ni han passat un procés de neteja i integració.

**10.** Les diferències principals entre els sistemes operacionals i els decisionals són les següents:

| Característica | Sistemes operacionals | Sistemes decisionals |
|---|---|---|
| Usuaris típics | Administratius | Analistes (executius) |
| Nombre d'usuaris | Milers | Centenars |
| Tuples a les quals s'ha accedit | Centenars | Milers |
| Objectiu del sistema | Execució del negoci | Anàlisi del negoci |
| Funcions principals | Operacions diàries (OLTP) | Presa de decisions (OLAP) |
| Disseny | Orientat a la funcionalitat | Orientat al tema |
| Característiques de les dades | Actuals i actualitzades, atòmiques, normalitzades, aïllades | Històriques, resumides (agregades), desnormalitzades, integrades |
| Ús | Repetitiu i rutinari (consultes predeterminades) | Esporàdic i innovador (consultes ad hoc) |
| Accés | R/W | Principalment lectura |

| Característica | Sistemes operacionals | Sistemes decisionals |
|---|---|---|
| Unitat de treball | Transaccions simples | Consultes complexes |
| Requeriments | Rendiment de transacci-ons + consistència de dades | Rendiment de les consul-tes i precisió de les dades |
| Mida | MB/GB | GB/TB |

# Glossari

**base de dades operacional** *f* Base de dades destinada a gestionar el dia a dia d'una organització, és a dir, a emmagatzemar la informació referent a l'operativa diària d'una institució.

***Big Data*** Conjunt d'estratègies, tecnologies i sistemes per a l'emmagatzematge, processament, anàlisi i visualització de conjunts de dades complexos.

***cloud*** Computació al núvol, coneguda també com a serveis al núvol, informàtica al núvol, núvol de còmput o núvol de conceptes (de l'anglès *cloud computing*), és un paradigma que permet oferir serveis de computació a través d'una xarxa, que usualment és Internet.

**dada** (definició des del punt de vista dels sistemes **decisionals**) m Mesura, observació feta i emmagatzemada en algun sistema.

**data governance** Vegeu govern de la dada.

**data vault** Conjunt de magatzems de dades flexibles i auditables en temps real.

**Data Warehouse** Vegeu magatzem de dades.

**Data Warehouse Appliances** Plataforma de maquinari i programari orientada a *datawarehousing* i processos analítics.

**factoria d'informació corporativa** *f* Conjunt d'elements de programari i maquinari que ajuden a l'anàlisi de dades per prendre decisions. Sigla FIC.

**FIC** *f* Vegeu factoria d'informació corporativa.

**gestió de dades mestres** *f* Metodologia que identifica la informació més crítica d'una organització i crea una única font fiable.

**govern de la dada** *m* Metodologia que té per objecte assegurar que les dades són sempre fiables i vàlides en cada context empresarial, que la qualitat es manté al llarg del temps i que existeixen mecanismes de control sobre qui pot fer què amb les dades a cada moment.

**magatzem de dades** *m* Bases de dades orientades a àrees d'interès de l'empresa que integren dades de diferents fonts amb informació històrica i no volàtil i que tenen com a objectiu principal recolzar en la presa de decisions en *Data Warehouse*.

**magatzem de dades corporatiu** *m* Conjunt de dades que guarda integrades totes les dades històriques de l'empresa.

**magatzem de dades departamental** *m* Conjunt de dades que resol les necessitats d'anàlisi d'un departament o conjunt d'usuaris.

**magatzem de dades operacional** *m* Conjunt de dades integrat i orientat al tema, però sense dades històriques. Se sol utilitzar com a pas intermedi en la construcció del magatzem de dades corporatiu.

**master data management** Vegeu gestió de dades mestres.

**metadada** *m* Dades sobre dades.

**OLAP** Sigles que fan referència a les eines d'anàlisi normalment multidimensional a *on-line analytical processing*.

**OLTP** Sigles d'*on-line transaccional processing*.

**SGBD** Vegeu sistema de gestió de bases de dades.

**sistema de gestió de bases de dades** *m* Programari que gestiona i controla bases de dades. Les seves funcions principals són les de facilitar el seu ús simultani a molts usuaris de diferents tipus, independitzar l'usuari del món físic i mantenir la integritat de les dades. Sigla SGBD.

**sistema de registre** *m* Font de cadascuna de les dades dels magatzems de dades, d'entre totes les fonts possibles.

**sistema informacional** *m* Sistema que dona suport als processos de presa de decisions per part dels analistes a l'organització.

**sistema operacional** *m* Sistema que ajuda a les operacions diàries del negoci d'una organització.

**sistema transaccional** *m* Sistema basat en transaccions de lectura/escriptura.

**transacció** *f* Conjunt d'operacions de lectura i/o actualització de la base de dades que acaba confirmant o cancel·lant els canvis que s'han dut a terme.

# Bibliografia

**Davenport, T.; Harris, J.** (2008). *Competing on Analytics*. Boston: Harvard Business School Press.

**Franco, J. M.; EDS-Institut Prométhéus** (1997). *El Data Warehouse - El Data Mining*. Barcelona: Gestió 2000.

**Gill, H. S.; Rao, P. C.** (1996). *Data Warehousing. La integració per a la millor presa de decisions*. Mèxic: Prentice Hall.

**Inmon, W. H.; Hackathorn, R. D.** (1994). *Using the data warehouse*. Nova York: Wiley.

**Inmon, W. H.; Strauss, D.; Neushloss, G.** (2010). *DW 2.0: The Architecture for the Next Generation of Data Warehousing*. Burlington, Mass.: Morgan Kaufman Series in Data Management Systems.

**Kimball, R.** (2002). *The Data warehouse toolkit: the complete guide to dimensional modeling*. Nova York: Wiley.