La construcció de la factoria d'informació corporativa

PID_00270637

Alberto Abelló Gamazo
Josep Curto Díaz
José Samos Jiménez
Juan Vidal Gil
David Díaz Arias

Temps mínim de dedicació recomanat: 7 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 a la UPC el programa de doctorat Erasmus Mundus IT4BI-DC. Els seus interessos d'investigació 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.

José Samos Jiménez

Doctor en Informàtica per la Universitat Politècnica de Catalunya. Professor titular del Departament de Llenguatges i Sistemes Informàtics de la Universidad 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.

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats per la professora: Àngels Rius Gavidia

Primera edició: febrer 2020

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

Índex

Introducció

La factoria d'informació corporativa (FIC) permet als analistes disposar de la informació que necessiten com a suport a la presa de decisions. Així i tot, generalment, la FIC no es pot considerar com un producte o una aplicació empaquetada que es pugui adquirir i, una vegada instal·lada a les nostres organitzacions, comenci a «fabricar» informació a partir de les dades de les fonts de dades operacionals.

Generalment, la FIC no es pot adquirir, s'ha de construir en les diferents organitzacions. Podem plantejar la construcció de la FIC segons diversos enfocaments, i també podem considerar variants en l'arquitectura de la FIC. En qualsevol cas, implementar aquesta arquitectura o les seves variants no és una operació trivial.

En aquest mòdul estudiarem diferents enfocaments alternatius per a la construcció de la FIC o les seves variants. Revisarem les implicacions que té començar la construcció des dels magatzems de dades departamentals i crear posteriorment el magatzem de dades corporatiu.

En l'execució de projectes veurem les dificultats que comporta crear la FIC en un únic projecte i l'alternativa de crear la FIC segons un conjunt de projectes autònoms. Analitzarem l'equip i perfils necessaris en la construcció de la FIC i l'evolució de l'entorn operacional una vegada construïda.

Posteriorment, s'abordarà la construcció del component d'integració i transformació, tenint en compte la construcció dels diferents components (transformació, depuració i integració) i posant l'accent en qüestions com la freqüència i la finestra d'actualització. Així mateix, s'analitzarà el paper de les eines de suport en la construcció d'aquest component.

Finalment, s'estudiarà detalladament la construcció dels diferents tipus de magatzem de dades: departamental, operacional i corporatiu. Es revisaran qüestions com ara el disseny del model de dades, la definició de la granularitat, l'organització de les dades, el seu aprovisionament i refrescament, així com l'addició de l'element temporal que permeti la historificació de les dades.

Objectius

En aquest mòdul es pretén oferir una visió global del procés de construcció de la FIC i de la construcció dels seus components. Mitjançant l'estudi, s'aconseguiran els objectius següents:

1. Entendre la problemàtica que suposa per a les organitzacions que no disposen de la FIC com a suport per a la presa de decisions i veure de quina manera es pot transformar un sistema operacional en un de decisional.

2. Conèixer els problemes que sorgeixen a l'hora d'intentar implementar variants de la FIC a organitzacions. Comprendre els avantatges i inconvenients dels diferents enfocaments.

3. Determinar com es pot estructurar el procés de construcció de la FIC en forma de projectes. Comprendre la dificultat del plantejament en un únic projecte i saber com implementar la FIC en un conjunt de projectes autònoms.

4. Tenir present l'evolució de l'entorn operacional en el desenvolupament de la FIC.

5. Comprendre els nous rols que apareixen en els equips de desenvolupament que intervenen en la construcció de la FIC.

6. Conèixer el procés de construcció del component d'integració i transformació, des de la construcció de cada component (transformació, depuració i integració) fins a la planificació dels processos d'actualització de dades.

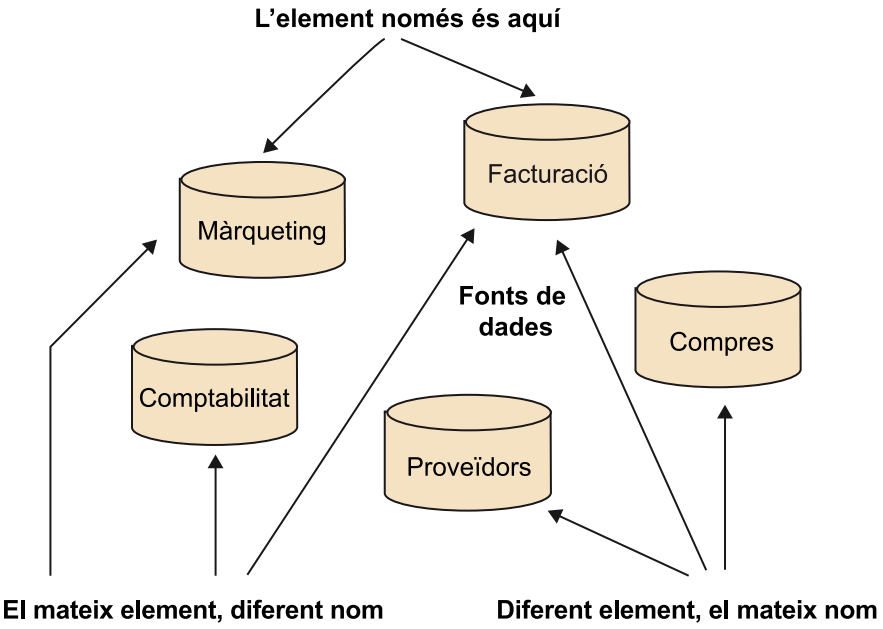
7. Conèixer com implementar els diferents tipus de magatzems de dades: departamental, operacional i corporatiu. Saber definir el model de dades, la seva organització, l'aprovisionament de dades i com afegir l'element temporal.

1. Transformació de dades des de l'entorn operacional al decisional

1.1. Suport a la presa de decisions des de l'entorn operacional

En el mòdul anterior, hem estudiat les característiques de l'entorn operacional. Sabem que les dades en un entorn operacional generalment estan orientades a les aplicacions o a la funcionalitat i desintegrades, a més de ser volàtils i no històriques.

Figura 1. Entorn operacional

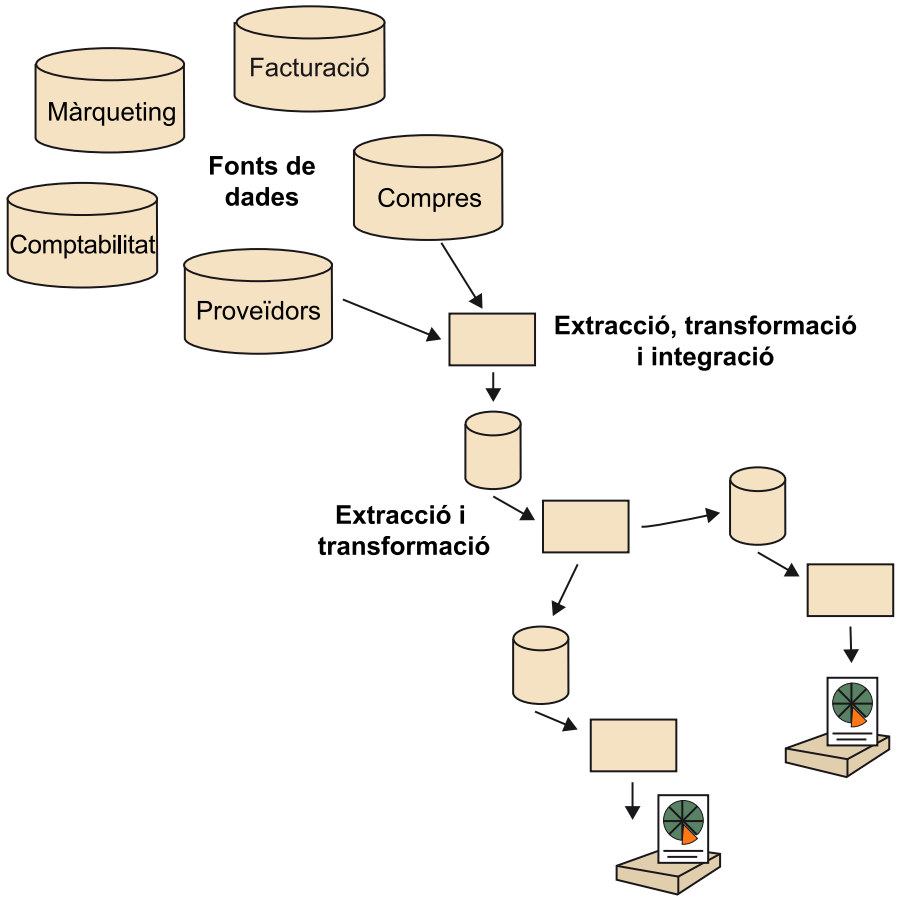


Cada aplicació operacional ofereix les funcionalitats operacionals per a les quals ha estat dissenyada. A més d'aquestes funcionalitats, o com a part d'elles, generalment també permet fer consultes sobre les seves dades i generar informes a partir d'aquestes dades, habitualment informes preestablerts. Aquests informes són els que poden usar els analistes com a suport en el procés en el qual es prenen decisions.

En alguns casos, la informació que s'obté en els informes preestablerts no és suficient i es requereix la generació de nous informes, possiblement amb la inclusió de dades de més d'una aplicació. Aleshores, la solució consisteix a desenvolupar un conjunt de programes d'extracció, transformació i integració de dades el resultat dels quals sigui l'informe desitjat. Generalment, aquests programes són desenvolupats pel Departament d'Informàtica de l'organització a petició dels analistes que faran servir la informació obtinguda.

Els programes desenvolupats transformen i integren les dades obtingudes de les bases de dades de les aplicacions i generen bases de dades intermèdies, o bé exclusivament transformen les dades de les aplicacions per adaptar-les a l'estructura requerida pels analistes. Per generar un informe, poden ser necessaris diferents passos de transformació i integració de dades. En la mesura en què es pugui, es tracta de reutilitzar el treball realitzat durant el desenvolupament d'informes previs, per exemple, accedint a alguna de les bases de dades intermèdies generades. D'aquesta manera, s'estalvia temps de desenvolupament de l'informe i d'execució per obtenir les dades requerides.

Figura 2. Obtenció d'informes en l'entorn operacional



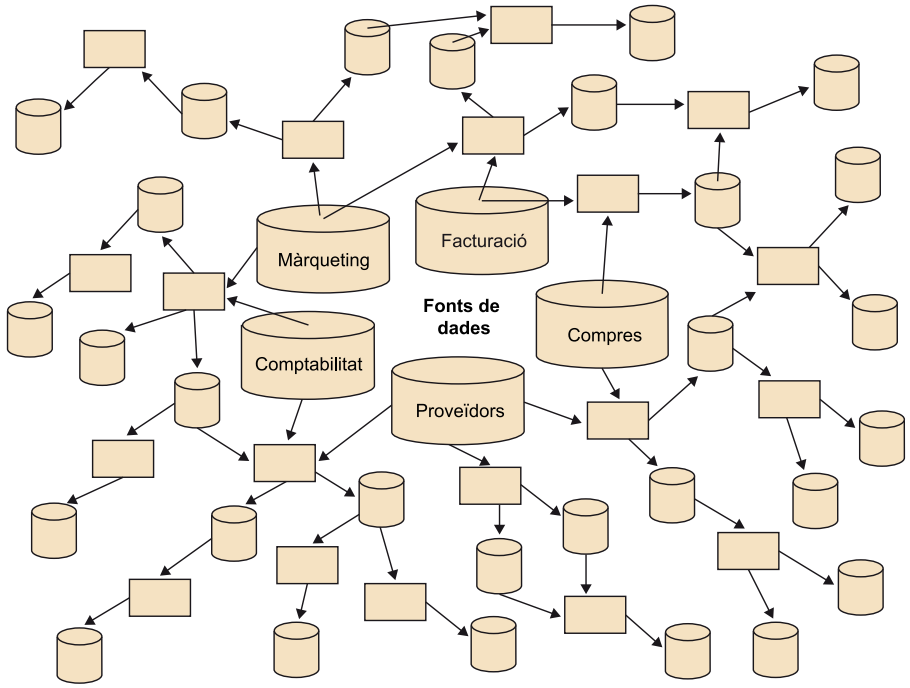
Els informes *ad hoc* que demanen els analistes es generen mitjançant programes desenvolupats a mida.

La necessitat d'informes a mida per part dels analistes no és un fet aïllat en les organitzacions, ja que les possibilitats ofertes pels sistemes operacionals en aquest sentit solen ser molt limitades. Això es produeix perquè l'objectiu principal d'aquests sistemes és el suport a l'operativa de l'empresa: les seves dades estan estructurades i organitzades amb aquesta missió. Així i tot, els analistes requereixen dades estructurades per al suport al procés de presa de decisions

que han de dur a terme, i aquestes dades es basen en d'altres registrades pels sistemes operacionals. Per tant, el desenvolupament de programes específics per a la generació d'informes a mida és bastant freqüent a les organitzacions.

Amb el pas del temps, la situació a la qual s'arriba en moltes organitzacions és similar a la mostrada a la figura 3. És a dir, partint de les fonts de dades dels sistemes operacionals, es construeix una estructura semblant a una teranyina composta per programes d'extracció, transformació i integració de dades, i també de bases de dades intermèdies que són utilitzades per altres programes per produir els informes requerits. Inmon denomina l'estructura resultant com l'entorn operacional de teranyina.

Figura 3. Teranyina d'un entorn operacional



1.2. Transformació de l'entorn operacional per satisfer les necessitats d'informació

L'entorn operacional de teranyina, resultat del desenvolupament d'informes a mida sol·licitats per part dels analistes, presenta greus inconvenients. El primer, que s'ha intentat reflectir visualment a la figura 3, és un problema de complexitat: el resultat del desenvolupament d'informes a petició constitueix un entorn complex, no planificat ni estructurat.

En el desenvolupament de qualsevol aplicació informàtica, complir els terminis d'entrega sol ser un objectiu molt important, que a vegades preval sobre d'altres, com, per exemple, el de generar la documentació necessària per al seu manteniment posterior. En el cas de les aplicacions desenvolupades per generar un informe, el termini de lliurament és especialment crític, ja que ha estat sol·licitat per un analista que el necessita com a suport per prendre una decisió, per la qual cosa és freqüent que no es duguin a terme les activitats de

documentació que serien desitjables. L'objectiu de cada equip de desenvolupament és generar l'informe requerit i per a això desenvolupa els programes necessaris, basant-se, en la mesura del possible, en els programes o bases de dades intermèdies prèviament desenvolupats. Tot i així, això últim no sempre és possible perquè no hi ha prou documentació ni control al respecte o simplement perquè és més ràpid i menys costós en temps de desenvolupament fer una cosa nova que modificar programes existents i que, en general, no estan prou documentats.

En relació amb això últim, tenim un problema de falta de productivitat: per a cada informe sol·licitat s'han de localitzar les dades necessàries (aquesta operació no és gens trivial, per la falta d'integració de les dades a les fonts de dades i a les bases de dades intermèdies) i desenvolupar els seus programes d'extracció, integració i transformació. Com que per fer aquestes operacions és molt difícil reutilitzar els esforços dedicats prèviament al desenvolupament d'altres informes, ens trobem en una situació en la qual, per desenvolupar cada informe, pràcticament hem de partir de zero.

Un dels problemes més greus d'aquest entorn és el de la falta de credibilitat (reflectit a la figura 3): a causa de la complexitat de l'entorn, no és improbable que un mateix informe (per exemple, els resultats del departament durant l'últim mes) s'hagi obtingut de dues maneres diferents, possiblement com a part d'un altre informe que conté informació addicional. Encara que s'hagi partit de les mateixes dades, es poden haver recorregut camins diferents per obtenir cadascun dels informes. El problema sorgeix quan la informació presentada en els dos informes no coincideix i genera desconcert als usuaris, la qual cosa provoca falta de confiança i credibilitat respecte al Departament d'Informàtica com a responsable dels informes generats.

No resulta estrany que es produeixi aquesta situació, ja que les dades de les fonts solen ser no integrades. N'hi ha prou que l'equip de desenvolupament d'un dels informes interpreti alguna dada de les fonts de dades o de les bases de dades intermèdies de manera diferent perquè es produeixin resultats com els que s'han descrit abans. La pèrdua de la confiança dels analistes en les dades amb les que treballen i en el departament que les ha generat és un problema molt greu, atès que és molt difícil de recuperar posteriorment.

L'entorn **operacional de teranyina** presenta problemes de complexitat, falta de productivitat i falta de credibilitat.

A part dels problemes descrits, la pregunta que ens hem de fer és la següent: pot satisfer les necessitats d'informació dels analistes l'entorn operacional de teranyina?

Respecte al temps requerit per obtenir la informació, el termini que transcorre des del moment en què l'analista fa la petició de l'informe fins que el rep segurament supera el que seria desitjable en un entorn tan canviant com en el que es mouen les organitzacions actualment.

D'altra banda, pel que fa al contingut dels informes, els analistes solen requerir l'evolució de diferents dades, i, per a això, requereixen realitzar una anàlisi de la informació històrica de l'organització. Tot i així, la informació dels sistemes operacionals generalment és informació no històrica, o conté una història molt limitada, ja que es requereix per a les operacions diàries de l'organització on generalment s'utilitzen dades de dates recents (informació no històrica).

L'entorn operacional de teranyina no satisfà els requeriments d'informació dels analistes.

Si l'entorn operacional presenta greus problemes i, a més, no satisfà les necessitats d'informació per part dels analistes, el podem transformar per corregir aquestes deficiències?

Respecte a la història de la informació emmagatzemada en els sistemes operacionals, una solució podria ser emmagatzemar la informació històrica requerida pels analistes. Això representaria un augment en la complexitat d'aquests sistemes. D'altra banda, si les aplicacions disponibles no emmagatzemen la història de les dades, haurien de modificar-se perquè ho fessin. Aquesta operació pot resultar molt costosa, sobretot en aplicacions antigues que hagin sofert canvis al llarg de la seva història, ja que, generalment, els canvis solen estar poc documentats i les modificacions requerides no són senzilles.

Per generar informes fiables més ràpidament i amb menys cost, s'hauria de reduir la complexitat de l'entorn operacional de teranyina. Per a això, principalment necessitaríem disposar d'un entorn que tingués les dades integrades. El problema és que les dades de l'entorn operacional, sobretot si aquest s'ha desenvolupat de manera gradual al llarg de la història o si s'han adquirit aplicacions empaquetades, solen ser no integrades. En aquest cas, també necessitaríem modificar totes les aplicacions per integrar-hi les dades.

