Introducció a l'emmagatzematge de dades

Àngels Rius Gavídia
Montse Serra Vizern
Josep Curto Díaz

PID_00189742

Universitat Oberta de Catalunya

Índex

Introducció

Fins ara hem estudiat les bases de dades relacionals que són les que, majoritàriament, avui dia hi ha implantades en la indústria. Aquest tipus de bases de dades dóna suport al negoci de l'organització i permet d'emmagatzemar les dades i el processament de la informació generada dia a dia. Tot això implica que estan dissenyades per a fer 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 a fer una factura, omplir un historial mèdic, gestionar una assegurança de vida, etc.

Aquesta assignatura presenta un nou tipus de bases de dades que estan orientades a donar suport a la presa de decisions en l'organització, són els anomenats *magatzems de dades*.

Magatzems de dades

Magatzem de dades és la traducció de *data warehouse*.

L'objectiu principal del magatzem de dades és treure rendiment de la informació emmagatzemada i això vol dir extreure les dades per a una anàlisi posterior que ajudi a prendre decisions. Això significa que aquest tipus de bases de dades té un enfocament diferent respecte a 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.

Finalment cal comentar que en aquests darrers anys han sorgit amb molta força el que s'anomena *bases de dades informatives*. Com el seu nom indica es tracta de bases de dades orientades a proporcionar informació (especialment per web). Ara bé, aquest tema no el tractarem en aquesta assignatura, ja que possiblement en un futur podria esdevenir una nova assignatura.

Objectius

Els materials didàctics inclosos en aquest mòdul s'orienten a aconseguir que l'estudiant assoleixi els 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 per què han aparegut.

3. Comprendre la importància dels processos de presa de decisions en el món dels sistemes d'informació i quin paper hi té el magatzem de dades.

4. Saber distingir les característiques i els objectius principals que tenen els magatzems de dades i saber-los diferenciar de les bases de dades operacionals.

1. Què és un magatzem de dades?

Els sistemes tradicionals sempre han tingut dificultats per a satisfer les necessitats informacionals d'una organització. Arran d'aquest fet sorgeixen solucions per a dotar les empreses o grups d'empreses d'eines capaces de solucionar les seves necessitats informacionals globals. Una de les solucions principals és el magatzem de dades.

L'avantatge principal dels magatzems de dades consisteix en la capacitat d'emmagatzemar la informació de manera homogènia i fiable en una estructura de dades jeràrquica pensada per a facilitar les consultes estratègiques de la direcció de l'organització.

El terme *magatzem de dades* ha estat concebut per Bill Inmon i R. D. Hackathorn. La definició proporcionada per Bill Inmon és la següent:

Lectura recomanada

W. H. Inmon; R. D. Hackathorn (1994). *Using the Data Warehouse*. Nova York: Wiley.

> El 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.

D'aquesta definició es desprèn el fet que som davant un nou tipus de bases de dades la importància del qual rau en el suport que pot oferir a les organitzacions des del punt de vista estratègic i que a primera vista sembla que no és gaire difícil de construir. La dificultat principal a l'hora de portar a terme la creació d'un magatzem de dades està en el fet de saber, *a priori*, quines dades es necessiten i de quina manera s'han d'organitzar.

Quantes empreses que volen portar a terme projectes d'aquests tipus tenen clares les dades que necessiten en el magatzem de dades? L'experiència ens indica que hi ha molt poques empreses que ho tinguin realment clar. Algunes d'aquestes no saben que no tenen prou acurades aquelles dades que volen introduir per a poder-ne treure resultats que serveixin per a donar suport a la presa de decisions.

2. Evolució històrica

Des de fa pocs anys (meitat dels anys noranta) assistim a l'explosió del fenomen dels magatzems de dades. El motiu principal que ha impulsat que apareguin és l'enorme quantitat de dades mal explotades.

Abans no hi havia dades mal gestionades? Aquesta recerca de la rendibilitat del negoci no hi era abans? Com es prenien les decisions estratègiques?

El primer intent important de desenvolupar un sistema d'informació diferent dels sistemes operacionals el va portar a terme IBM a mitjan dècada dels setanta (Art Benjamin, Canadà). L'objectiu que es va marcar va ser la possibilitat de generar informes sense la necessitat que els programadors dels sistemes operacionals fossin els qui haguessin de fer un nou desenvolupament cada vegada. Per això es van utilitzar unes eines flexibles d'integració i generació d'informes *ad hoc* a partir de la informació continguda en els fitxers dels sistemes operacionals. D'aquesta manera s'aconseguien una sèrie de beneficis: usuaris més ben servits i estalvi de costos.

Aquest primer *centre d'informació* aconseguia desplaçar l'esforç de manteniment dels sistemes d'un 77% a un 33%, a causa del fet que, almenys en aquell moment, la majoria dels manteniments se centraven a corregir, millorar o crear nous informes.

El 1980 Steven Alter va diferenciar, formalment, dos tipus de sistemes d'informació: els EDP i els DSS. Els EDP estan destinats al processament de les dades operatives de l'empresa i els DSS estan orientats a la presa de decisions. Aquests dos tipus de sistemes tenen una clara intersecció: els MIS, que permeten d'elaborar informes estàndard per als directius.

EDP, DSS i MIS

EDP: processament electrònic de dades (*electronic data processing*).

DSS: sistemes de suport a la decisió (*decision support systems*).

MIS: sistemes d'informació per a la gestió (*management information systems*).

Un altre fet significatiu en l'evolució dels sistemes d'informació és el naixement, al principi dels anys vuitanta, de la informàtica d'usuari final, a partir de la introducció de l'ordinador personal (PC) i de la primera revolució ofimàtica: fulls de càlcul, tractament de textos, bases de fitxers personals, etc.

Aquesta revolució va permetre d'augmentar la productivitat individual enfront dels desenvolupaments tradicionals, ja que a partir d'aquestes eines informàtiques, els mateixos usuaris, o persones molt properes a ells, eren capaços, per mitjans poc "consistents" moltes vegades, d'elaborar informes amb més qualitat i més ràpidament que les àrees de desenvolupament.

Van començar a coexistir dos tipus d'informàtics: els especialistes de la programació que es dediquen al desenvolupament i els usuaris finals de determinades àrees de negoci amb coneixements mínims d'ofimàtica. Com a conseqüència d'aquests dos perfils informàtics i de la diferència de criteris entre aquests, es va desencadenar la necessitat de crear els centres d'informació.

Un centre d'informació està format per una sèrie de fitxers i taules que resideixen, normalment, en una partició del mateix ordinador central en què s'executen els sistemes operacionals. Aquestes taules solen ser una còpia de fitxers operacionals que es consideren importants i que s'han bolcat als centres d'informació segons s'ha necessitat.

L'accés a les dades dels centres d'informació es fa per mitjà d'eines flexibles que utilitzen "pseudoprogramadors", que són els responsables d'elaborar i repartir els informes, i també d'extreure fitxers per als usuaris finals. Aquests fitxers es manipulen en els PC de manera individualitzada, per mitjà d'eines informàtiques, últimament en bases de dades personals, per a elaborar gràfics, simulacions, estadístiques, previsions, etc.

Amb la introducció de les xarxes locals, la manipulació de les dades que provenen dels centres d'informació s'ha pogut fer en alguns casos en l'àmbit departamental. La creixent potència i abaratiment dels PC ha permès d'escalar aquestes "manipulacions" a autèntics sistemes de mida i potència elevades.

Durant els anys vuitanta s'imposen les bases de dades relacionals i als anys noranta el seu desplegament arriba al seu màxim abast. Com hem comentat anteriorment, aquestes bases de dades presenten mancances si es volen fer servir per a prendre decisions i, per aquest motiu, sorgeix un altre tipus de sistemes d'informació: els magatzems de dades.

Tot seguit podem veure, des de dos punts de vista diferents, què afavoreix que aparegui aquest nou tipus de sistema d'informació.

Punt de vista socioeconòmic

A causa del nou marc mundial en les relacions comercials establerts per la creació de l'Organització Mundial del Lliure Comerç l'any 1995, els requeriments de les organitzacions han canviat.

La globalització

El fet de la globalització no solament ha tingut incidència en la informàtica, sinó també en qualsevol entorn de la societat.

Arran del nou marc mundial, les relacions comercials tenen les característiques següents:

- Un mercat global (globalització) que es caracteritza per la supressió de barreres proteccionistes aranzelàries.

- Una competència més gran entre els diferents sectors.

- Una disminució dels marges d'explotació, la qual cosa implica menys guanys per a cada operació i, en alguns casos, les conseqüents reduccions de plantilla.

- L'augment del nombre d'operacions per a obtenir guanys similars als anteriors.