Per tant, la situació és la següent: disposem d'un conjunt d'aplicacions, que formen l'entorn operacional i que satisfan les necessitats operacionals de l'organització, però que no satisfan les necessitats d'informació dels analistes. Per intentar corregir aquesta situació, necessitaríem modificar totes les aplicacions que hi ha per integrar les dades i emmagatzemar la seva història. Aquesta operació pot resultar inviable per la seva complexitat i cost.

D'altra banda, si ens fixem en les necessitats dels diferents tipus d'usuaris, els analistes necessiten fer consultes molt complexes (per exemple, evolució dels resultats del departament en els últims anys) i obtenir els resultats de manera immediata. Tot i així, els usuaris dels sistemes operacionals necessiten conèi-

xer de la manera més fidel possible la situació actual del sistema modelat (per exemple, situació actual del departament). Per tant, els sistemes operacionals s'han d'estructurar per reflectir tots els canvis que es produeixin i aquests canvis poden ser molt freqüents (per exemple, interessa que el saldo d'un compte estigui actualitzat en el moment en què es produeixi qualsevol canvi).

Les necessitats dels analistes i dels usuaris dels sistemes operacionals són contraposades: difícilment un sistema es pot dissenyar i configurar per ser òptim tant en la resposta a consultes complexes requerida pels analistes com en l'execució de modificacions sobre les dades en què es basen.

- Un mateix sistema no pot satisfer alhora les necessitats operacionals de l'organització i les d'informació dels analistes.

- Encara que fos possible fer-ho, els sistemes operacionals són molt costosos de modificar per integrar totes les dades i emmagatzemar la seva història.

- Per tant, els analistes de les organitzacions requereixen sistemes específics com a suport en el procés de presa de decisions.

Els nous sistemes requerits pels analistes necessiten obtenir les dades a partir dels sistemes operacionals existents i evitar modificacions. La solució consisteix a construir sistemes amb aquesta finalitat específica, que estiguin dissenyats per optimitzar el tipus d'operacions que necessiten fer els analistes i que funcionin en plataformes especialment configurades per a tal fi.

D'altra banda, seria convenient que els analistes disposessin de sistemes que els permetessin obtenir directament la informació requerida, en lloc d'aconseguir-la mitjançant peticions al Departament d'Informàtica. Això es pot aconseguir dissenyant els nous sistemes mitjançant un model de dades que estigui especialment orientat a aquesta finalitat, que és el model de dades multidimensional i que els dotarà de més autonomia.

Finalment i a efectes de concurrència d'usuaris pot ser molt diferent el volum d'usuaris i els moments d'accés a un sistema operacional que a un sistema orientat al treball dels analistes.

La solució per oferir suport a les necessitats d'informació dels analistes consisteix a incorporar el concepte de **magatzem de dades** a l'organització mitjançant la implementació de la FIC.

1.3. Diferències entre un entorn operacional i un d'informacional

Com s'ha comentat anteriorment, existeixen notables diferències entre un entorn operacional i un entorn que cobreix les necessitats d'informació dels analistes. Aquest segon entorn, que es basa en la FIC, és un entorn informacional. Un resum de les diferències és el següent:

1) Necessitats d'informació: els analistes de la FIC s'orienten a la consulta i anàlisi de dades, mentre que els usuaris de l'entorn operacional treballen més orientats a l'operativa i al dia a dia del negoci.

2) Ubicació de les dades: els entorns operacionals són heterogenis i la seva informació és de diferent naturalesa i distribuïda dins de l'organització, mentre que a la FIC l'orientació va dirigida a centralitzar informació de diferent naturalesa, però complementària.

3) Vigència de les dades: els entorns operacionals guarden una finestra temporal reduïda orientada a donar servei a les consultes i transaccions més recents en el temps, mentre que la FIC haurà d'assumir una finestra temporal més gran que suporti anàlisis sobre un rang temporal més ampli.

4) Tipus d'operacions que cal realitzar: actualització atòmica (pocs registres) de tipus transacció en l'entorn operacional (exemple, alta d'un client) i consulta i actualitzacions massives a la FIC (exemple, obtenir l'evolució de la cartera de clients en els últims sis mesos).

2. Estratègies en la construcció de la FIC

2.1. Diferents enfocaments en la construcció de la FIC

L'experiència dels projectes *Data Warehouse* demostra que, encara que el concepte de la FIC sigui un concepte amb una arquitectura ben definida, la seva implementació ha donat lloc a diferents enfocaments o variants, a vegades per simplificar, a vegades per falta de metodologia i a vegades amb l'objectiu d'aconseguir resultats tangibles de manera més ràpida.

La complexitat que pot arribar a tenir l'arquitectura de la FIC segons el negoci i la mida d'una organització, fa convenient conèixer diferents enfocaments d'abordatge d'aquesta mena de projectes. A continuació, coneixerem i analitzarem aquests enfocaments en la construcció de la FIC.

2.1.1. Enfocament basat en la construcció de magatzems de dades departamentals

La majoria dels analistes treballen directament sobre els magatzems de dades departamentals. És a dir, per als analistes d'informació, aquests són la «façana» de la FIC, el que veuen els usuaris. La resta dels components de la FIC proporciona les dades als magatzems de dades departamentals i freqüentment són invisibles per als seus usuaris. Per tant, és habitual que la percepció que tenen els usuaris de la FIC sigui, de manera exclusiva, l'oferta pels magatzems de dades departamentals.

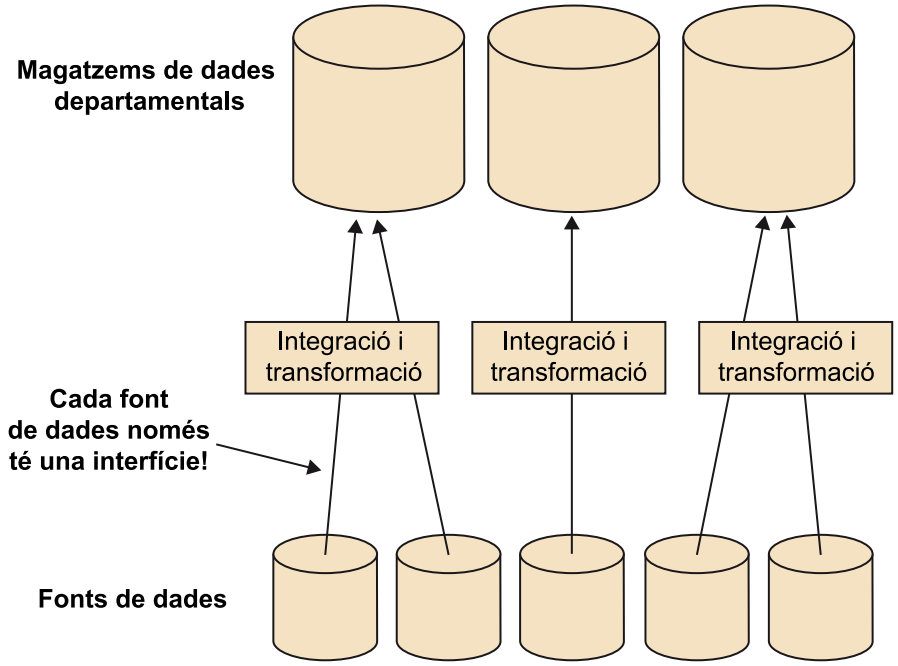
Aquesta percepció, a vegades, és compartida pels desenvolupadors, en alguns casos, per desconeixement. En molts llibres i altres materials de formació, així com en presentacions de venedors de diferents eines, l'únic component que s'estudia quan es parla de magatzems de dades és el que correspon als departamentals, i no es tenen en compte la resta dels components de la FIC.

En altres casos, encara que es conegui l'arquitectura de la FIC, s'ignora de manera conscient per perseguir els resultats immediats i, aparentment, construir només magatzems de dades departamentals és més barat, fàcil i ràpid que construir la FIC.

Això últim és cert quan només necessitem un magatzem de dades departamental o bé un nombre reduït d'aquestes dades. Generalment, aquesta és la manera d'incorporar el concepte de magatzem de dades a les organitzacions. Per començar, es planteja la construcció d'un magatzem de dades departamental perquè així podem obtenir resultats de manera immediata. Si partim de les dades de les fonts de dades, les transformem i integrem en un magatzem de

dades departamental dissenyat segons algun model multidimensional, que es caracteritza per ser un model totalment orientat a la consulta. És a dir, sens dubte, la manera més ràpida i barata de construir un primer magatzem de dades departamental consisteix a centrar-se de manera específica en la seva construcció, en lloc de construir prèviament la FIC. A més del magatzem de dades departamental, també haurem construït un component d'integració i transformació específic per obtenir les dades que emmagatzemem.

Figura 4. Magatzems departamentals amb component d'integració i transformació específic.



Una vegada tenim un magatzem de dades departamental, no és estrany que usuaris d'aquest departament o d'altres requereixin la construcció de nous magatzems de dades per satisfer les seves necessitats específiques d'informació. Com que els diferents departaments generalment treballen amb diferents dades, encara que puguin compartir-ne algunes, és freqüent que els nous projectes es desenvolupin de manera independent respecte als anteriors. Així doncs, per a cada projecte: es dissenya el magatzem de dades departamental i es construeix un component d'integració i transformació, segons els seus requeriments.

La construcció de magatzems de dades de manera independent és l'enfocament «natural» o intuïtiu que es duu a terme en moltes organitzacions, per desconeixement de la FIC o, coneixent-la, per intentar estalviar costos al començament. És un enfocament vàlid quan es tracta de construir magatzems de dades departamentals totalment independents.

El problema sorgeix quan els magatzems de dades per construir no són totalment independents; és a dir, quan hi ha dades comunes que han d'incloure's en diferents magatzems de dades departamentals. En aquestes situacions, ens trobem el següent:

1) Els diferents components d'integració i transformació treballen sobre les fonts de dades comunes: generem múltiples interfícies per a les mateixes dades. Això representa un problema de cost en temps de desenvolupament, manteniment i execució. Això últim es produeix perquè s'accedeix diverses vegades a les mateixes dades, una vegada per a cada magatzem de dades departamental que les utilitza. Hi ha fonts de dades origen compartides entre magatzems.

2) Cada magatzem de dades ha pogut fer la seva pròpia interpretació de les mateixes dades. Així doncs, tenim una falta d'integració de dades comunes en els diferents magatzems de dades departamentals.

Exemple de falta d'integració de les dades

Una dada denominada benefici es pot interpretar de manera diferent i tenir significat diferent en cadascun dels magatzems de dades departamentals en els quals es defineixi. En un es pot tractar del benefici abans d'impostos i en un altre, després d'impostos. Si comparem el benefici aconseguit en cadascun dels departaments dels respectius magatzems de dades, podem obtenir resultats no reals o erronis.

El problema de la falta d'integració s'accentua quan hi ha dades replicades entre les fonts de dades, situació que es produeix amb freqüència, i cada magatzem de dades departamental fa la seva integració. En aquest cas, a més de les diferents possibilitats d'interpretació de les dades de les fonts, s'han d'afegir les diferències que es puguin produir quan s'integren les dades de manera diferent a cada magatzem de dades departamental.

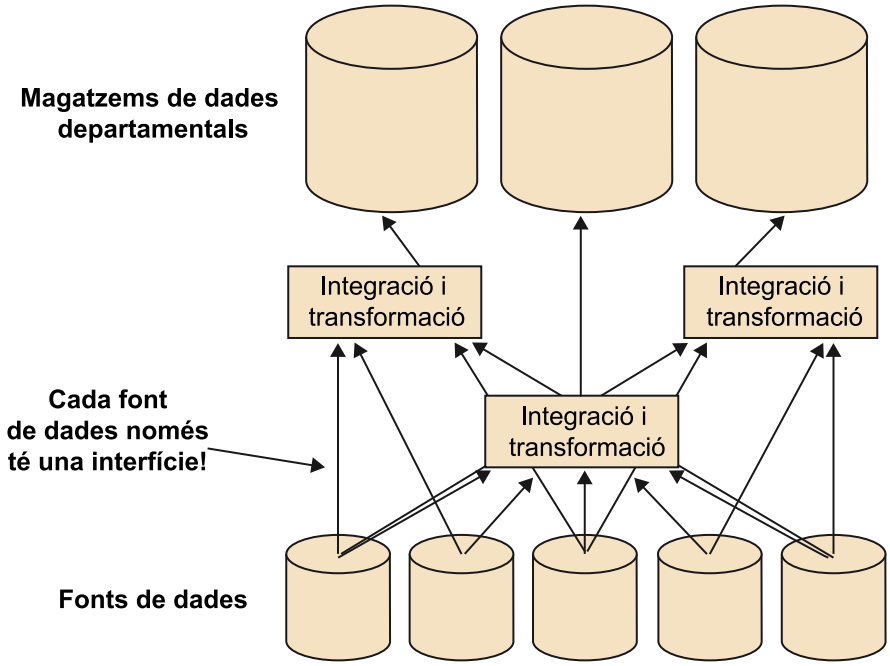
Exemple de diferents interpretacions en la integració de les dades de les fonts

En una organització disposem d'una aplicació de màrqueting amb dades de clients potencials (clients del passat, clients actuals i possibles clients) i una altra de facturació amb les dades dels clients actuals. En un magatzem de dades departamental que necessita treballar amb les dades dels clients, s'ha partit de les dades de clients en l'aplicació de màrqueting i s'han completat amb les dades disponibles en l'aplicació de facturació. Així i tot, en un altre magatzem de dades departamental s'ha procedit de manera inversa: s'han pres com a base les dades dels clients en l'aplicació de facturació i s'han completat amb les dades que hi ha a l'aplicació de màrqueting. El resultat global en tots dos casos no ha de coincidir; de fet, és més probable que les bases de dades de clients resultants en els dos magatzems de dades departamentals siguin molt diferents.

El problema de fonts de dades comunes entre magatzems departamentals és més habitual del que puguem estimar, ja que hi ha moltes fonts que seran crítiques per analitzar el nostre negoci i apareixeran en els diferents magatzems departamentals. Un exemple poden ser els clients, els productes, els proveïdors, els comptes corrents, etc.

Per les raons esmentades anteriorment, en moltes organitzacions el problema de construcció de la FIC es redueix al de construcció de manera independent d'un conjunt de magatzems de dades departamentals, malgrat que hi hagi dependències entre ells. En aquests casos, l'arquitectura que s'obté com a resultat és la que es mostra a la figura 5.

Figura 5. Fonts de dades comunes entre magatzems departamentals



Aquesta arquitectura pot resultar vàlida quan es comença a implantar el concepte de magatzem de dades a l'organització, quan només es disposa d'uns quants magatzems de dades departamentals i aquests realment són independents entre ells. No obstant això, l'habitual és que els magatzems departamentals tinguin dependències entre si i que en creixi el número amb el temps. Què podem fer en aquesta situació?

Els problemes d'integració i de multiplicitat d'interfícies sobre les fonts de dades se solucionen mitjançant la construcció del magatzem de dades corporatiu.

2.1.2. Construcció del magatzem de dades corporatiu *a posteriori*

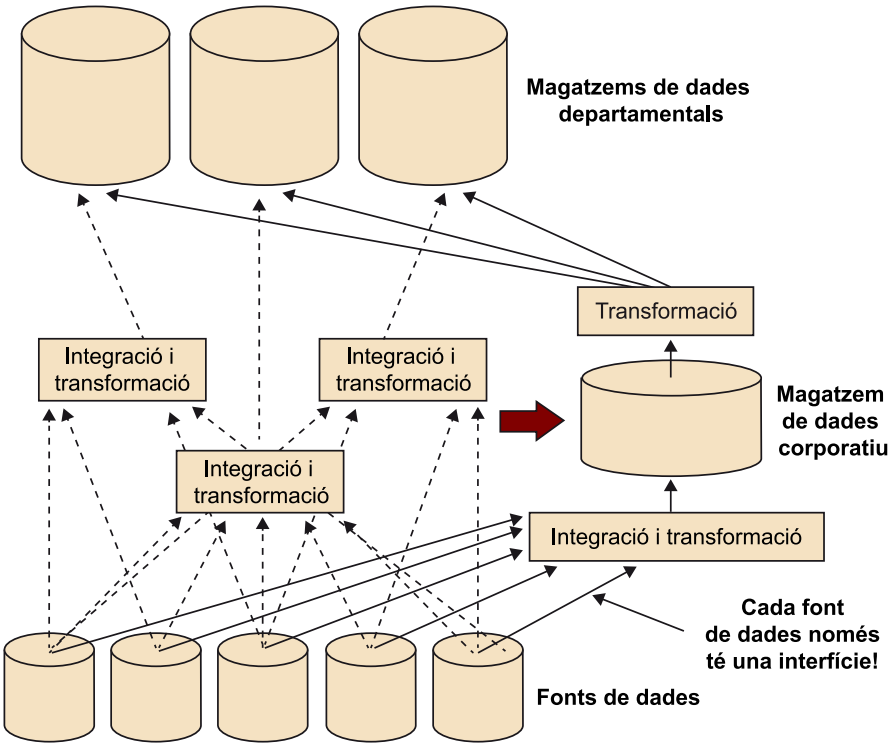
L'objectiu de construir un magatzem de dades corporatiu és tenir un repositori centralitzat i integrat de la informació de la companyia. En un magatzem de dades departamental ens centrem en la seva consulta, però en un magatzem de dades corporatiu és un punt crític el fet de tenir un emmagatzematge centralitzat i integrat. Per a la construcció del magatzem de dades corporatiu, orientat a l'emmagatzematge de les dades més que no pas a la seva consulta, es fa el següent:

1) Es redueix el nombre d'interfícies a les fonts de dades: només és necessària una interfície per a cada font per portar les dades de la font d'origen al magatzem de dades corporatiu.

2) Les dades estan integrades en el magatzem de dades corporatiu, la qual cosa evita tenir-ne múltiples interpretacions.

Si partim de zero, la solució consisteix a construir el magatzem de dades corporatiu abans o durant la construcció dels diferents magatzems de dades departamentals. Tot i així, si ja tenim diferents magatzems de dades departamentals construïts, hem de transformar l'arquitectura basada exclusivament en aquests per incloure el magatzem de dades corporatiu.

Figura 6. Magatzem de dades corporatiu i magatzems de dades departamentals



El problema per incloure el magatzem de dades corporatiu quan ja tenim construïts diferents magatzems de dades departamentals rau en el fet que hem de realitzar un procés de reenginyeria sobre els diferents components d'integració i transformació, així com sobre els magatzems de dades departamentals prèviament, per adaptar-los a la nova arquitectura. Aquesta transformació pot resultar molt complexa i costosa.

Particularment, pel que fa a la transformació dels magatzems de dades departamentals, el principal problema que es presenta és el de la integració de les dades: si s'han construït d'una manera no integrada, se les ha de transformar, i, en alguns casos, aquesta transformació en l'àmbit tècnic presenta problemes

de gestió i no és gens trivial. A més, tenim el problema afegit que els canvis fets sobre els magatzems de dades departamentals són visibles per als usuaris finals i afecten directament el seu treball.

Exemple de correcció per falta d'integració de dades en els magatzems de dades departamentals

Si la dada «benefici» de l'exemple utilitzat anteriorment en un o diferents magatzems de dades departamentals s'interpreta com a «benefici abans d'impostos» i la interpretació acceptada en l'àmbit corporatiu és del «benefici després d'impostos», la nova interpretació s'ha d'incorporar als magatzems de dades departamentals que no la consideressin. Per tant, els usuaris han d'adaptar-se al nou significat a l'hora de generar els informes que necessitin.

A causa dels problemes que sorgeixen a l'hora de construir el magatzem de dades corporatiu després d'haver construït els magatzems de dades departamentals, es recomana construir-lo prèviament o al mateix temps. És a dir, és més adequat planificar la construcció de la FIC des d'un principi.

Un plantejament intermedi entre la construcció de magatzems de dades departamentals independents i la seva construcció de manera conjunta amb la FIC és la creació de magatzems departamentals amb dimensions conformades.

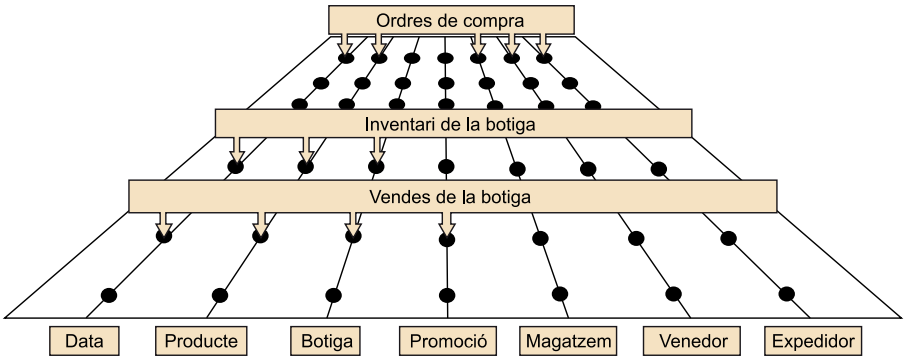
Una **dimensió conformada** és una entitat del model de dades compartit per diversos magatzems de dades. Per exemple, l'entitat Client o l'entitat Producte poden ser compartides pel magatzem departamental de màrqueting, el d'operacions i el de finances, ja que tots han de realitzar les seves anàlisis a partir dels mateixos productes i clients.

És important tenir una única entitat per a tots els magatzems departamentals que asseguri una única versió i un únic procés de transformació i càrrega que l'actualitzi. Aquest plantejament va ser proposat per Ralph Kimball dins de l'arquitectura coneguda com a *Enterprise bus matrix* en lloc de basar-la en la FIC com planteja Inmon.

La identificació de dimensions conformades implica una visió global de l'organització, fins i tot en el cas d'estar creant un magatzem de dades departamental. És necessari analitzar quins processos de negoci de la nostra organització poden ser analitzats utilitzant les entitats que creem en el nostre magatzem departamental. Per exemple, en el magatzem departamental financer utilitzarem l'entitat Producte, però aquesta entitat s'aplica també a l'anàlisi d'altres processos de negoci, com pot ser el procés de vendes *online* o els processos de fabricació, propis d'altres departaments. Així mateix, a l'hora de crear un nou magatzem departamental quan ja n'existeixen d'altres, convé revisar l'existència de dimensions conformades en altres magatzems departamentals, amb la finalitat de reutilitzar-les al nou magatzem.

A la figura 7 es representa aquesta arquitectura (*entreprise datawarehouse bus matrix*) proposada per Kimball, que difereix notablement de la FIC proposada per Inmon. La proposta de Kimball es recolza en la creació de magatzems de dades departamentals que posteriorment queden connectats per les dimensions comunes, mentre que la proposta d'Inmon planteja la creació conjunta dels magatzems departamentals amb la FIC. En aquesta assignatura ens centrarem en la construcció basada en la FIC d'Inmon, encara que en diferents moments es farà referència al plantejament de Kimball.

Figura 7. *Enterprise bus matrix*



Font: www.kimballgroup.com

2.1.3. Combinació del magatzem de dades operacional i el magatzem de dades corporatiu

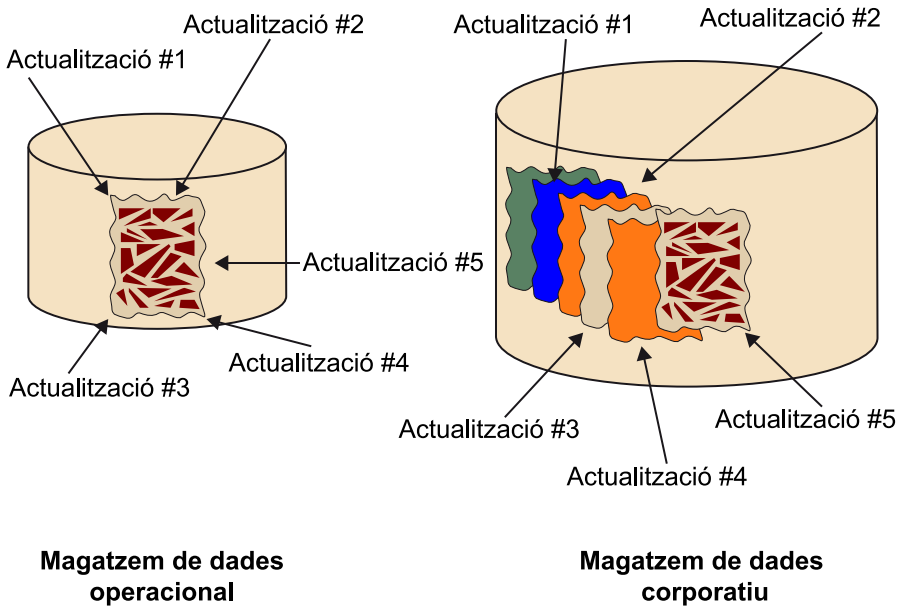
El magatzem de dades operacional i el magatzem de dades corporatiu tenen característiques semblants: tots dos estan orientats al tema i integrats. Es diferencien en el fet que les dades del magatzem de dades operacional són volàtils i no històriques, mentre que les dades del magatzem de dades corporatiu són no volàtils i històriques.

El magatzem de dades operacional emmagatzema una imatge actualitzada de les dades integrades de l'organització. El magatzem de dades corporatiu emmagatzema una pel·lícula formada a partir de les diferents imatges de les dades de l'organització, és a dir, la informació està historiada.

Exemple de magatzem de dades operacional i corporatiu en un operador de telecomunicacions

Per exemple, en un magatzem operacional d'un operador de telecomunicacions interessa saber la tarifa actual d'un client, però en un magatzem de dades corporatiu, interessa saber les diferents tarifes que ha tingut un client des de la seva alta a la companyia i el període de temps durant el qual ha tingut cada tarifa. Al magatzem de dades corporatiu s'analitzarà la història del client a la companyia.

Figura 8. Magatzem de dades operacional i corporatiu



Podem combinar la construcció dels dos magatzems de dades en una única estructura? En teoria sí, però a la pràctica no és convenient. La raó principal per no combinar-los és que han d'estar dissenyats per suportar diferents tipus d'operacions, que realitzen diferents tipus d'usuaris.

L'objectiu principal del magatzem de dades operacional és mantenir una visió integrada i totalment actualitzada de les dades operacionals. Sobre aquests, s'executaran moltes operacions d'actualització i consultes simples que seran dutes a terme principalment per oficinistes. El seu disseny i configuració estaran orientats per realitzar aquest tipus d'operacions d'una manera òptima.

D'altra banda, el magatzem de dades corporatiu no requereix que les seves dades estiguin totalment actualitzades: n'hi ha prou que estiguin actualitzades segons les necessitats dels analistes, que són els seus usuaris (en alguns casos, n'hi haurà prou amb què s'actualitzin de manera periòdica: setmanalment o mensualment). Aquest magatzem està especialment dissenyat per a què les actualitzacions s'emmagatzemin com a imatges noves, de manera que s'anirà formant una pel·lícula de les dades. Està doncs configurat per optimitzar les consultes de les imatges de les dades, no per fer modificacions sobre aquestes.

A vegades, ens podem trobar que els magatzems de dades operacionals tenen un nivell de granularitat més detallat que el magatzem de dades corporatiu, ja que aquest utilitza informació agregada.

Si combinem les dues estructures, el resultat serà una base de dades amb molts registres (hem de guardar la història emmagatzemada al magatzem de dades corporatiu), que ha d'estar configurada per fer modificacions sobre les dades (una cosa requerida pel magatzem de dades operacional). D'aquesta manera, qualsevol transacció serà molt costosa, ja que es poden combinar consultes

complexes amb actualitzacions, i el volum de dades que han de tractar les diferents transaccions serà molt gran perquè haurà de considerar les dades del magatzem de dades corporatiu.

El magatzem de dades operacional i el magatzem de dades corporatiu tenen objectius diferents i estan dissenyats per aconseguir-los de manera òptima. Si els combinem en una estructura comuna, es degradarà el temps de resposta per a les dues funcionalitats.

2.1.4. La FIC sense el magatzem de dades operacional

El magatzem de dades corporatiu conté totes les dades que necessiten els analistes, que s'obtenen directament a partir de les fonts de dades operacionals i també a partir del magatzem de dades operacional, que al seu torn les obté de les fonts de dades operacionals tal com es pot veure a la figura 9. El procés d'integració i transformació de les dades de les fonts operacionals per incloure-les al magatzem de dades corporatiu o al magatzem de dades operacional és comú a tots dos, i per aquest motiu n'hi ha prou amb transformar les dades del magatzem de dades operacional per adaptar-les a les estructures del magatzem de dades corporatiu.

Figura 9. FIC amb magatzem de dades operacional

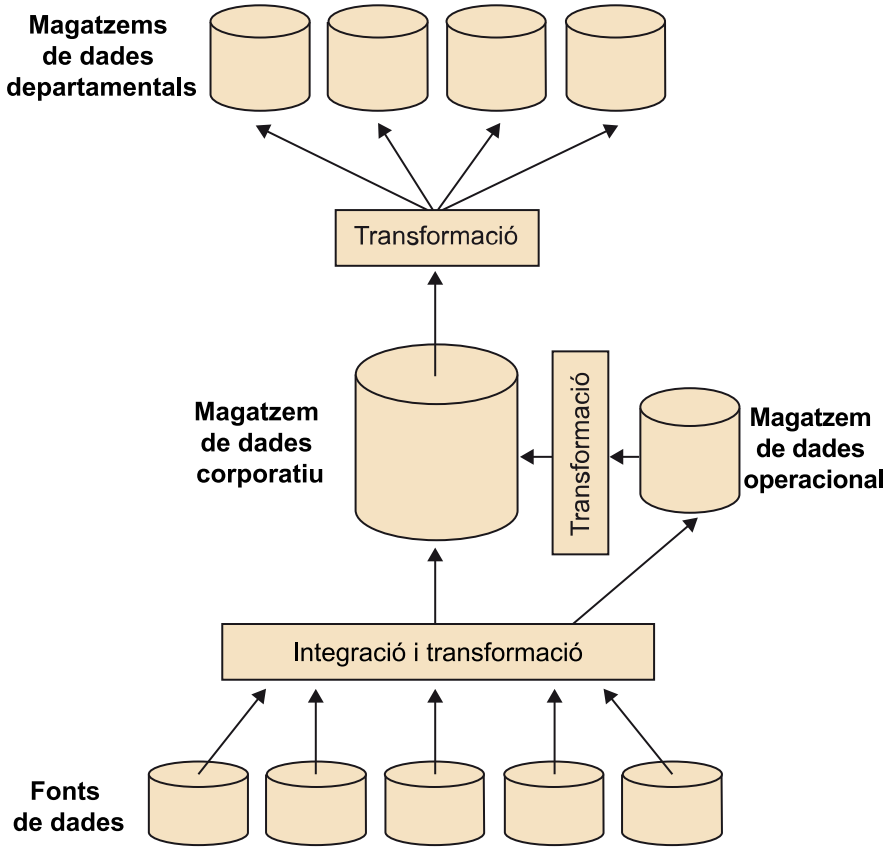
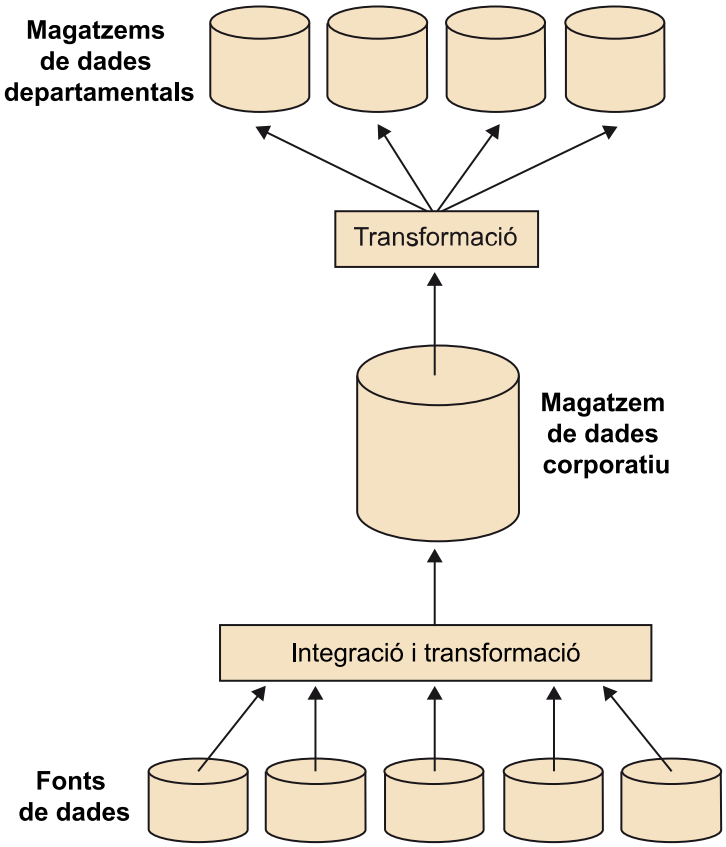


Figura 10. FIC sense magatzem de dades operacional



A diferència del magatzem de dades corporatiu, el magatzem de dades operacional no és estrictament necessari a la FIC. És a dir, tot i que convé disposar del mateix per les funcionalitats que proporciona als usuaris, i també per permetre construir més fàcilment el magatzem de dades corporatiu, no es necessita de manera estricta per construir el magatzem de dades corporatiu ni tampoc per construir els magatzems de dades departamentals, que són els que proporcionaran la funcionalitat principal als analistes (vegeu la figura 10).

Les dues arquitectures representades a les figures 9 i 10 també són vàlides, encara que l'arquitectura de la figura 10 ofereix menys funcionalitats que la de la figura 9.

El **magatzem de dades operacional** és una estructura opcional a la FIC, per proporcionar informació als analistes.

Malgrat que el magatzem de dades operacional és opcional a la FIC, en algunes organitzacions serà més necessari que en d'altres, depenent de diferents factors, entre els quals destaquen els següents:

- La mida de l'organització: com més gran sigui l'organització, més probable és que necessiti el magatzem de dades operacional. A les organitzacions

petites, els problemes causats per la falta d'integració de les dades operacionals solen ser menors.

- La naturalesa dels negocis: si l'organització necessita accedir de manera immediata a informació integrada i actualitzada, per exemple, perquè interacciona directament amb els clients o perquè la necessita en el procés de fabricació, és molt probable que necessiti el magatzem de dades operacional.

- També pot afectar el fet que la companyia tingui una orientació analítica important i en els diferents nivells de l'organització (també el més operatiu) s'apliqui l'anàlisi i es necessiti donar suport a decisions operatives.

- Finalment, depèn de la quantitat de les aplicacions operacionals i del grau d'integració que hi ha entre elles. En una organització amb un conjunt petit d'aplicacions operacionals que estan molt integrades, el magatzem de dades operacional serà menys necessari que en una altra amb gran quantitat d'aplicacions poc integrades.

2.1.5. La FIC amb *Staging Area*

Una *Staging Area* o àrea de maniobres és una zona de treball temporal de la FIC, situada entre les fonts de dades dels sistemes operacionals i el *Data Warehouse*.

Malgrat ser una peça opcional de la FIC, hi ha escenaris en els quals és realment útil plantejar-se l'ús d'aquest espai temporal. És habitual implementar la *Staging Area* amb una base de dades, tot i que també es podria utilitzar un conjunt d'arxius temporals com a zona de treball.

L'ús d'una *Staging Area* simplifica el procés d'extracció del component d'integració de la FIC, especialment quan és necessari realitzar càlculs intermedis per reduir la complexitat de la dada, homogeneïtzar-la des de múltiples orígens heterogenis o per realitzar processos de neteja de les dades que millorin la seva qualitat.

Un altre avantatge del seu ús és la reducció de l'impacte sobre els sistemes operacionals que pot provocar, quan s'executen tasques ETL (*Extract Transform Load*) pesades. Per exemple, quan, des dels sistemes de *Business Intelligence*, ens connectem als sistemes operacionals per carregar o actualitzar la informació del *Data Warehouse*, sent perceptible fins i tot pels usuaris finals.

Aquest problema es pot veure agreujat quan no sigui possible realitzar les càrregues en horaris de baixa activitat, o quan, per realitzar la càrrega de l'*Staging Area*, sigui necessari realitzar càlculs previs sobre les dades del sistema operacional.