- La fidelització necessària dels clients amb la conseqüent millora del servei postvenda.

Totes aquestes característiques obliguen els sistemes d'informació a readaptar-se:

- La necessitat no solament de conèixer el dia a dia del comportament de l'organització, sinó també d'anticipar-se als canvis que la nova dinàmica comercial genera implica la necessitat d'emmagatzemar informació nova.

- En les organitzacions es creen i destrueixen departaments i seccions de manera tan ràpida que produeixen importants canvis en els models organitzatius. Aquests canvis no estan previstos en els sistemes d'informació tradicionals.

- Es fa necessari obtenir informació més detallada, dinàmica i per períodes de temps.

Punt de vista informàtic

Les eines informàtiques, a partir del 1994, han progressat de tal manera tant des del punt de vista quantitatiu com des del punt de vista qualitatiu que són possibles tantes operacions analítiques com consultes d'usuari final sobre volums de dades enormes.

L'arribada de les arquitectures client/servidor ha estat el que ha convertit els magatzems de dades en una autèntica disciplina per a la professió.

3. Característiques d'un magatzem de dades

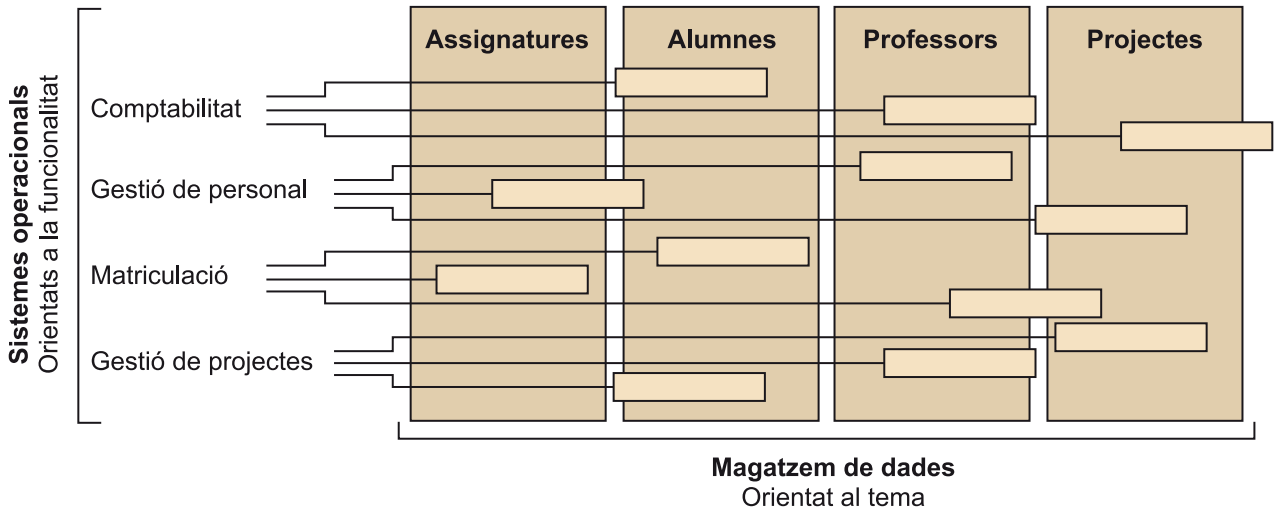
Com s'ha vist anteriorment, el magatzem de dades representa una novetat en el tractament de la informació. Per a poder dur a terme un tractament adequat de la informació, el magatzem de dades ha de complir un conjunt de característiques: que sigui orientat al tema, que les dades estiguin integrades i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

3.1. Orientat al tema

Aquesta primera característica fa referència a les directrius dels dissenyadors dels magatzems de dades. El disseny dels sistemes operacionals és donat per un conjunt de requeriments, ja que es construeixen per satisfer una necessitat concreta i ben coneguda. Així, 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 requeriments concrets que tenen, ni l'ús que es pot arribar a fer de les dades que guardem (això es decidirà molt després, quan aparegui la necessitat de fer un estudi concret). Conseqüentment, l'única cosa que el dissenyador pot considerar en aquest cas són les àrees o 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 (per la qual s'utilitzaran), sinó dividida per temes d'interès.