En aquest escenari, l'*Staging Area* ens permet fer una càrrega «en brut» el més ràpid possible de les dades del sistema operacional cap a l'àrea de maniobres. Això redueix dràsticament el temps de connexió amb el sistema operacional i, per tant, l'impacte negatiu en el seu rendiment. Posteriorment, tots els càlculs necessaris abans d'enviar la dada al *Data Warehouse* i tots els processos de neteja de la dada es realitzaran sobre la còpia de l'*Staging Area*, totalment desconnectada de la font de dades original.

Finalment, un altre avantatge important de l'*Staging Area* és la tolerància a errors que aporta al component d'integració. Si, per algun motiu, es genera un error en els processos de transformació o càrrega, podríem reiniciar-los, sense necessitat de repetir el procés d'extracció i, per tant, sense afectar de nou als sistemes operacionals.

Un aspecte negatiu de l'ús de l'*Staging Area* és l'increment del temps total d'execució dels ETL, ja que inevitablement, estarem afegint més processos intermedis. No obstant això, els beneficis globals que aporta gairebé sempre compensen aquest cost en temps d'execució.

2.2. Construcció de la FIC mitjançant un sol projecte

Si una organització reconeix la necessitat de la FIC, generalment considera que necessita tots els seus components i, a més, que els necessita de manera immediata, la tendència habitual és plantejar la construcció de la FIC mitjançant un sol projecte. El problema principal que planteja aquest projecte és la complexitat.

Generalment, en aquest entorn, els analistes encara no coneixen clarament les funcionalitats que els poden oferir els magatzems de dades departamentals. La idea inicial que tenen sobre les seves necessitats d'informació canvia quan descobreixen les possibilitats que els ofereixen els magatzems de dades. En aquestes condicions, difícilment poden transmetre els seus requeriments als desenvolupadors de la FIC. L'abast funcional de la FIC és complicat de delimitar. D'altra banda, localitzar les dades que es necessiten en les aplicacions operacionals no és una tasca fàcil. Si a més es pretén desenvolupar tot el conjunt de la FIC en un sol projecte, aquest serà de dimensions i complexitat massa grans en comparació amb els que se solen desenvolupar a les organitzacions.

D'altra banda, precisament a causa de la complexitat del projecte, es fa molt difícil justificar el cost de desenvolupament de la FIC segons el benefici que aporta a l'organització, ja que els costos són molt grans i el benefici, tot i que és clar, no s'ha avaluat prou.

El risc de projecte pel que fa a terminis i costos és alt. En un projecte d'aquesta envergadura, tindrem un abast de projecte molt difícil de determinar amb exactitud. D'altra banda, no és senzill gestionar les expectatives i la presa de requisits d'un alt nombre d'analistes de diferents departaments involucrats,

on les visions i expectatives poden ser molt dispars. Així mateix, també existeix el risc del canvi de requeriments amb el temps, situació habitual en un projecte de gran abast al qual els canvis en el negoci de la companyia poden afectar directament.

El resultat de plantejar el desenvolupament de la FIC com un sol projecte és un projecte molt gran i complex, els objectius del qual no estan totalment clars, els requeriments poden canviar, i els seus costos són difícilment justificables pel que fa a l'organització. En aquesta situació, és més probable que sigui un fracàs.

2.3. Construcció de la FIC mitjançant projectes autònoms

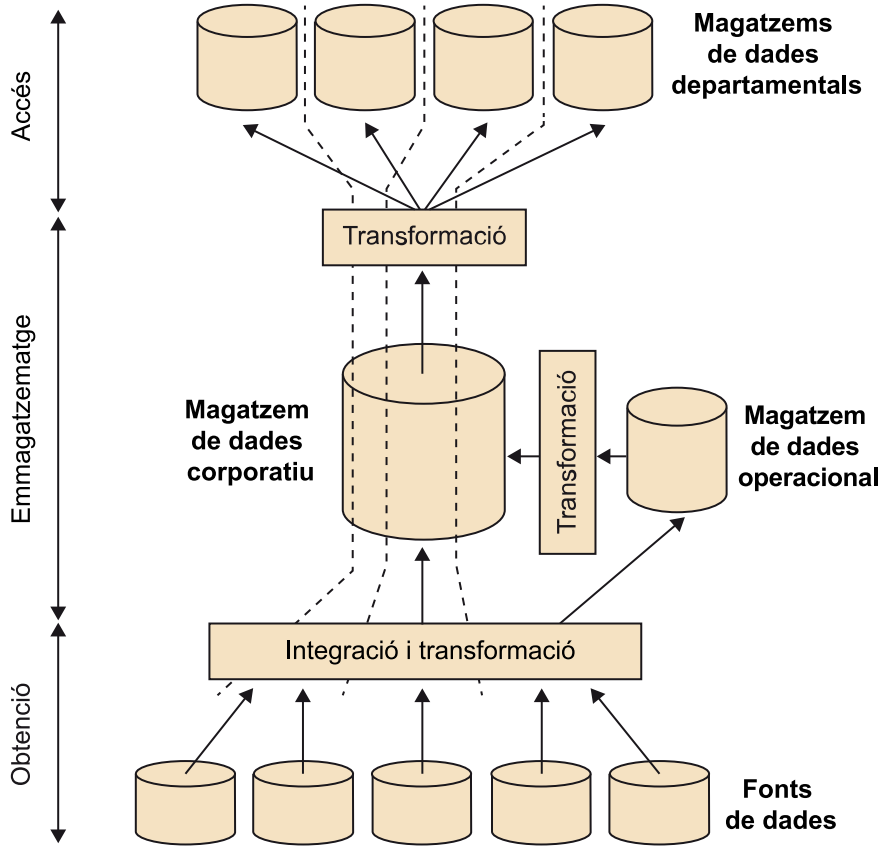
En apartats anteriors, hem vist que intentar construir la FIC com un sol projecte és massa complex. D'altra banda, si dividim l'arquitectura de la FIC horitzontalment, segons la funcionalitat dels components, no es redueix la complexitat tant com seria desitjable i, a més, és difícil satisfer els requeriments dels analistes d'informació.

Per tant, una altra possibilitat consisteix a dividir l'arquitectura de la FIC de manera vertical, és a dir, dividir la construcció de la FIC en forma de projectes, de manera que facin el següent:

- Que proporcionin un valor per si mateixos a l'organització perquè el seu cost sigui justificable.

- Que siguin complets i autònoms en la mesura en què es pugui; és a dir, que no necessitin altres projectes per entrar en producció.

Figura 11. Divisió de la FIC en projectes autònoms



A la figura 11, es mostra una representació de l'estructura dels anomenats projectes. És a dir, a cada projecte s'hi han de construir els components següents:

- El component d'accés: el magatzem de dades departamental amb les dades que necessiten els analistes d'aquest projecte o bé l'eina d'accés al magatzem de dades corporatiu o al magatzem de dades operacional.

- El component d'emmagatzematge: la part del magatzem de dades corporatiu i/o magatzem de dades operacional que conté les dades a les quals es necessita accedir, i que encara no ha estat construïda per cap projecte anterior.

- El component d'obtenció de les dades: el component d'integració i transformació que aconsegueix les dades necessàries per al projecte, a partir de les fonts de dades origen. Es considera que un dels problemes més importants que hi ha a la construcció de la FIC és el de l'obtenció de les dades, és a dir, la localització de les dades necessàries i la construcció del component d'integració i transformació. Aquest és un dels motius més freqüents de fracàs dels projectes de construcció de la FIC.

Cada projecte consta dels components d'accés, emmagatzematge i obtenció de les dades provinents de les fonts de dades (és una pràctica habitual que cada component estigui desenvolupat per persones o equips diferents). Es tracta d'un projecte complet que pot entrar en funcionament de manera independent de la resta dels altres projectes existents.

D'aquesta manera, cada projecte intenta satisfer els requeriments d'un grup d'analistes de manera independent, de manera que sigui justificable el cost del projecte segons els beneficis que aporta i es redueixi la complexitat del desenvolupament de la FIC.

Sota aquest plantejament de divisió vertical, en els apartats següents analitzarem els projectes en els quals es dividiria la construcció de la FIC.

2.3.1. El primer projecte: projecte global de desenvolupament

En un primer moment, la visió que es té de la FIC és dispersa. Coneixem la seva arquitectura i, a més, la seva aportació com a suport d'informació per als analistes és positiva per a l'organització. Així i tot, no coneixem encara amb detall els magatzems de dades departamentals que es necessiten en l'organització, ni el que li aportaran.

El primer objectiu és dividir la construcció de la FIC en projectes, de manera que cada projecte:

- Correspongui a un problema concret.
- Tingui a un responsable en l'organització: l'analista que utilitzarà el sistema com a resultat del projecte.
- Ofereixi un benefici tangible a l'organització: podem conèixer quin serà el cost del projecte i el benefici que aquest aportarà a l'organització.

Generalment, cada projecte definit es correspondrà amb un magatzem de dades departamental que s'ha de desenvolupar. El seu responsable serà l'analista (o un dels analistes) que utilitzarà el magatzem de dades departamental, i el benefici es calcularà segons els objectius que es pretenguin aconseguir amb el magatzem de dades desenvolupat i el seu cost estimat de desenvolupament.

Per definir els objectius del projecte, haurem de mantenir reunions amb els analistes d'informació, que seran els seus usuaris i responsables del projecte. Així i tot, no n'hi haurà prou amb les reunions amb els usuaris. De manera addicional, perquè la definició del projecte sigui realista, ens haurem d'assegurar que les dades requerides pels usuaris estan a l'organització, a les fonts de dades

operacionals, i que tenen el grau de detall i la qualitat requerits. És a dir, haurem de fer una revisió global de tots els components de cada projecte: obtenir, emmagatzemar i accedir a les dades requerides pels usuaris.

És a dir, l'objectiu del primer projecte és transformar la visió dispersa que es té de la FIC en una visió concreta en forma de projectes, de manera que cadascun d'ells satisfaci uns objectius determinats i aporti un valor concret a l'organització. Aquesta visió que conformen tots els projectes de la FIC ha de ser el més global possible, de manera que cada projecte individual tingui en compte el seu impacte en la resta dels projectes i viceversa, l'impacte de la resta dels projectes en el mateix projecte. En el desenvolupament d'aquesta mena de projectes sol ser útil afegir a la visió departamental la visió de processos de negoci, atès que un procés de negoci serà transversal a diversos departaments. El conjunt dels magatzems departamentals per desenvolupar a la FIC haurà de donar una visió integrada i completa dels processos de negoci.

Exemple de procés de negoci: facturació

La facturació és un procés de negoci que afecta diversos departaments com finances, control de gestió, vendes. Cadascun d'aquests departaments podrà tenir el seu propi magatzem de dades departamental de la FIC, però si ens centrem en la informació de facturació, encara que cada magatzem ofereixi una perspectiva diferent, la visió global ha de donar una informació de facturació completa i íntegra.

Aquest primer projecte és el més important en el desenvolupament de la FIC, ja que la resta dels projectes en depenen. Així i tot, generalment resulta molt difícil justificar-ne l'execució a l'organització: cal plantejar un projecte d'estudi que analitzi tota l'organització i el benefici de la qual estigui reflectit exclusivament per l'èxit dels futurs projectes que es desenvolupin. Disposar d'un patrocinador amb prou nivell a l'organització facilita el desenvolupament del primer projecte i de la FIC.

Els beneficis del projecte global de desenvolupament no són immediats i, per aquest motiu, és més fàcil de justificar i dur-lo a terme, si disposem d'un patrocinador a l'organització que ofereixi suport al desenvolupament de la FIC.

2.3.2. Desenvolupament de projectes autònoms

Mitjançant el desenvolupament del primer projecte, el projecte global de desenvolupament, hem dividit el desenvolupament de la FIC en forma de projectes.

Habitualment, són projectes que es completen en uns sis mesos, encara que posteriorment és possible fer-hi ampliacions.

Tot i que no és fàcil d'aconseguir, és molt important que els projectes siguin autònoms; és a dir, que cada projecte tingui perfectament delimitats els seus objectius i no depengui del desenvolupament d'altres projectes en curs. Això és així perquè es tracta de projectes els requeriments dels quals són canviants i els usuaris no tenen totalment clar què necessiten ni tampoc què els pot oferir la FIC. Si un projecte no és autònom i, per exemple, necessita les dades d'un altre projecte que està en desenvolupament, és molt probable que sorgeixin conflictes entre ells, ja que segurament les necessitats de dades del projecte no autònom canviaran durant el desenvolupament, i per a l'altre projecte adaptar-se a aquestes necessitats canviants serà més costós del que s'esperava inicialment.

Resulta especialment important que els projectes siguin autònoms quan es planifica el seu desenvolupament per part de diferents empreses de serveis, sobretot si aquestes el duen a terme amb un pressupost tancat.

Generalment, molts projectes estaran interrelacionats, treballaran sobre les mateixes dades. Per tant, si volem que dos projectes que treballen sobre les mateixes dades siguin autònoms, hauran de desenvolupar-se de manera seqüencial.

Aquesta última condició sovint no és fàcil d'imposar, ja que es necessita disposar dels magatzems de dades departamentals el més aviat possible. Tot i així, és raonable si es vol evitar un conflicte entre els projectes. Si es desenvolupen de manera paral·lela projectes no autònoms, serà crític que el desenvolupament del component d'integració i transformació es faci de manera que no es repeteixin elements entre projectes ni que tampoc sorgeixin conflictes.

El conflicte principal entre els projectes desenvolupats de manera paral·lela sorgeix en obtenir les dades en el component d'integració i transformació.

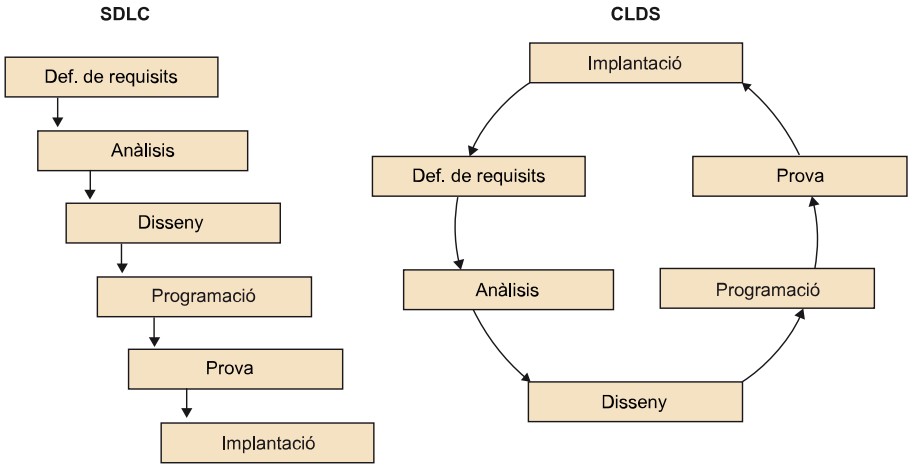
Tots els projectes han de considerar el desenvolupament dels components d'obtenció i emmagatzematge de les dades i accés a aquestes. Així i tot, en alguns casos no s'han de desenvolupar tots els components perquè ja estan desenvolupats. És a dir, es poden plantejar projectes que es dediquin a explotar dades que estan disponibles al magatzem de dades corporatiu (projectes que només tenen component d'accés). També pot haver-hi projectes la missió dels quals sigui ampliar el conjunt de dades que hi ha al magatzem de dades corporatiu (projectes que només tenen components d'obtenció i emmagatzematge). En aquests casos, la situació és diferent d'aquella en què es duia a terme un desenvolupament parcial de l'arquitectura de la FIC, ja que ara s'estudien tots els components de cada projecte, però alguns d'aquests ja estan desenvolupats. De la mateixa manera, en el moment de dur a terme alguns projectes, ens trobarem taules de bases de dades que podrem utilitzar i

que ja estan desenvolupades i actualitzades pel component de transformació i càrrega d'altres projectes ja implementats. Generalment ens trobem taules compartides per molts projectes; es tracta de taules que seran crítiques per la FIC i que acabaran essent entitats mestres amb un tractament especial a la FIC.

Per desenvolupar cadascun dels projectes, podem aplicar la metodologia de desenvolupament en cascada o SDLC, o la metodologia en espiral, també anomenada CLDS.

Les fases de la metodologia de desenvolupament en espiral són les mateixes que les de la metodologia en cascada. La diferència entre les dues és l'ordre d'execució de les seves fases, així com la forma en què es realitzen, iterativa en espiral i seqüencial en cascada.

Figura 12. Fases de la metodologia de desenvolupament



És convenient aplicar la metodologia de desenvolupament CLDS quan no es tenen clars els requeriments del sistema que cal desenvolupar i aquests no es poden descobrir de manera immediata per mitjà d'entrevistes convencionals amb els usuaris.

Les fases de la metodologia de desenvolupament CLDS són les següents:

a) Es comença per implantar una primera versió del sistema als usuaris: aquesta podria ser un model en paper o un prototip de sistema.

b) Els usuaris proven aquesta versió.

c) Es duu a terme el desenvolupament necessari per obtenir, emmagatzemar i analitzar les dades de la versió de prova.

d) Una vegada desenvolupats els programes necessaris, es realitza un disseny formal del sistema.

e) S'analitzen els resultats del disseny, i es reformulen i reprogramen si és necessari.

f) Com a últim pas, s'entenen els requeriments del sistema.

Aquests passos es repeteixen fins a tenir desenvolupat un sistema que compleixi les necessitats dels usuaris. Els usuaris van descobrint les seves necessitats mitjançant l'ús de les versions preliminars del sistema que van refinant-se successivament, fins a aconseguir una versió final.

En l'entorn de construcció d'un magatzem de dades departamental, és freqüent que els analistes no tinguin una idea clara de les característiques del sistema que necessiten fins que el veuen funcionant. En aquest cas, la metodologia de desenvolupament CLDS és més adequada, ja que permet fer refinaments successius del sistema fins a definir clarament els requeriments del sistema que es desitja.

2.4. Evolució de l'entorn operacional

Quan es construeix la FIC, alguns elements de l'entorn operacional deixen de ser útils i es poden deixar d'utilitzar.

2.4.1. Evolució en l'entorn operacional de teranyina

L'entorn operacional havia evolucionat a partir del desenvolupament de programes d'extracció i emmagatzematges temporals de dades en el que havíem denominat entorn operacional de teranyina. Quan es construïa cadascun dels projectes autònoms en els quals s'ha dividit la construcció de la FIC, part dels programes d'extracció i dels emmagatzematges temporals de dades, és a dir, part de la «teranyina» deixa de tenir utilitat, ja que la seva funció passa a ser exercida pel nou magatzem de dades departamental i, per tant, es pot procedir al seu desmantellament.

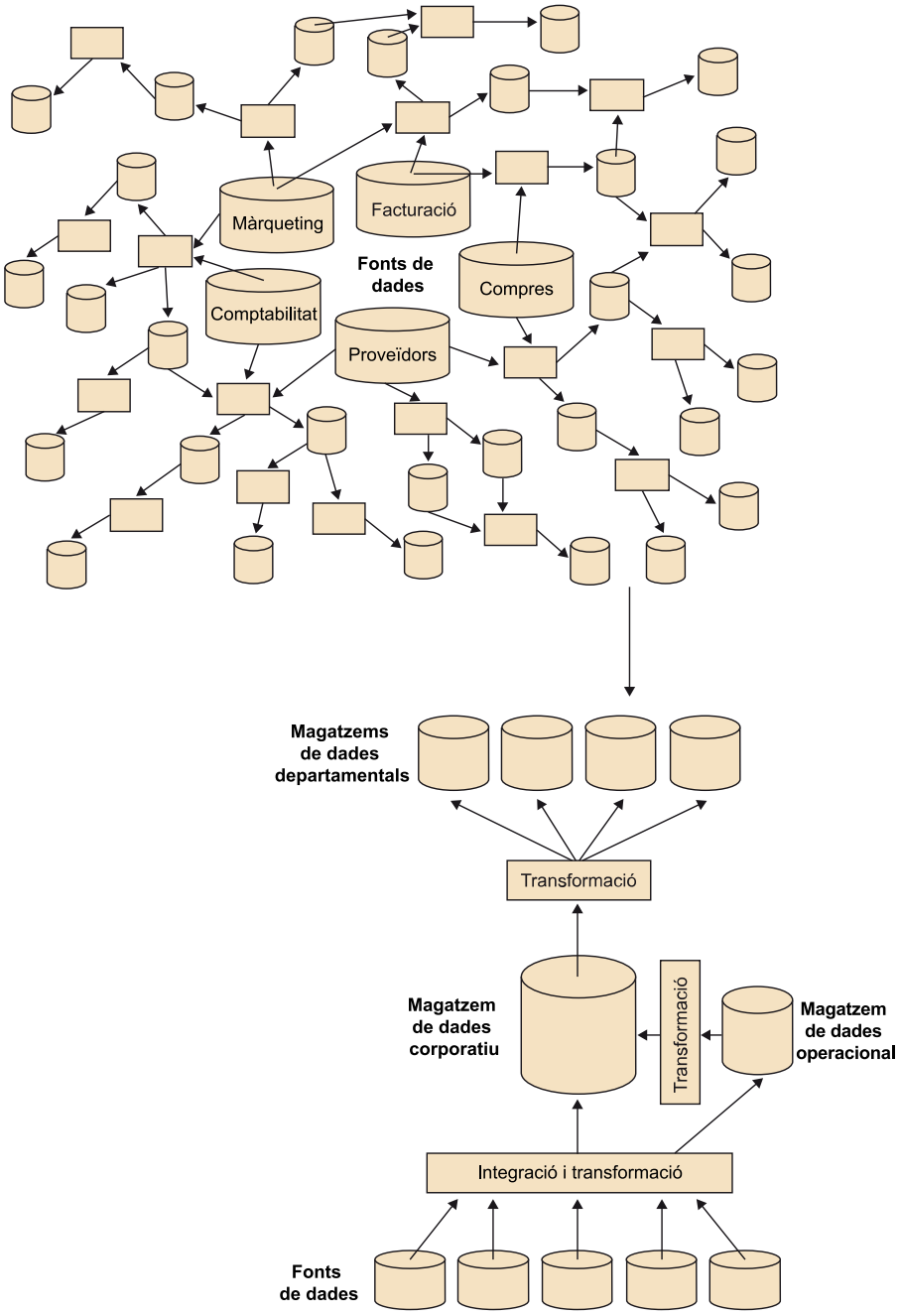
El **desmantellament de la teranyina** consisteix a localitzar els programes d'extracció i els emmagatzematges temporals de dades que deixen de ser útils i evitar que es tornin a executar o crear.

És convenient que les operacions de desmantellament es realitzin de manera progressiva dins de cada projecte de desenvolupament de la FIC. Per desenvolupar el component d'integració i transformació, s'haurien d'analitzar les diferents fonts de dades, la qual cosa implica determinar i documentar la part de la teranyina que queda coberta i passa a ser redundant amb el desenvolupament del nou projecte.

El desmantellament de la teranyina es duu a terme de manera progressiva. A l'hora de construir un magatzem de dades departamental, podem desmantellar la part de la teranyina de l'entorn operacional que exercia la funció del nou sistema construït.

El desmantellament de la teranyina de l'entorn operacional s'ha de tenir en compte quan es planifica la construcció de la FIC. Si no es desmantella, l'organització estarà malgastant els recursos que consumeixen els programes d'extracció i els emmagatzematges temporals de dades.

Figura 13. Desmantellament de l'entorn operacional de teranyina



Podem veure el resultat de desmantellar l'entorn operacional a la figura 13. Passem de tenir l'entorn operacional de teranyina a tenir la FIC, on les aplicacions operacionals actuen com a fonts de dades i l'extracció de les dades es fa de manera controlada i sistemàtica mitjançant el component d'integració i transformació.

Un dels principals beneficis de la implantació de la FIC és l'eliminació de la teranyina de l'entorn operacional. El seu desmantellament porta un estalvi de temps en eliminar el treball de manteniment d'un entorn d'estructura complexa. Un magatzem de dades no només porta beneficis per la millora en l'emmagatzematge de les dades i dels processos d'anàlisi que optimitzaran la presa de decisions i la detecció d'oportunitats, sinó que implica l'eliminació de processos i treballs recurrents que aportaven una informació que pecava de qualitat i que generalment era costosa d'obtenir.

2.4.2. Altres canvis en l'organització

Amb la FIC es produeix una evolució del tipus d'informes que utilitzen els analistes de l'organització. La FIC, principalment mitjançant els magatzems de dades departamentals, permet als analistes generar els informes a mida. Les aplicacions operacionals i l'entorn de teranyina creat a partir d'aquestes permetien als usuaris obtenir informes estructurats de format fix.

La FIC permet als analistes incrementar de manera considerable l'ús d'informes a mida, com a suport al procés de presa de decisions, en detriment del nombre d'informes de format fix oferts per les aplicacions operacionals.

D'una banda, els equips dedicats al desenvolupament i manteniment de programes per a la generació d'informes (els programes d'extracció que formaven la teranyina de l'entorn operacional) s'hauran d'adaptar a la nova situació i reubicar en algun altre departament. Així mateix, en el cas que algun dels informes es generés de manera manual (situació bastant freqüent en algunes organitzacions) i ara passés a generar-se dins de la FIC, les persones encarregades d'aquesta tasca s'hauran de reubicar per dur a terme altres tasques en l'organització.

La FIC, a més d'afectar la manera de treballar dels analistes, afectarà els equips de desenvolupament que es dedicaven a generar els informes, que ara passen a generar-se dins de la FIC.

2.5. Ús del sistema de processament analític en línia (OLAP) a la FIC

Una vegada construïts el *Data Warehouse* o els *Data Marts* (magatzems departamentals), necessitem eines que ens permetin realitzar consultes i analitzar la informació que contenen els magatzems de dades de manera fàcil i productiva.

En l'actualitat, existeixen dues tendències pel que fa a l'explotació de dades. Aquests dos models poden ser complementaris i no necessàriament excloents, doncs, com és habitual, cadascun d'ells aporta certs avantatges i inconvenients al sistema de suport a la presa decisions.

2.5.1. Magatzems de dades amb sistemes OLAP

Les eines OLAP permeten als analistes realitzar consultes complexes sobre grans volums de dades carregats prèviament als magatzems de dades departamentals, sense necessitat de tenir coneixements tècnics avançats. Perquè el sistema OLAP funcioni correctament, els *Data Marts* han de ser dissenyats segons el model multidimensional, la qual cosa permet crear magatzems de dades multidimensionals, també coneguts com a cubs OLAP.

El principal avantatge d'aquest model és la robustesa, consolidació i validació de les dades que aporten valor i qualitat al sistema analític, tant a la mateixa dada com al resultat de les múltiples consultes realitzades. Com que es tracta d'un model preconfigurat segons les necessitats i requeriments dels mateixos analistes, es pot garantir que els càlculs i mètriques obtingudes seran vàlids per a tota l'organització i es reforça la idea de la dada única, on bàsicament es pretenen reduir o eliminar ambigüitats i diferents interpretacions sobre una mateixa dada, la qual cosa donaria resultats diferents per a un mateix context.

Per contra, el principal inconvenient dels sistemes OLAP i el motiu principal pel qual algunes organitzacions es plantegen prescindir-ne és l'alt cost de desenvolupament que implica obtenir una solució OLAP robusta. A més del temps necessari per a la seva implementació i posterior manteniment, és necessari tenir a l'organització tècnics amb coneixements en aquestes tecnologies, o bé subcontractar aquest desenvolupament. Per treure el màxim profit al model OLAP, també serà necessari que els analistes coneguin llenguatges de consulta específics com a MDX.

En qualsevol cas, si les condicions de l'organització són les adequades, disposar d'un sistema OLAP de consultes multidimensionals aportarà gran valor i coneixement sobre les seves pròpies dades, facilitant enormement la feina dels analistes i millorant el procés de suport a la presa de decisions, en els diferents àmbits de l'empresa.

2.5.2. Magatzems de dades sense OLAP

Amb l'auge de les tecnologies associades al fenomen del *Big Data*, en els darrers anys, han aparegut noves eines d'anàlisi que, d'alguna manera, han anat superant les limitacions clàssiques dels models OLAP. Aquestes eines denominades *self-service BI* estan orientades als analistes perquè puguin accedir directament als magatzems de dades i a uns altres orígens de dades, sense necessitat d'esperar que els departaments de TI preconfigurin un entorn OLAP d'anàlisi, que, com hem vist, pot arribar a ser un procés bastant costós.

Certament, aquestes eines de *self-service BI* també requereixen certa formació tècnica, però estan dissenyades perquè es puguin arribar a utilitzar sense necessitat d'aprendre a programar ni tenir coneixements tècnics avançats, reduint la dependència dels analistes amb els departaments de TI. Una altra característica important d'aquest tipus d'eines és que permeten realitzar anàlisis i consultes avançades de manera visual, la qual cosa facilita el traspàs del coneixement cap a la resta dels professionals de l'organització, que no necessàriament tindran coneixements tècnics ni analítics. Algunes eines d'aquest tipus són Power BI (Microsoft), Tableau o QlikView, entre moltes altres.

En l'àmbit de la ciència de dades, l'ús d'eines gràfiques avançades també facilita el descobriment de patrons ocults a les dades, que, d'una altra manera, resultaria molt més difícil de trobar, així com realitzar estudis i investigar nous escenaris.

Hem vist que, amb l'ús d'eines de *self-service BI*, els analistes tenen més llibertat per explotar les dades, però el resultat d'aquestes anàlisis és més personal i no necessàriament seguiran els criteris de qualitat establerts per l'organització. Es pot donar el cas que dos analistes obtinguin resultats diferents en un mateix problema causat per interpretacions diferents i, en el pitjor dels casos, pot arribar a invalidar els dos resultats.

Vegem un exemple

Suposem que un hospital necessita calcular la mètrica «Estada mitjana» dels pacients ingressats a traumatologia en un període de temps. Aquest càlcul bàsic pot ser utilitzat amb posterioritat per calcular costos, contractar personal, organitzar calendaris, aprovisionar fàrmacs, etc.

Aparentment, sembla un càlcul simple i podríem dir que simplement necessitem calcular «Quants dies han estat ingressats cadascun dels pacients», per posteriorment calcular la mitjana.

No obstant això, aquest requeriment dona peu a diferents interpretacions de «Què és una estada hospitalària?»:

- Són els dies que el pacient ha estat ingressat?
- Són les nits que el pacient ha pernoctat a l'hospital?
- Si el pacient ingressa un dia a les 23.00 h, aquest primer dia compta com a estada?
- Si donen l'alta hospitalària al pacient un dia a les 08.00 h, aquest últim dia compta com a estada?

- Quantes hores ha d'estar ingressat com a mínim un pacient perquè es consideri una estada?

Sembla clar que, si l'organització no estableix un criteri únic (típicament recollit en els sistemes OLAP mitjançant mètriques calculades) i cada analista fa una interpretació diferent del concepte «Estada mitjana» en la seva anàlisi *self-service BI*, les diferències en els càlculs derivats poden ser significatives, arribant fins i tot a invalidar-los tots. El responsable (usuari final), que ha de prendre una decisió basant-se en aquesta informació, no tindrà la certesa de disposar d'informació de qualitat ni útil.

2.5.3. Models mixtos o complementaris

Moltes organitzacions disposen de sistemes OLAP consolidats, però no volen renunciar als beneficis i agilitat dels sistemes de *self-serviceBI*. En aquests casos, es dissenya un model mixt, on els dos models es complementen. El sistema OLAP ofereix la robustesa i qualitat de les explotacions periòdiques típicament automatitzades, i les eines *self-service BI* ofereixen als analistes de dades i negoci la possibilitat d'explotar i explorar les dades amb més flexibilitat.

En un model mixt, els analistes tenen les eines per explorar les dades i analitzar totes les possibles interpretacions. Després d'un procés de deliberació i consens dels resultats, aquests es validen per un comitè d'experts i finalment es traspassen al sistema OLAP. D'aquesta manera, es comparteix un únic criteri en tota l'organització i la dada precalculada pot ser utilitzada per qualsevol professional d'empresa, amb garanties de qualitat, fins i tot sense haver participat en el procés de validació. Disposar d'aquesta informació en el sistema OLAP també permet automatitzar certs processos que han de ser recalculats periòdicament.

2.6. Perfils a l'equip de gestió i desenvolupament de la FIC

A més dels perfils habituals en qualsevol equip de desenvolupament de projectes, a l'equip de desenvolupament de la FIC apareix un conjunt de perfils que no són habituals en altres projectes i que estudiarem a continuació juntament amb els perfils de gestió característics de la FIC.

2.6.1. L'administrador de la FIC

L'administrador de la FIC és el seu responsable davant de l'organització. La seva missió és que la FIC s'adapti a les necessitats de l'organització. Per aquest motiu, ha de conèixer les possibilitats que ofereix, així com les necessitats d'informació de l'organització, i controlar que les satisfaci prou.

Per a cada projecte que es desenvolupi, s'ha d'assegurar que compleixi els requeriments d'àmbit i funcionalitat, així com els que fan referència a cost i temps de desenvolupament.

L'administrador de la FIC és el màxim responsable davant de l'organització que la FIC compleixi al llarg del temps els seus requeriments d'informació.

2.6.2. Els analistes de requeriments de negoci

L'analista de requeriments de negoci és el responsable del primer projecte: el projecte global de desenvolupament. Comptarà amb un equip d'analistes per al seu desenvolupament. La seva missió és identificar els requeriments d'informació per part de l'organització i planificar el desenvolupament de la FIC de manera que aquests requeriments es compleixin.

L'administrador de la FIC és el màxim responsable del seu entorn, mentre que els analistes de requeriments de negoci són els interlocutors entre els usuaris de la FIC i l'equip que l'ha de construir i mantenir.

Els analistes de requeriments de negoci recullen les necessitats d'informació dels usuaris de la FIC i les transmeten a la resta de l'equip de desenvolupament.

A més de ser responsables del projecte global de desenvolupament, també ho són dels diferents projectes autònoms en els quals es divideix la construcció de la FIC, que es desenvoluparan posteriorment.

2.6.3. L'arquitecte de la FIC

L'arquitecte és el responsable del disseny de la FIC, dissenya la seva arquitectura i es responsabilitza dels projectes de desenvolupament d'infraestructura.

L'**arquitecte de la FIC** analitza les fonts de dades de l'organització i dissenya els esquemes de dades i l'estructura de la FIC segons els requeriments d'informació dels usuaris i les possibilitats que ofereixen les fonts de dades.

L'arquitecte de la FIC treballa de manera conjunta amb l'analista de requeriments de negoci en el projecte global de desenvolupament de la FIC, especialment per definir els requeriments d'infraestructura dels projectes.

2.6.4. El patrocinador de la FIC a l'organització

El patrocinador de la FIC és una figura política l'objectiu de la qual és aconseguir el suport necessari en l'organització per desenvolupar-la i salvar els obstacles interns que sorgeixin.

L'àmbit dels projectes de desenvolupament de la FIC comprèn tota l'organització. Requereix integrar les dades dels diferents departaments i, una vegada la FIC ja funciona, pot canviar la manera de treballar d'equips importants de personal. Per tot això, és freqüent que els responsables d'alguns departaments es mostrin poc acostumats a compartir les dades o, generalment, que alguns gestors d'alt nivell a l'organització mostrin la seva oposició a la FIC pels canvis que representa. El patrocinador de la FIC ha d'interaccionar amb tots ells per intentar aconseguir el seu suport o, almenys, reduir la seva oposició.

D'altra banda, segons el plantejament de construcció de la FIC mitjançant el desenvolupament de projectes autònoms presentat en aquest mòdul, els diferents projectes de construcció de magatzems de dades departamentals estan justificats, ja que aporten un benefici a l'organització més gran que el cost que representa el seu desenvolupament. Tot i així, altres projectes molt importants, com, per exemple, el projecte global de desenvolupament i el projecte de desenvolupament de la infraestructura, no tenen una justificació immediata del seu cost. El patrocinador de la FIC a l'organització tindrà com a missió aconseguir els fons necessaris per al desenvolupament d'aquests projectes.

Tenint en compte el tipus de persones amb les quals ha d'interaccionar i la tasca que ha de dur a terme, interessa que el patrocinador de la FIC en l'organització sigui un directiu de prou nivell perquè la seva tasca resulti més fàcil. A més, ha d'estar convençut dels avantatges que la FIC pot proporcionar a l'organització.

El **patrocinador de la FIC** en l'organització és una figura política i té per objectiu aconseguir l'èxit en el desenvolupament de la FIC. Generalment, es tracta d'un directiu d'alt nivell que coneix els avantatges que aporta la FIC a l'organització i que s'encarrega d'obtenir els fons necessaris i de resoldre els problemes interns que sorgeixin a l'organització.