Com es veu en la figura superior, cada sistema operacional (en aquest cas, d'una universitat) accedeix exactament a les dades que li calen i 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. Però probablement no accedirà a totes perquè algunes, com ara les notes dels estudiants, no li calen. En canvi, un magatzem de dades desa les dades segons els possibles temes que es poden analitzar. Tingueu en compte que no sabem quina utilitat concreta es donarà a les dades emmagatzemades. Simplement es guarden per quan calgui analitzar-les (però, per ara, no sabem ni quan ni com caldrà fer-ho). A més, no guardarem totes les dades dels sistemes operacionals, perquè algunes no pertanyen a cap tema d'anàlisi que ens interessi (per exemple, els números de telèfon dels estudiants).

3.2. Integració de dades

Sabem que els sistemes operacionals de les empreses són heterogenis: funcionen sobre maquinari i programari diferent, utilitzen models de dades diferents (uns orientats a l'objecte, d'altres relacionals, etc.) i presenten el negoci des de diferents punts de vista (finances, vendes, gestió de personal, etc.). Per tant, el primer pas per a oferir totes les dades als analistes ha de ser integrar tots aquests sistemes, de manera que els analistes, malgrat que les dades provinguin de fonts diferents, ho 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'informàtics, sinó d'usuaris no experts als quals s'ha de facilitar la feina. 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 a mencionar-ne només alguns, podríem parlar d'unificar els tipus i estructures de dades, definir claus primàries comunes, 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).

Integrar

Integrar no és simplement posar les dades en un repositori comú. Aquestes dades han de passar un procés d'integració i transformació que veurem detalladament més endavant en el mòdul "La factoria d'informació corporativa".

A més, cal mencionar que els magatzems de dades disposen d'un component que ajuda a integrar: les metadades.

Les metadades

Estudiarem les metadades detalladament en el mòdul "La factoria d'informació corporativa", però podem avançar dient que permeten de simplificar i automatitzar l'obtenció de la informació des dels sistemes operacionals fins als sistemes informacionals i, per tant, són bàsiques per al procés d'integració.

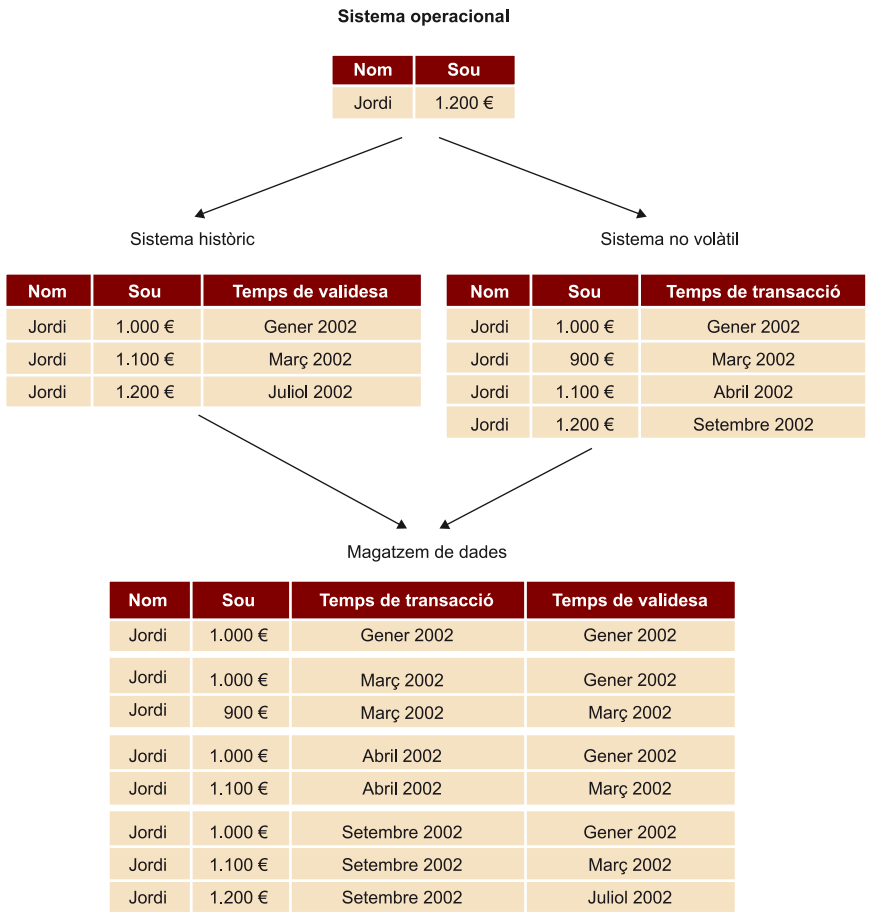
3.3. Informació històrica i no volàtil

Les dues darreres 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 ocorre un cert esdeveniment al món real (la historicitat). L'altre ens indica quan tenim constància d'aquest fet en la nostra base de dades (la no-volatilitat).

La historicitat és important per a analitzar com han evolucionat les coses, per a poder veure una pel·lícula en comptes 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 a poder saber si un cert informe es va fer tenint en compte unes dades o unes altres. La no-volatilitat implica que no hi hagi les operacions de modificar i esborrar pròpiament dites. Les dades no s'esborren o modifiquen, sinó que s'insereixen les correccions i la data en què s'han fet.

Exemple d'historicitat i no-volatilitat



Nom	Sou	Temps de transacció	Temps de validesa
Jordi	1.000 €	Gener 2002	Gener 2002
Jordi	1.000 €	Març 2002	Gener 2002
Jordi	900 €	Març 2002	Març 2002
Jordi	1.000 €	Abril 2002	Gener 2002
Jordi	1.100 €	Abril 2002	Març 2002
Jordi	1.000 €	Setembre 2002	Gener 2002
Jordi	1.100 €	Setembre 2002	Març 2002
Jordi	1.200 €	Setembre 2002	Juliol 2002

En aquesta figura, podem veure la diferència que hi ha si guardem el sou del Jordi en un tipus de sistema temporal o en un altre. Primer de tot, veiem que si desem el sou en un sistema relacional operacional (no temporal), només ens cal un tuple que conté el sou actual del Jordi. Si disposem d'un sistema històric, veiem com ha evolucionat el sou al llarg del temps i sabem en quin moment és vàlid cada valor. En canvi, si el sistema és simplement no volàtil, veiem quins són els nostres coneixements/creences segons les transaccions que s'hagin executat en cada moment. Per exemple, ens permet de saber que, al març, crèiem (erròniament) que el Jordi cobrava 900 euros (això no va coincidir amb la realitat en cap moment; per tant, un sistema històric no ho reflecteix). Finalment, veiem que en un magatzem de dades tenim tots dos tipus d'informació temporal. Per a veure l'interès d'això, fixem-nos en l'últim tuple. Ens indica que al setembre vam saber que el Jordi cobrava 1.200 euros des del juliol (cosa que no podem reflectir en cap dels sistemes anteriors). Noteu la diferència entre el volum de dades en un sistema operacional i en un magatzem de dades (en aquest cas, una tupla enfront de vuit tuples).

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

4. Objectius d'un magatzem de dades

En aquest apartat enumerarem diferents objectius que un magatzem de dades hauria d'assolir o complir, a mès de mencionar els objectius tant en l'àmbit empresarial com en el tècnic.

Els objectius més destacats són els següents:

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

Ajudar en la presa de decisions

Aquest seria l'objectiu principal del magatzem de dades, ja que aquest ha de ser un instrument útil per a donar suport en la presa de decisions. Aquest sistema s'ha de construir basant-se en els requeriments d'aquelles persones que l'han d'utilitzar.

> Quan es construeix un magatzem de dades, el punt més important que s'ha de tenir en compte és el d'obtenir la informació necessària sobre el negoci amb què posteriorment es puguin desenvolupar noves oportunitats que permetin de millorar els resultats de l'empresa.

El fet que se segmentin les dades, que aquestes estiguin disponibles amb més rapidesa, que es puguin fer anàlisis financeres "fàcilment" sense dependre dels informàtics i que estiguin depurades fa que la presa de decisions sigui més fàcil i que aquesta es faci amb més fonament.

Segmentar les dades de negoci

Tota organització necessita saber el seu posicionament dins el mercat, procurar liderar en el sector i projectar-se adequadament cap al futur.

Segmentar les dades de negoci permet fer particions segons un o més criteris.

Exemple de segmentació

Un exemple clàssic és saber quins dels productes de l'empresa són els que es venen més, en quins intervals temporals i en quina situació geogràfica.

La segmentació també serveix per a avisar on hi pot haver futurs problemes que se'ns escapen a primer cop d'ull.

Exemple empresa de venda de material elèctric

Imaginem que en una empresa de material elèctric la majoria de gent que hi va a comprar són lampistes. Si nosaltres segmentem les vendes per trams d'edat ens podem trobar que la majoria de lampistes tenen entre seixanta i seixanta-cinc anys. És clar que d'aquí a cinc anys el negoci haurà baixat molt si no fem alguna cosa.