2.6.5. El gestor de canvis organitzacionals

La implantació de cadascun dels projectes de desenvolupament de la FIC representa un canvi en la manera de treballar en part de l'organització. El gestor de canvis organitzacionals té com a responsabilitat principal gestionar l'impacte de la FIC a l'organització.

D'una banda, cada projecte de desenvolupament de la FIC representa un canvi en la manera de treballar dels usuaris del nou projecte. Anteriorment, obteníem els informes del Departament d'Informàtica o mitjançant persones que els elaboraven manualment, i amb freqüència els informes tenien forma de llistats. Ara, amb la FIC, poden generar ells mateixos els informes que necessiten i obtenir-los directament en temps real. Atès que a moltes persones no els agrada cap mena de canvi, encara que aquest sigui beneficiós a curt termini, s'ha de fer una tasca de màrqueting per facilitar l'acceptació dels canvis. Aquesta tasca és responsabilitat del gestor de canvis organitzacionals.

D'altra banda, el canvi en la manera de generar els informes pot desplaçar diferents treballadors del Departament d'Informàtica o d'altres departaments, que anteriorment eren els encarregats de fer la tasca que ara passa a estar coberta per la FIC. Aquests treballadors seran molt útils com a membres de l'equip de desenvolupament de la FIC o en altres activitats dins de l'organització. El gestor de canvis organitzacionals ha d'intentar reduir la incertesa d'aquests treballadors i col·laborar per determinar quines seran les seves noves funcions, així com per minimitzar els possibles conflictes que es puguin produir.

El **gestor de canvis organitzacionals** té com a responsabilitat que els membres de l'organització acceptin els canvis que implica la FIC.

2.6.6. El gestor de canvis de les metadades

Les metadades són un component molt important de la FIC, ja que en descriuen l'estructura i contingut i, a més, permeten interconnectar la resta dels components entre si.

Tenint en compte la seva importància, interessa disposar d'una persona responsable de les metadades de la FIC. La seva missió és assegurar que reflecteixin la situació actual de la FIC i siguin accessibles tant per als analistes d'informació com per als equips de desenvolupament que ho requereixin. A més, ha d'assegurar que les metadades puguin ser enteses pels diferents usuaris que hi accedeixen. Tenint en compte la varietat d'eines que generen i utilitzen metadades, no resulta una tasca senzilla.

El **gestor de metadades** és responsable de la situació i de l'accés a les metadades a la FIC.

2.6.7. Els analistes de la qualitat de la dada

Els analistes de la qualitat de les dades tenen com a responsabilitat assegurar que aquelles que s'han obtingut de les fonts de dades operacionals satisfacin els requeriments d'informació de l'organització.

S'encarreguen de verificar que les dades s'han obtingut, transformat i carregat de la manera requerida i el resultat és de la qualitat esperada; és a dir, que les dades en els diferents magatzems de dades de la FIC són les adequades i són correctes.

En el cas de detectar dades incorrectes, la solució no és corregir-les on s'han detectat, sinó trobar la font d'aquesta incorrecció i solucionar allà el problema que s'hagi produït.

Els **analistes de la qualitat de les dades** han de detectar aquelles que no s'adaptin al grau de qualitat requerit per la FIC. La seva missió no consisteix a corregir aquestes dades, sinó a identificar els motius de la baixa qualitat i recomanar accions que condueixin a la solució d'aquest problema.

La quantitat necessària d'analistes de la qualitat de les dades dependrà del grau de qualitat de les dades operacionals i del grau requerit. Durant les fases de desenvolupament de la FIC, segurament seran necessaris diferents analistes de qualitat. Quan la FIC estigui en funcionament, n'hi hauria d'haver prou amb una sola persona, fins i tot a temps parcial.

2.6.8. L'administrador de bases de dades

A l'organització hi ha d'haver personal d'administració de bases de dades per a l'entorn operacional, però també convé que existeixi personal específic per a l'entorn de la FIC.

Tot i que el sistema de gestió de bases de dades utilitzat en l'entorn operacional pot ser el mateix que el que s'ha utilitzat en els diferents magatzems de dades, les seves característiques de configuració varien de manera radical entre els dos entorns. Per aquest motiu, interessa disposar d'administradors especialitzats en l'entorn dels magatzems de dades que siguin capaços d'obtenir el màxim rendiment dels sistemes i que no hagin de canviar contínuament la manera de pensar a l'hora d'administrar tant l'entorn operacional com el dels magatzems de dades.

Així mateix, com que la FIC obté les seves dades a partir de l'entorn operacional, és adequat que els responsables tècnics dels dos entorns siguin persones diferents, de manera que la integritat de les dades de la FIC no es vegi afectada

per possibles problemes de l'entorn operacional. Si l'administrador dels dos entorns fos únic, davant de qualsevol problema es pot produir un conflicte de prioritats; mentre que la màxima prioritat per a un administrador de bases de dades propi serà la FIC.

L'**administrador de bases de dades** ha de monitorar de manera contínua els processos d'obtenció i accés a dades, i configurar els sistemes perquè aquests es facin de manera òptima.

2.6.9. Especialistes a obtenir i accedir a les dades

A més de les figures estudiades en els apartats anteriors, l'equip de desenvolupament de la FIC estarà format per especialistes en les operacions específiques en les quals es descomponen els projectes de desenvolupament de la FIC: obtenció i emmagatzematge de les dades i accés a aquestes dades.

Pel que fa a l'obtenció de les dades, particularment si aquesta es fa mitjançant l'ús d'alguna eina de suport, l'equip de desenvolupament comptarà amb els especialistes desenvolupadors del component d'integració i transformació. Aquests coneixen el contingut i les possibilitats de les fonts de dades i també els procediments i les eines per obtenir les dades a partir d'aquests.

Pel que fa al component d'accés, l'equip de desenvolupament comptarà amb especialistes en les eines d'accés a les dades i metadades que utilitzaran els diferents tipus d'analistes d'informació de l'organització. Aquests especialistes construiran els mètodes d'accés necessaris per satisfer les necessitats dels usuaris en aquest sentit.

Tal com ja s'ha esmentat, el component d'integració i transformació és un component clau de la FIC, i la necessitat de recursos en aquest component ha de quedar ben coberta.

2.6.10. L'enginyer de dades (Data Engineer)

En l'actualitat, han aparegut nous rols professionals com el de l'enginyer de dades. En realitat, són perfils tècnics ja coneguts però que, amb la revolució del *Big Data*, s'han especialitzat en l'enginyeria i arquitectura dels sistemes de gestió de dades. Aquests professionals s'encarreguen de preparar i processar la informació, garantint la qualitat de la dada i establint les relacions necessàries.

2.7. Usuaris de la FIC

Una vegada tenim construïda la FIC a la nostra organització, o com a mínim algun dels projectes autònoms identificats, és el moment de donar accés a tots els usuaris analistes perquè realitzin les seves consultes sobre els diferents ma-

gatzems de dades. Existeixen diversos perfils d'usuaris d'una FIC, però tots ells són, en major o menor grau, usuaris amb coneixements avançats en anàlisi, explotació o visualització de dades.

És necessari diferenciar entre usuari de la FIC i usuari final. Típicament, l'usuari de la FIC és un analista de dades que analitza i prepara la informació que serà consultada per l'usuari final dins del procés de suport a la presa de decisions. Els usuaris finals poden ser molt variats, però és habitual trobar rols com els de direcció, comandaments intermedis, gestors, responsables d'àrea, etc. És a dir, professionals que necessiten dades actualitzades i de qualitat per poder prendre les decisions de gestió, organització, operatives o estratègiques que considerin oportunes, segons les seves responsabilitats. No obstant això, cada vegada és més habitual que qualsevol tipus de professionals, sense un rol de gestió clarament definit, sol·licitin informació als analistes per millorar i optimitzar els seus propis processos productius.

En els darrers anys, aquests perfils s'han anat popularitzant i especialitzant, i han donat lloc a diferents classes de professions en l'anàlisi de dades.

2.7.1. L'analista de dades (Data Analyst)

És el professional amb capacitat per comprendre les dades de l'usuari final, sintetitzar-les, contextualitzar-les i obtenir informació útil per a l'organització. No disposa de la formació d'un *Data Scientist*, però té una visió àmplia del negoci, moltes vegades com a resultat d'anys d'experiència, que li permet fer tasques analítiques avançades.

2.7.2. El científic de dades (Data Scientist)

És el professional amb coneixements de negoci i tècnics, capaç de realitzar consultes i anàlisis complexes, creuant diferents orígens de dades. Un científic de dades ha de ser capaç d'extreure coneixement de les dades, identificar patrons de comportament, implementar algorismes de processament de dades i definir models predictius. Per fer el seu treball, un *Data Scientist* ha de tenir, entre d'altres, coneixements matemàtics, estadístics, de programació, de bases de dades, de mineria de dades, de modelatge i de visualització de dades. Una altra faceta important del científic de dades és la capacitat i habilitat de comunicació per poder explicar amb claredat i sense ambigüitats el resultat del seu treball.

2.7.3. L'analista de negoci (Business Analyst)

És el professional expert en analítica de negoci amb formació en modelatge i visualització de dades mitjançant eines visuals: gràfics, imatges, infografies… Aquest tipus de visualitzacions faciliten enormement el traspàs i assimilació del coneixement dins de les organitzacions.

2.7.4. El responsable de dades (Chief Data Officer)

És el líder de l'estratègia de gestió i anàlisi de dades de l'organització. Podríem dir que té un perfil similar al del CIO (*Chief Information Officer*), típicament el director de TI, però especialitzat en la gestió de la dada.

3. Desenvolupament del component d'integració i transformació

En aquest apartat, estudiarem les principals solucions que se solen aplicar en el desenvolupament d'algunes de les activitats del component d'integració i transformació. Ens centrarem en l'obtenció de les actualitzacions de les dades, les transformacions, la integració i l'actualització de les dades del magatzem de dades, així com el suport que ofereixen algunes eines del mercat per a la seva implementació i la importància de les metadades.

3.1. Construcció dels components d'extracció i obtenció de dades

El procés per obtenir les dades a partir de les fonts de dades origen i actualitzar les dades dels magatzems de dades es duu a terme mitjançant un conjunt d'aplicacions que s'executen amb aquesta finalitat. Aquest procés es divideix en dues fases:

- Obtenir la imatge inicial.
- Obtenir les actualitzacions.

Aquest component també és conegut com a ETL, sigles del terme en anglès *Extract Transform Load*.

3.1.1. Obtenir la imatge inicial

La imatge inicial s'obté amb un conjunt d'aplicacions que generalment s'executa una sola vegada. El resultat d'aquesta fase és una imatge de la situació actual dels sistemes operacionals, obtinguda mitjançant un buidatge de les seves respectives bases de dades. Normalment, la imatge inicial s'obté sense dificultat; encara que, segons les característiques dels sistemes operacionals, això no sempre és així.

Exemple d'obtenció de la imatge inicial de les dades

En un sistema operacional basat en una base de dades relacional, és possible obtenir les dades de manera immediata, mitjançant un buidatge amb les eines que el sistema de gestió de base de dades ofereix per fer-ho. Tot i així, en una aplicació empaquetada desenvolupada sobre el sistema de fitxers, és possible que només puguem accedir a les dades mitjançant les pantalles de l'aplicació i es requereixi, en aquest cas, un desenvolupament complex i costós en el qual s'hagin d'anar capturant les dades presentades a les diferents pantalles.

Per reflectir al magatzem de dades l'evolució que tenen, partint de la seva imatge inicial, hem d'anar obtenint les actualitzacions que es van produint. Ens trobarem diferents tipus de dades segons la seva estructura:

1) Dades estructurades: amb estructura de dades coneguda, s'emmagatzemen principalment en bases de dades relacionals. La manipulació es fa per mitjà de gestors de bases de dades, i les consultes, mitjançant SQL.

2) Dades semiestructurades: encapsulades en fitxers semiestructurats com l'XML5 o l'SGML6. En aquesta situació, és possible treballar amb el context de negoci, la qual cosa proporciona gran valor a les organitzacions. Actualment trobem bases de dades especialitzades en XML per manipular aquest tipus de dades, i també tècniques com el *web-mining* (mineria de dades aplicada a la web) que permeten recuperar informació de pàgines web.

3) Dades no estructurades: encapsulades en objectes sense una estructura predefinida (àudio, vídeo, PDF o Word) que requereix l'ús de tècniques especials com el *text-mining* (mineria de dades aplicada a fitxers de text) o *l'information retrieval* (tècniques, amb freqüència estadístiques, aplicades a trobar informació relacionada amb un concepte en fitxers).

Actualment, les dades no estructurades o semiestructurades, és a dir, encapsulades en fitxers XML, SGML o fins i tot en objectes sense una estructura predefinida (PDF o Word), també són fonts d'alt valor potencial per a les organitzacions. Permeten des de conèixer informació dels competidors fins a conèixer en profunditat els clients. En el primer cas, per exemple, la informació relativa als preus del competidor es troba a la seva pàgina web, però on realment hi ha l'avantatge competitiu és en l'automatització d'aquest procés.

En aquesta situació, els procediments d'extracció d'informació estàndard, com l'ETL, amb freqüència no són suficients i cal utilitzar tècniques de recuperació d'informació que utilitzen mètodes de reconeixement de patrons, mostreig estadístic, probabilitat, etc. Això complica l'obtenció de la imatge inicial de les dades, que no sempre està focalitzada a les dades internes de l'organització.

D'altra banda, hem de tenir en compte la sèrie històrica que cal obtenir. Generalment, els sistemes operacionals només guarden una imatge de les seves dades o bé una història reduïda d'aquestes dades.

Si les diferents imatges emmagatzemades en els sistemes operacionals s'han anat perdent a mesura que s'han fet modificacions, només podrem disposar de la història que emmagatzemem a partir del moment en què es construeixen els magatzems de dades.

En alguns casos, per diferents motius (per exemple, per motius legals), pot haver-hi una història més extensa de les dades, a vegades fora dels sistemes operacionals, encara que obtinguda a partir d'aquests sistemes. En cada cas, haurem de valorar si és útil per als analistes disposar als magatzems de dades de la història que hi havia abans; en cas positiu, en lloc de partir de la imatge inicial, partiríem d'una pel·lícula inicial.

Dades històriques en un banc

En un banc, el sistema operacional de gestió de moviments dels comptes només guarda les dades dels últims dotze mesos. Les dades dels mesos anteriors fins a un total de cinc anys s'han d'emmagatzemar per motius legals. Tot i així, aquests moviments històrics no s'emmagatzemen en el sistema operacional, sinó que mensualment s'extreuen de la base de dades del sistema i s'emmagatzemen en un mitjà d'emmagatzematge més econòmic. Tot i que aquestes dades romanen accessibles dins de l'organització, només s'hi accedeix de manera puntual, per motius operacionals, no per analitzar-les.

Per obtenir la imatge inicial, haurem de desenvolupar un conjunt d'aplicacions d'obtenció de les dades de les fonts de dades, que generalment s'executaran una sola vegada. En algun cas i atès que també cal desenvolupar un procés per realitzar les actualitzacions de dades, podem aprofitar aquest procés per a l'obtenció de la imatge inicial. Per exemple, si el procés d'actualització està obtenint les dades de la font origen filtrades per un rang temporal (dies, setmanes, mesos), ens podem plantejar executar aquest procés d'actualització N vegades per a tots els períodes que componguin la sèrie històrica de la imatge inicial.

3.1.2. Mètodes per obtenir les actualitzacions de les dades

La manera d'obtenir les dades per a les actualitzacions dependrà dels requeriments dels analistes sobre els magatzems de dades, i també de les possibilitats ofertes per les fonts de dades.

A vegades l'obtenció d'actualitzacions consistirà en un simple filtratge sobre la informació origen, com pot ser l'obtenció de dades corresponents a un rang temporal.

Exemple d'actualització mitjançant filtratge

Un operador de telecomunicacions anomenat ATEL té un magatzem de dades en el qual en una de les taules guarda les dades de les portabilitats d'operador realitzades pels clients en els quals l'operador ATEL sigui origen o destinació d'aquesta portabilitat. Els analistes de negoci analitzen les portabilitats del client i la data límit de les seves anàlisis és el dia laborable anterior. En aquest cas, l'obtenció d'actualitzacions podria ser un filtratge de dades sobre la font origen en la qual obtinguem les portabilitats realitzades el dia anterior.

En altres ocasions, l'obtenció de les actualitzacions no es podrà realitzar mitjançant un simple filtratge, pel fet que no existeix un rang temporal que defineixi unívocament les dades que cal actualitzar o perquè una freqüència d'actualització alta ens obligui a utilitzar mètodes de detecció d'actualitzacions. A la figura 14, es representen els diferents mètodes per obtenir les actualitzacions:

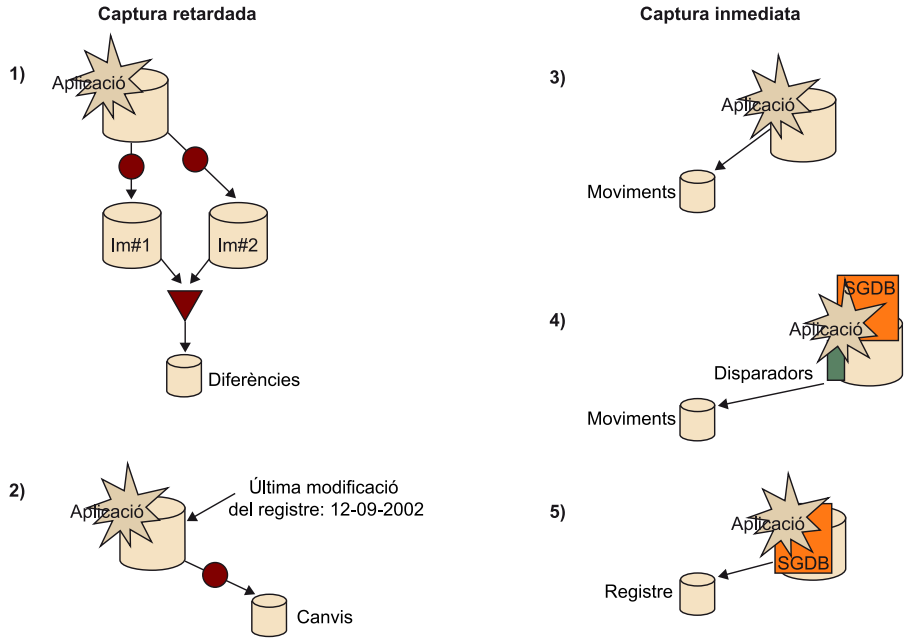
1) Comparació d'imatges (cas 1 de la figura 14): algunes fonts ofereixen la possibilitat d'obtenir un buidatge de les seves dades de manera massiva. Un exemple d'aquest tipus de fonts són els fitxers ordinaris. Si treballem amb

bases de dades relacionals, el buidatge es pot obtenir mitjançant consultes que seleccionen les dades que ens interessin. Per a aquest tipus de fonts de dades, podem obtenir les actualitzacions que s'hagin produït comparant imatges successives que s'hagin obtingut.