Per tant, la segmentació ens permet de fer un estudi més profund de les dades i consegüentment establir comparatives que ens ajudaran a valorar-les d'una manera més correcta i a detectar tendències.

Gestionar el coneixement de l'empresa

També és possible veure el magatzem de dades com a contenidor de dades de l'empresa, les quals seran sotmeses a una anàlisi posterior.

Des d'aquest punt de vista, el magatzem de dades pot ser útil en els casos següents:

- Per a gestionar empreses amb sistemes distribuïts, ja que exerceix funcions d'integració de la informació.
- Per a gestionar empreses amb sistemes heterogenis, ja que exerceix funcions d'homogeneïtzació.
- Per a gestionar empreses amb sistemes centralitzats, perquè poden fer la selecció adequada al tema/usuari.
- Per a gestionar processos de fusió empresarial, si es fa servir amb la finalitat de consolidar.

Depurar les dades

Una de les mancances que hi ha quan dissenyem un magatzem de dades és la poca fiabilitat d'algunes de les dades que té l'empresa i la seva redundància, amb el consegüent perill que es generin problemes d'integritat en el magatzem de dades.

L'excusa que representa la creació d'un magatzem de dades passa per millorar, adequar i racionalitzar les dades que hi ha en el sistema operacional.

A la pràctica, aquest fet representa la millora d'alguns processos operacionals que en condicions normals no s'hauria pogut fer, però amb l'excusa de la implantació d'un magatzem de dades és possible fer-ho.

5. Comparativa: magatzem de dades 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-nos 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 involucri la inserció en una determinada taula d'una tupla que reflecteixi aquest fet.

Per tant, en cadascuna de les operacions (de manera general) s'involucren molt poques dades, però és cert que el volum global és 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 que el sistema gestor fes la transacció que reflecteixi un reintegrament per a poder 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 a això necessita analitzar l'evolució de l'índex de morositat de l'últim any dels seus clients, no ha de tenir en compte si un determinat client ha vingut aquest matí a fer moviments en el 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 detall.

Com es pot comprovar, la funció que fa cadascuna de les bases de dades en els exemples anteriors és ben diferent. En el primer cas es tracta d'una base de dades operacional i en el segon cas, 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 què 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 sobre la base d'aquesta informació.

Continuem amb els exemples

És clar, per tant, que els fets que la base de dades operativa té no són els que el director necessita. De tota manera, 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 directiva de les institucions o empreses ha de manipular d'una manera habitual i que ha causat que hagin aparegut en el mercat eines d'ajuda en la presa de decisions.

5.1. Diferències en l'emmagatzematge, disseny i estructuració de les dades

Temporalitat

Les dades s'han de guardar el temps que calgui. 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.

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.

El volum dels magatzems de dades

Als Estats Units quan es refereixen al volum dels magatzems de dades sempre parlen de *terabytes* de dades.

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 se solen donar diferents nivells. Aquest fet ens indica que algunes vegades tenim les dades implícitament duplicades.

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, dins aquest període, una sola vegada.

Estructuració

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

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

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

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

En el disseny de les bases de dades operacionals, ha de ser 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.

5.2. Diferències en el tractament de la informació

Explotació de la informació

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

En els magatzems de dades, les consultes solen ser imprevistes. N'hi pot haver de predefinides, però la varietat de possibilitats que hi ha fa impossible preveure quines seran les necessitats dels usuaris finals. A més, aquestes consultes estan orientades a àrees d'interès del negoci que sovint són canviants.

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 què s'actualitzen les dades. Per contra, en el cas dels magatzems de dades, aquest temps ha de ser ràpid, però no instantani, ja que el temps no és crític.

Temps de resposta

No és gens fàcil que les respostes a les peticions que es fan als magatzems de dades siguin ràpides quan parlem de terabytes de dades.

5.3. Diferències de funcionalitats

Activitats

L'activitat de les bases de dades operacionals és del dia a dia; en definitiva, és l'operativitat per al funcionament de l'empresa. Per tant, seran aplicacions fàcils de fer anar, no s'haurà de pensar gaire en les opcions que hi ha, i ràpides.

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

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 que és important és la dada actual, mentre que en el cas del magatzem de dades la importància està en les dades històriques.