2) Fonts amb una empremta de temps (cas 2 de la figura 14): en aquest cas, les fonts emmagatzemen per a cada registre modificat el moment en el qual es va dur a terme la modificació. D'aquesta manera, obtenir les darreres modificacions és immediat: n'hi ha prou amb aconseguir els registres marcats amb una empremta de temps posterior a la de l'últim conjunt de modificacions obtingut. Si volem recuperar les operacions d'esborrament, caldrà realitzar un esborrament lògic en el qual es marqui el registre, però no s'elimini físicament.

Aquests dos mètodes es denominen de captura retardada, ja que no s'intenta capturar les modificacions en el moment en què es produeixen, sinó posteriorment. Tenen l'inconvenient que poden perdre moviments: si es fan diferents modificacions sobre un registre i aquestes no s'obtenen de manera immediata després de produir-se cadascuna, per mitjà dels mètodes anteriors obtindrem un resum de totes les operacions fetes en una única operació de modificació.

Figura 14. Tècniques d'obtenció d'actualitzacions



Exemple de pèrdua de moviments de les dades

A la imatge inicial del saldo d'un compte obtenim un valor de 1.000, i posteriorment es produeix un ingrés de 500 i un reintegrament de 700. Si en aquell moment tornem a agafar la imatge del saldo del compte per qualsevol dels dos mètodes anteriors, obtenim un valor de 800; si comparem els saldos, obtenim que s'ha produït un reintegrament de 200 i hem perdut el detall dels moviments fets.

3) Anàlisi del fitxer delta de moviments (cas 3 de la figura 14): algunes aplicacions emmagatzemen les modificacions fetes sobre les seves dades en un fitxer de moviments denominat fitxer delta. Generalment, s'utilitza per auditar les operacions. Podem utilitzar aquest fitxer per obtenir les modificacions fetes sobre les dades que ens interessin.

4) Detecció de moviments mitjançant disparadors (cas 4 de la figura 14): si el sistema en el qual es basa la font de dades ofereix la possibilitat de definir disparadors o altres funcionalitats de bases de dades actives, els podem utilitzar per obtenir les modificacions sobre les dades que ens interessin, activant-los quan es produeixin esdeveniments de modificació sobre aquestes dades. Aquests disparadors poden comunicar els canvis produïts a les parts interessades perquè actuïn de manera adequada. Una manera habitual d'actuar, en aquest cas, és la creació d'un fitxer delta de moviments que es pot analitzar posteriorment.

5) Anàlisi del fitxer de registre (cas 5 de la figura 14): algunes fonts de dades, basades en SGBD que ho permeten, emmagatzemen les operacions que s'hi realitzin en un fitxer de registre, generalment per motius de recuperació davant d'errors, o bé perquè sigui possible desfer modificacions. Si analitzem aquest fitxer, es poden obtenir les modificacions fetes sobre les dades que ens interessin. Aquesta solució té l'inconvenient que, encara que es generi a la font de dades, el fitxer de registre no és de fàcil accés, ja que està pensat per ser utilitzat pel SGBD i el seu format no sempre és de domini públic.

Aquests tres últims mètodes d'obtenció de modificacions sobre les dades, a diferència dels dos primers, tenen com a característica principal que no perden cap dels moviments que es produeixen. Se'ls denomina mètodes de captura immediata, ja que obtenen els moviments en el precís instant en el qual es produeixen.

L'avantatge que presenta l'últim cas és que no és invasiu contra la base de dades d'origen, ja que no accedeix ni modifica l'estructura de les taules d'origen. Qualsevol dels mètodes anteriors sí que ho és, i pot afectar el rendiment de la base de dades d'origen. L'inconvenient d'aquest últim cas, com s'ha dit, és la dificultat per interpretar el fitxer de registre que podrà ser diferent segons l'SGDB amb el qual treballem. Alguns SGDB disposen d'eines per interpretar de manera senzilla aquest fitxer, però solen ser eines propietàries i no disponibles en tots els SGDB.

Un altre aspecte que cal considerar en els diferents mètodes és la independència respecte a l'SGBD. Tal com s'ha comentat anteriorment, el cas 5 no és independent de l'SGDB, com tampoc no ho és la creació de disparadors del cas 4. La resta de casos si que són independents de l'SGDB.

6) Mètodes de recuperació d'informació semiestructurada: les tècniques d'accés i consulta són les adequades per extreure informació estructurada, encapsulada en format XML i SGML. Malgrat que trobem diferents aproximacions, no es recomana l'ús de llenguatge de programació ni la creació de rutines *ad hoc*. Els mètodes recomanats són els següents.

a) Utilització d'eines ETL especialitzades que inclouen els mecanismes de manipulació de fitxers XML i també la possibilitat d'analitzar la diferència entre les versions dels documents.

b) Ús de motors especialitzats d'indexació d'informació semiestructurada com Apache Lucene.

7) Mètodes de recuperació d'informació no estructurada. En el cas de la informació no estructurada, també hi ha diferents aproximacions. Els diferents models que trobem són aquests:

a) Booleà: mètode de recuperació simple fonamentat en la teoria algebraica booleana. Estableix una rellevància binària de manera que un document és rellevant o no ho és. La seva simplicitat ha motivat que caigui en desús.

b) Vectorial: compara la consulta amb el text en un document i genera un vector. Vectors que apunten a la mateixa direcció són similars i el grau de versemblança és la funció de l'angle entre els vectors. És un dels mètodes actuals més populars i fiables. En particular, motors Java com Apache Lucene l'utilitzen.

c) Model LSI (*Latent Semantic Indexing*): l'objectiu és calcular la versemblança entre la consulta i el document mitjançant la similitud entre conceptes, i no entre paraules.

d) Mètodes millorats: els mètodes anteriors s'han millorat amb l'ús de motors semàntics, xarxes semàntiques o l'anàlisi de la regressió.

3.1.3. Criteris de selecció del mètode per obtenir les actualitzacions de les dades

Quin mètode utilitzarem d'entre els mètodes que hem presentat per obtenir modificacions? La resposta a aquesta pregunta dependrà en gran mesura de les funcionalitats ofertes per la font de dades i pels requeriments dels usuaris dels magatzems de dades. Els criteris següents no són estrictes, però ens poden servir com a orientació:

a) Si la font de dades disposa d'un fitxer delta de moviments, generalment aquesta serà l'opció que triarem, ja que aquest fitxer sol ser directament accessible i fàcil d'analitzar.

b) Si no existeix fitxer delta, s'analitza el registre en el cas que el generi la font de dades. En aquest cas, l'accessibilitat és menor i la dificultat per analitzar-lo, més gran.

c) Si la font de dades està basada en una base de dades que permet crear disparadors, podrem utilitzar aquest mètode. La definició de disparadors farà que les transaccions siguin més costoses en temps d'execució; a més, el cost de desenvolupament d'aquesta solució sol ser més gran que el de les anteriors. Tot i així, podem obtenir les modificacions en el mateix moment en què s'han produït, sense necessitat de fer una anàlisi posterior.

d) Si la font de dades emmagatzema una empremta de temps per a cada modificació feta, es preferirà aquesta solució a la de comparació d'imatges, ja que és més immediata i el cost per obtenir les modificacions és molt menor. Tot i que el cost també pot ser menor en aquest cas que en casos anteriors, el fet que perdem moviments farà que aquesta solució sigui menys atractiva que aquelles que no presenten aquest inconvenient. Si volem recollir les operacions d'esborrament, haurem d'activar un mecanisme a les taules origen com pot ser l'esborrament lògic.

e) L'opció que se sol triar com a últim recurs és la de comparar imatges. Presenta els inconvenients que amb aquesta podem perdre moviments i a més el cost en temps d'execució pot ser bastant més elevat que el de les altres solucions, depenent de la mida de la font de dades. Tot i així, aquesta opció és immediata, no exigeix cap funcionalitat especial a les fonts de dades (sempre la podem implementar) i, per tant, és la triada habitualment. D'altra banda, moltes eines ETL disposen de components per realitzar aquesta comparació.

Exemple de problema d'obtenció de les modificacions de les dades

Volem obtenir les dades d'una font que no crea un fitxer delta, que està basada en una base de dades al fitxer de registre de la qual no es pot accedir i que no permet la definició de disparadors, a més, tampoc emmagatzema una empremta de temps amb les modificacions. En aquesta situació, l'alternativa més immediata és la de comparar imatges. Una altra possibilitat seria la de modificar la font de dades per a què generi un fitxer delta de modificacions; tot i així, això no sempre és factible (per exemple, si la font de dades és una aplicació empaquetada) o pot resultar massa costós, depenent de la mida de l'aplicació.

3.2. Construcció dels components de transformació, integració i depuració de dades

Quan ja hem obtingut les dades de les diferents fonts, és necessari preparar la informació abans de carregar-la al magatzem de dades. Per a això seguirem els passos següents:

a) Cada conjunt de dades pot tenir una estructura diferent depenent de la font de la qual procedeix. Les hem de transformar per adaptar-les a l'estructura de l'esquema del magatzem de dades en el qual s'emmagatzemaran: operacions de selecció de columnes, filtratge de registres, agregacions, etc.

b) Hem de depurar els errors o conflictes que puguem trobar dins de les dades de cadascuna de les fonts.

c) Hem d'integrar les dades amb la depuració d'errors o conflictes entre dades de fonts diferents.

d) Hem de crear les columnes derivades de la informació d'origen i que siguin necessàries per al magatzem de dades.

Com a resultat d'aquest procés, obtindrem un conjunt de dades directament utilitzable per actualitzar el magatzem de dades corresponent.

A continuació, comentarem alguns detalls d'aquestes operacions per separat. Això no significa que es facin de manera seqüencial, sinó que se'n poden combinar o intercalar algunes segons les necessitats.

3.2.1. Transformació de les dades

Les transformacions que cal fer sobre les dades per preparar la informació, segons s'ha comentat a l'apartat anterior, poden ser molt variades. Entre les més freqüents, trobem les següents:

- Canviar el format o el tipus de dades (per exemple, els camps de data).
- Canviar la codificació (per exemple, EBCDIC a ASCII).
- Reestructurar els camps (per exemple, fusionar o dividir camps o canviar el seu ordre relatiu).
- Canviar les unitats o els codis de representació (per exemple, canvis de moneda).
- Canviar el grau d'agregació (per exemple, calcular les vendes mensuals a partir de les diàries).
- Calcular camps derivats (per exemple, calcular l'edat a partir de la data de naixement).
- Generar claus subrogades: creació de claus subrogades (claus internes) que ens permetran guardar diferents versions d'una clau de negoci els atributs de la qual canvien en el temps.
- Discretitzar valors: variables contínues que discretitzem quan les classifiquem segons rangs de valors (per exemple, en el consum de clients, passem d'un valor continu a rangs de consum: 'alt', 'mitjà', 'baix').

- Encriptar camps per qüestions de seguretat.

- Afegir informació temporal (per exemple, període de validesa de les dades).

Una de les transformacions que generalment sempre s'ha de fer és l'última esmentada, afegir a les dades informació temporal. S'haurà d'afegir la informació sobre el període de validesa de les dades o el moment en el qual s'hagi registrat la modificació (o en el qual s'hagi detectat), segons sigui requerit pel magatzem de dades corresponent. D'aquesta manera, seqüenciarem les imatges obtingudes per anar formant la pel·lícula que emmagatzema el magatzem de dades.

Exemple d'informació temporal i dates de vigència

En un magatzem de dades d'una companyia d'assegurances, que guardi informació dels sinistres podem tenir la dimensió pòlisses, en la qual emmagatzemem el capital assegurat de la pòlissa, la província del prenedor, i altres dades associades a la pòlissa que poden canviar amb el temps, i que en cada canvi generaran un registre nou (és a dir, una nova versió) a la taula de pòlisses. D'aquesta manera, podem tenir d'una banda la taula amb les dades dels sinistres de les pòlisses que hem d'associar a la taula amb les dades pròpies de les pòlisses. Tenir versionades les dades de la taula de pòlisses permetrà associar a cada sinistre la versió corresponent de la pòlissa a la data corresponent (exemple, data de sinistre).

Les transformacions asseguren la correcta adequació de les dades d'origen a les taules del magatzem de dades.

3.2.2. Depuració de les dades

L'objectiu de depurar les dades obtingudes de les diferents fonts és millorar-ne la qualitat. Algunes de les incidències més comunes que es produeixen són les següents:

- Detectar i corregir valors inconsistents (per exemple, un atribut edat amb un valor de tres-cents cinquanta).

- Afegir valors per defecte als camps amb valors no definits. Generalment, es fa d'acord amb criteris marcats pel magatzem de dades al qual es destinen segons la font de dades. El valor subministrat pot ser constant, calculat o, en alguns casos, pot interessar deixar-ho sense definir.

- Detectar i corregir informació duplicada. A vegades és difícil de detectar, ja que es tenen diferents representacions del mateix valor (per exemple, diferents maneres d'escriure el nom d'un carrer a les dades del domicili). Serà més freqüent trobar informació duplicada entre diferents fonts de dades, però també la podem trobar dins d'una mateixa font.

- Detectar i corregir errors d'integritat referencial entre entitats relacionades (per exemple, no pot donar-se el cas d'incorporar al magatzem de dades

informació de vendes d'un producte que no existeix a l'entitat productes del nostre magatzem).

Caldrà definir l'acció que és necessari realitzar en el moment de detectar la incidència. Existeixen diferents opcions:

a) Rebutjar informació d'origen completa (per exemple, fitxer complet).

b) Rebutjar el registre erroni.

c) Corregir el registre erroni i inserir-lo al magatzem de dades.

Igualment, serà necessari activar el mecanisme de comunicació de la incidència detectada (alertes, notificació via correu electrònic, etc.).

Exemple de depuració de dades

És habitual que en un magatzem de dades tinguem falta de sincronització en l'actualització de dades que han d'estar íntegres entre si, com poden ser les dades de vendes i les dades de productes. Per exemple, podem tenir un codi de producte que rebem al fitxer de vendes i que no existeix a la taula mestra de productes. Una possible solució és integrar el registre amb valor «desconegut». En aquest cas, no rebutgem el registre del fitxer de vendes, l'integrem però amb valor «desconegut» o un codi que establim per al valor indeterminat d'aquesta dimensió (exemple: P9999). Aquest valor indeterminat, si existeix en el mestre de productes, és un registre que serveix per aglutinar tots els productes no classificats. El registre de vendes erroni s'actualitza al magatzem de dades i es visualitza en els informes, però classificat en un valor «desconegut», fins que el mestre de productes s'actualitzi i el fitxer de vendes es reprocessi. Aquesta solució ens permet carregar el registre, encara que no quedi ben classificat en els productes. Aquesta opció és millor que rebutjar el registre, ja que a l'hora de rebutjar perdem la informació i podríem tenir un percentatge de registres considerables afectats per aquesta incidència.

Així mateix, hem de decidir si les correccions sobre les dades errònies han de realitzar-se en el component de transformació i integració o en els sistemes d'origen. És preferible, que els errors se solucionin en origen, ja que si els solucionem en el component de transformació i integració l'error seguirà estant en origen i continuarem rebent informació errònia.

L'objectiu de la depuració de les dades és garantir la qualitat de les que incorporem al magatzem de dades.

3.2.3. Integració de les dades

Les dades provinents de diferents fonts s' han d'integrar entre si i amb les dades del magatzem de destinació. La manera d'aconseguir-ho pot variar.

El **procés d'integració** serà diferent depenent de si fem la càrrega inicial del magatzem de dades o en fem una actualització.

A més del volum de dades que cal tractar, la diferència principal rau en el fet que a les actualitzacions, per fer la integració, podem utilitzar les correspondències entre les dades de les fonts i les del magatzem de dades prèviament establertes en la càrrega inicial o en actualitzacions anteriors. Generalment, en el procés de càrrega inicial es farà una integració de totes les dades prèvia a la càrrega al magatzem de dades. D'altra banda, quan es fa l'actualització, és possible que no estiguin disponibles les dades de totes les fonts alhora i que interessi integrar les dades de les diferents fonts per separat al magatzem de dades.

El problema principal amb el qual ens trobem consisteix a detectar quines dades representen el mateix concepte.

Si les diferents fonts de dades utilitzen com a clau el mateix camp de l'entitat (per exemple, NIF), es poden relacionar sense dificultat, excepte per errors en les dades. El problema sorgeix quan cada font de dades empra la seva clau (per exemple, un codi generat) i no hi ha camps comuns que puguin servir com a clau alternativa per establir relacions entre si o, si n'hi ha, els seus valors es representen de manera diferent entre les fonts.

Durant el procés d'integració, es transformaran les dades per homogeneïtzar la seva representació i s'eliminarà la informació duplicada.

S'hauran d'establir els procediments adequats per propagar les correccions fetes fins als sistemes operacionals dels quals procedeixen les dades. Aquestes correccions seran especialment rellevants després d'obtenir les dades per a la càrrega inicial del magatzem de dades, però també s'hauran de tenir en compte les efectuades en cadascuna de les actualitzacions de les dades.

Dades depurades

Si hem dedicat un esforç considerable per integrar les dades de clients de diferents sistemes operacionals, depurant-los i eliminant duplicats, és raonable utilitzar les dades depurades en els sistemes operacionals, en lloc de continuar utilitzant-les amb errors. Per aquest motiu, s'haurà de definir un sistema per propagar les correccions fetes a les dades des del component d'integració i transformació fins als sistemes operacionals dels quals procedeixen.