Usuaris

En les bases de dades operacionals, els usuaris solen ser molts. 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, els usuaris són molt pocs, solen tenir accés a determinades dades agrupades i/o acumulades i solen ser a la part alta de l'empresa: direcció, màrqueting, planificació estratègica, control de gestió, etc.

5.4. 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, diversos factors condicionen l'evolució dels magatzems de dades:

a) Creixement exponencial de l'univers digital. Els usuaris i les xarxes de sensors dupliquen anualment les dades de les organitzacions, i aquestes sovint no són estructurades. Aquest creixement no solament planteja un repte pel que fa a l'emmagatzematge, sinó també a la gestió i manipulació de les dades. Ens referim, doncs, a un problema que té tres dimensions: 1) velocitat de generació de les dades, 2) volumetria de les dades i 3) variabilitat de les dades.

b) Noves tècniques de modelització. Daniel Linstedt va publicar l'any 2000 una nova tècnica anomenada *data vault*. L'objectiu d'aquesta tendència era la creació de magatzems de dades flexibles i auditables en temps real.

c) 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:

- Anàlisi contínua de dades[1]: mitjançant fluxos continus de dades, permet 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 a reduir el nombre d'incidències. Aquestes accions podrien consistir a avisar el personal de manteniment o canviar el comportament dels elements de la xarxa.

(1)En anglès, *data streaming*.

- Processament d'esdeveniments complexos[2]: 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 poden automatitzar ofertes dirigides a clients que segueixen un mateix patró.

(2)En anglès, *complex event processing*.

- Bases de dades en memòria[3]: mitjançant la memòria d'un servidor que utilitza tècniques OLAP, aquestes bases de dades permeten analitzar dades de gran volumetria en temps real. Sovint aquesta tecnologia dóna suport a les tecnologies anteriors.

(3)En anglès, *in-memory*.

- Hadoop, MapReduce i altres tecnologies equivalents: empreses com Google, Amazon o Facebook gestionen diàriament gran quantitat de dades que han de ser introduïdes en el sistema i consultades en temps real. Amb aquesta finalitat, sovint 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 solament utilitza el llenguatge SQL).

d) 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 a facilitar la creació d'aquest tipus d'avantatges competitives.

Resum

En aquest primer mòdul hem fet una introducció al concepte de magatzem de dades per a tenir els fonaments suficients en la resta de l'assignatura.

Primer, hem explicat què és un magatzem de dades. Com hem vist, no és un concepte nou, ja que implícitament s'estava fent servir encara que amb altres eines: els centres d'informació són els precursors del magatzem de dades.

Posteriorment, hem definit el magatzem de dades segons Inmon i n'hem repassat les característiques principals: orientació al tema, integració, no-volatilitat i dades històriques.

Hem vist que els magatzems de dades no són un altre tipus d'organització de bases de dades, sinó que donen 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.

Finalment, i per a remarcar la idea anterior, s'han comparat les bases de dades operacionals amb els magatzems de dades i hem vist que les diferències són realment molt importants.

Activitats

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

a) Expliqueu quins serien els objectius d'aquest projecte.

b) Quines dades creieu que serien rellevants per a aconseguir-lo?

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

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 sorprèn algun d'aquests? Per què?

c) Compartiu aquestes experiències en el 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 són molt ràpids.

a) Aquesta estructura ens serviria per a fer processos per a prendre decisions?

b) Si és que no, quines diferències caldria implementar per a construir un magatzem de dades?

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.

2.b) Des del punt de vista de 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.

Glossari

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

centre d'informació *m* Conjunt de fitxers i bases de dades precursor dels magatzems de dades que es basava en dades de l'entorn operacional per a treure informació i l'emmagatzemava per a usuaris i processos. La informació no estava compartida.

client/servidor *m* Entorn mitjançant el qual s'estableixen relacions entre agents per mitjà d'una xarxa de transmissió de dades, de manera que els agents clients reclamen serveis oferts per agents servidors.

data warehouse *Vegeu* magatzem de dades.

globalització *f* Procés que inclou aspectes generals de l'economia que afecta molt el desenvolupament dels magatzems de dades.

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 que tenen com a objectiu principal fer de suport en la presa de decisions.
en *data warehouse*

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*. EUA: Harvard Business School Press.

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

Gill, H. S.; Rao, P. C. (1996). *Data Warehousing. La integración para la mejor toma de decisiones*. Mèxic: Prentice Hall.

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