3.3. Construcció del component d'actualització de les dades en els magatzems de dades

Quan ja hem obtingut les dades a partir de les fonts de dades, es transformen, depuren i integren si és necessari i, finalment, es transporten al magatzem de dades per procedir a carregar o actualitzar les dades que hi ha.

3.3.1. Mètodes d'actualització dels magatzems de dades

L'actualització de les dades dels diferents magatzems de dades es pot dur a terme de les maneres següents:

1) Càrrega: mitjançant l'operació de càrrega, el magatzem de dades de destinació passa a contenir exclusivament les dades que s'indiquen en aquesta operació. Si prèviament contenia altres dades, aquestes són substituïdes per les noves.

2) Addició: l'operació d'addició permet afegir les dades indicades al magatzem de dades. Podria donar-se el cas que algunes de les dades que es desitja afegir siguin ja al magatzem de dades. Si es produeix aquesta situació, es pot triar una de les dues alternatives següents:

a) Fusió destructiva: la fusió destructiva permet afegir les dades indicades a l'operació a les que ja hi havia prèviament d'una altra manera. Si la clau d'un registre de les dades indicades en l'operació coincideix amb la clau d'alguna de les dades existents, el registre previ és substituït pel registre nou. Els nous registres la clau dels quals no coincideix amb d'altres que ja hi havia s'afegeixen directament al magatzem de dades.

b) Fusió constructiva: la diferència respecte a la destructiva consisteix en el fet que, quan coincideix la clau d'algun dels registres que s'ha d'afegir amb la d'un registre existent, marca aquests registres però no els substitueix amb els valors nous. S'insereixen els nous registres i els anteriors queden marcats. D'aquesta manera, posteriorment es pot actuar sobre els registres de manera convenient.

3.3.2. Selecció del mètode d'actualització

Per carregar un magatzem de dades a partir de les obtingudes dels sistemes operacionals, utilitzarem una combinació de les operacions que hem estudiat a l'apartat anterior. Farem servir l'una o l'altra depenent de la situació de les dades que s'han de carregar i del magatzem.

La càrrega inicial de les dades es farà mitjançant una operació de càrrega, ja que partirem d'un magatzem de dades buit. Aquesta operació, fins i tot, ens permetrà crear de manera automàtica les taules del magatzem de dades, si no estaven creades anteriorment.

L'actualització de les dades es farà amb una combinació de les altres tres operacions (addició, fusió destructiva i fusió constructiva). Amb freqüència, n'hi haurà prou amb realitzar operacions d'addició, ja que anirem emmagatzemant noves imatges de les dades. Per tant, no trobarem dades duplicades.

En alguns casos, interessa actualitzar alguns camps de les dades existents a partir de les noves dades obtingudes. Aleshores, s'utilitzarà la fusió constructiva.

Exemple de fusió constructiva

Al magatzem de dades, per a cada registre s'indica mitjançant dos camps de tipus data (data d'inici, data final) el període de validesa de la resta dels seus camps. En afegir un registre nou, el camp data final roman no definit. Si es produeix una modificació sobre qualsevol camp del registre a la font de dades operacionals, el que farem serà afegir un nou registre al magatzem de dades amb els valors dels camps actualitzats (el camp data final no estarà definit). No obstant això, a més, haurem d'actualitzar el camp data final de la versió prèvia del registre per indicar que aquests valors ja no són vàlids. Aquesta operació d'actualització es durà a terme mitjançant l'operació de fusió constructiva.

De manera addicional a l'operació de càrrega inicial, es distingeixen dues formes d'actualització dels diferents tipus de magatzems de dades:

1) Refrescament total: totes les dades del magatzem de dades s'obtenen de nou i es carreguen mitjançant l'operació de càrrega. Aquesta situació es pot donar per carregar les dades dels magatzems de dades departamentals a partir del magatzem de dades corporatiu.

2) Manteniment incremental: es tracta de minimitzar el temps requerit per fer l'actualització, per a això es parteix de les dades existents i es busquen procediments per actualitzar-les segons les noves dades. Per fer el manteniment incremental, utilitzarem la resta de les operacions estudiades. Generalment, aquesta serà la manera d'actualitzar les dades en els diferents magatzems de dades.

Els processos d'actualització a les taules del magatzem de dades poden ser processos costosos en el cas de treballar amb grans volums de dades. Existeixen tècniques per optimitzar el rendiment d'aquests processos i les eines de suport solen tenir utilitats per implementar-los. Algunes d'aquestes tècniques es basen en la desactivació de restriccions de les taules o utilitats de la base de dades a fi de millorar els processos d'actualització (desactivació d'índexs, claus primàries, escriptura en log de la base de dades) i altres tècniques, a l'hora de dividir el conjunt de registres per actualitzar en diversos subconjunts i executar en diversos fils els processos d'actualització, posant al dia cada subconjunt en un fil independent.

3.4. Freqüència i finestra d'actualització

3.4.1. Freqüència d'actualització en un magatzem de dades

Cada magatzem de dades té uns requeriments d'actualització propis i, a més, aquests no necessàriament han de ser homogenis per a totes les dades dins d'un magatzem de dades.

Requeriments de freqüència d'actualització de diferents camps

En un magatzem de dades operacional d'un banc, disposem de les direccions dels clients i de les dades del saldo dels comptes. En el cas que es produeixi algun canvi en l'adreça d'un client en un sistema operacional, ens interessa que s'actualitzi al magatzem de dades operacional al final del dia, ja que és exclusivament llavors quan utilitzem les adreces per generar les etiquetes de la correspondència que s'envia als clients. Tot i així, qualsevol canvi en el saldo d'algun dels seus comptes registrat en una aplicació operacional interessa que s'actualitzi el més ràpid possible al magatzem de dades operacional, ja que s'utilitzarà aquest camp per determinar si s'autoritzen o no al client les operacions de crèdit fetes amb les seves targetes.

Per tant, per a cada magatzem de dades i cadascuna de les seves dades s'haurà de definir la freqüència d'actualització. Com més forts siguin aquests requeriments, el cost d'actualització serà més elevat. La tendència en les empreses és a tenir una dada cada vegada més actualitzada, la qual cosa implica una freqüència d'actualització cada vegada més gran. El que es busca és una actualització pròxima al temps real. Aconseguir-ho no és trivial i implica la utilització de tècniques avançades tant en la detecció de l'actualització com la seva propagació al magatzem de dades. D'altra banda, l'actualització pròxima al temps real pot generar confusió en l'analista de negoci respecte a l'actualització d'una dada determinada i pot trobar-se resultats diferents en executar un informe dues vegades al llarg del dia. En aquests casos, el que se sol fer és dividir la taula del magatzem en dues: la primera és una taula que s'actualitza en procés *batch* (partició estàtica) amb una periodicitat definida (per exemple diària) i la segona té una actualització pròxima al temps real (partició dinàmica). Aquesta segona taula conté els canvis que s'han produït des de l'última actualització de la partició estàtica i s'actualitza diverses vegades durant el dia. El fet de separar en dues particions té l'avantatge que l'actualització de la partició dinàmica (diverses vegades al dia) no afecta les consultes que es realitzen sobre la partició estàtica i que els usuaris tenen més clar el nivell d'actualització que es trobaran quan accedeixen a una partició o a l'altra.

Generalment, el magatzem de dades operacional és el que té uns requeriments més estrictes en aquest aspecte. En molts casos, l'ideal és que l'actualització sigui immediata, però això no sempre és factible. En aquest sentit, depèn de les possibilitats que ofereixin les fonts de dades.

El magatzem de dades corporatiu subministra les dades als magatzems de dades departamentals; per tant, en aquest sentit, els requeriments dels diferents magatzems departamentals quedaran reflectits en el magatzem de dades corporatiu.

Per a cada dada dels diferents magatzems de dades s'ha de definir la freqüència d'actualització: cada quant s'ha d'actualitzar a partir de les fonts de dades.

3.4.2. Finestra d'actualització del magatzem de dades

L'objectiu principal dels magatzems de dades és donar suport al procés de presa de decisions. Per tant, les operacions que se solen realitzar al respecte són de consulta. Per optimitzar l'execució d'aquest tipus d'operacions, els SGBD sobre els quals estan implementats els magatzems de dades estan configurats de manera que només permeten fer operacions de consulta.

Si es vol fer una altra tipus d'operacions diferents de les de consulta –per exemple, les d'actualització del magatzem de dades–, cal parar els sistemes, canviar la configuració, fer les operacions necessàries i restaurar la configuració perquè les consultes continuïn essent òptimes. Durant aquest temps, el magatzem de dades no està disponible per als seus usuaris.

Disponibilitat del magatzem de dades corporatiu

En el cas extrem de requerir una disponibilitat total (vint-i-quatre hores al dia cada dia de l'any), una possible solució és tenir la base de dades del magatzem de dades replicada i actualitzar una de les còpies mentre s'utilitza l'altra. Una vegada actualitzada aquesta, es pot utilitzar mentre s'actualitza l'altra.

La **finestra d'actualització** d'un magatzem de dades és el temps necessari per fer les operacions que l'actualitzen. Durant aquest període de temps, el magatzem de dades no està operatiu per realitzar consultes.

Relacionat amb la finestra d'actualització, tenim el concepte de la finestra d'extracció; aquesta fa referència al temps necessari per obtenir les modificacions de les dades a partir de les fonts de dades.

La **finestra d'extracció** és el temps necessari per obtenir els moviments a partir de les fonts de dades.

Segons el mètode d'obtenció de les modificacions, la finestra d'extracció serà més o menys àmplia. Generalment, el mètode d'obtenció de modificacions a través de comparació d'imatges és el que requereix una finestra d'extracció més àmplia, ja que les operacions que ha de fer consumeixen molt temps. A més, juntament amb el mètode d'obtenció de modificacions basat en fonts amb empremta de temps, requereix que es faci un procés a les plataformes de les fonts de dades. Això últim pot resultar problemàtic si aquestes plataformes estan sobrecarregades de treball.

En el cas de la resta dels mètodes, n'hi ha prou amb obtenir el fitxer de registre o els respectius fitxers de moviments i analitzar-los en una plataforma diferent. En el mètode de comparació d'imatges, la comparació també es pot dur a terme en una altra plataforma, però l'extracció de cadascuna de les imatges

(només se n'ha d'extreure una cada vegada) s'ha de fer a la plataforma de la font de dades. Finalment, si s'usa el mètode basat en l'empremta de temps, generalment l'anàlisi de les dades s'ha de fer a la plataforma que les conté.

El component d'integració i transformació ha de tenir presents els requeriments de disponibilitat de les fonts de dades, així com els dels magatzems de dades. Per aquest motiu, és molt important minimitzar el temps de procés dels diferents passos, com s'ha assenyalant quan s'han presentat, de manera que es puguin executar dins de les finestres d'extracció i actualització disponibles en cada cas.

Exemple de finestra d'actualització

És molt habitual la realització de processos d'actualització del magatzem de dades en processos *batch* nocturns. Per exemple, en un banc podem realitzar l'actualització dels moviments dels clients en un procés *batch* diari, de manera que els usuaris cada dia tenen la «foto» de tancament del dia anterior. Aquesta finestra d'actualització no afecta les consultes que es realitzen durant el dia i el refrescament de les dades s'ajusta molt bé al ritme natural del negoci (moviments diaris).

3.5. Eines de suport al desenvolupament

Les necessitats dels diferents usuaris canvien amb el temps, especialment les dels analistes; per tant, és habitual que es produeixin canvis tant a les fonts de dades com als magatzems de dades. Per les seves característiques d'element intermediari entre la resta dels elements de la FIC, el component d'integració i transformació es veurà afectat pels canvis en qualsevol dels altres elements. Per tant, a més de complir amb les funcions que hem explicat abans, aquest component ha de ser prou flexible perquè es pugui adaptar als canvis que es produeixin en qualsevol dels components amb els quals interacciona i, a més, ho ha de fer de manera immediata.

El component d'integració i transformació està format principalment per programari. Per desenvolupar qualsevol component de programari, es presenten dues alternatives:

- Desenvolupament manual.
- Desenvolupament automàtic o amb suport automàtic.

3.5.1. Funcionament de les eines

Si el que necessitem fonamentalment en el component d'integració i transformació és reaccionar als canvis de manera immediata, la solució més adequada serà dur a terme un desenvolupament amb suport automàtic. Com que aquesta necessitat és àmpliament reconeguda, al mercat hi ha eines orientades de manera especial al desenvolupament d'aquest component. Aquestes eines ofereixen un conjunt de transformacions tipus de les dades, així com altres funcionalitats específiques per suportar les operacions requerides.

Aquest tipus d'eines permeten implementar el component de transformació i integració utilitzant una interfície gràfica que ens permet definir des de les fonts origen tot el flux de procés fins a la taula destí del magatzem de dades. Tots els mapatges, transformacions i abocaments definits són guardats en un repositori de metadades. Se sol treballar en dos nivells: de flux de procés, entenent-lo com a treball, i de transformació, com a pas d'aquest flux. Les configuracions es realitzen a aquests dos nivells. Gairebé totes les eines guarden certa similitud en la manera de dissenyar els fluxos. Existeixen moltes transformacions estàndard (lectura, filtratge, unió de conjunts de dades, ordenacions, agregacions, actualització de taules, etc.) habituals en el component de transformació i integració que venen predefinides a les eines i l'únic necessari és parametritzar-les.

Exemple de pas predefinit en eina

És molt habitual crear passos de lectura en el component de transformació i integració. Aquesta lectura es pot realitzar des de fitxer pla, Excel, XML, base de dades, etc. Les eines incorporen diferents connectors per obtenir dades de moltes tipologies de fonts. Per exemple, per llegir dades d'una taula d'un SGBD determinat, en el pas predefinit de lectura de BBDD cal indicar l'SGDB, el servidor on s'allotja la base de dades, el nom de la base de dades, les credencials i compondre la consulta que recuperarà les dades.

3.5.2. Avantatges i inconvenients de les eines

Tant el fet d'utilitzar una eina com de dur a terme un desenvolupament manual presenten una sèrie d'avantatges i inconvenients. Atesa la situació de cada organització, s'haurà de determinar la solució més adequada.

Els **avantatges** de fer el desenvolupament manual són els següents:

- Podem començar el desenvolupament immediatament.
- Sigui com sigui el nostre entorn, els desenvolupaments manuals es poden adaptar a qualsevol tipus de situació.
- Més flexibilitat a l'hora d'implementar lògica de negoci complexa.
- Més facilitat per trobar perfils que coneguin el llenguatge que s'ha d'utilitzar.

I els **inconvenients** són aquests:

- Hi ha molts programes que és necessari construir, i tots tenen una estructura semblant: lectures, filtratges, encreuaments, unions, etc. Es pot plantejar l'opció de reutilitzar codi creant funcions, però aquestes també cal desenvolupar-les i mantenir-les.
- Les metadades associades als programes han de desenvolupar-se de manera explícita, com una tasca addicional, per la qual cosa presenten el mateix problema que la majoria dels desenvolupaments manuals. Directament no

es desenvolupen les metadades per reduir el cost o el termini d'execució o, si es desenvolupen, no s'actualitzen amb els canvis que es van produint.

- Els programes requereixen bastant temps per desenvolupar-se.

- Els requeriments canvien constantment i els programes s'hi han d'adaptar. Per tant, les modificacions són molt freqüents i solen ser costoses. No existeixen metadades que ajudin a realitzar l'anàlisi d'impacte d'un canvi.

- És complicat realitzar un llinatge de dades per conèixer, partint d'una dada final, totes les transformacions realitzades. Cal anar al codi per obtenir la traçabilitat de la dada.

- És més difícil marcar unes bones pràctiques i uns estàndards a l'equip de desenvolupament. És habitual, com passa en qualsevol altra aplicació, trobar-nos amb processos difícils de mantenir.

- El cost total de desenvolupament és molt alt.

Els principals **avantatges** de l'ús d'eines són els següents:

- Els programes d'extracció i transformació es poden construir ràpidament, sobretot en aquells passos comuns molt estesos.

- Gràcies al fet que emmagatzemen les definicions fetes en forma de metadades, els programes generats es poden mantenir ràpidament i fàcilment.

- Les metadades associades es produeixen i es mantenen de manera automàtica.

- És possible realitzar un llinatge de dades per conèixer, partint d'una dada final, totes les transformacions que ha sofert, tenint en compte la informació guardada a les metadades.

- És possible realitzar l'anàlisi d'impacte per conèixer les transformacions afectades per un canvi en la informació origen.

- Els desenvolupaments tenen més portabilitat. En el cas de canviar de plataforma de font o de destinació de les dades, les metadades que defineixen les correspondències i transformacions continuen essent vàlides. N'hi ha prou amb modificar els paràmetres dels passos de transformació, tenint en compte els nous paràmetres de les plataformes de destinació.

- És més senzill marcar unes bones pràctiques i uns estàndards a l'equip de desenvolupament. Els desenvolupaments són més estàndard i es mante-

nen més bé perquè queden guardats en fluxos de procés més senzills de seguir.

- Els costos de desenvolupament es redueixen de manera significativa.

D'altra banda, les eines del mercat presenten diferents **inconvenients**:

- Estan preparades per ser utilitzades en entorns molt generals; si el nostre entorn té característiques particulars, com és habitual que passi, cal adaptar-les. En el cas de suportar-la, aquesta operació pot ser molt complexa i costosa.

- És precís dedicar un període a la formació dels desenvolupadors que les utilitzaran.

- Presenten limitacions a l'hora d'implementar lògica de negoci complexa.

- El seu preu és molt elevat en el cas d'eines líders. Existeix l'opció d'emprar eines *open source* que cada vegada presenten un grau de maduresa més gran. Tot i així, el seu cost es compensa amb l'esforç de desenvolupament que estalvien.

Per tant, per començar a desenvolupar un magatzem de dades, la solució que solen adoptar moltes organitzacions és desenvolupar el programari del component d'integració i transformació de manera manual (o amb el suport automàtic de qualsevol altre desenvolupament). Aquesta solució pot resultar adequada al principi, per construir algun magatzem de dades departamental. No obstant això, en àmbits més amplis o per a diferents magatzems de dades, els costos de desenvolupament i les limitacions pel que fa al temps de resposta davant de canvis es disparen. Per aquest motiu, és recomanable adquirir una eina de suport des del principi, i planificar el temps necessari per adaptar-la al nostre entorn. El preu de les llicències és un factor que cal considerar, però aquesta inversió pot suposar un estalvi en terminis i sempre podem avaluar l'opció de programari lliure, ja que les eines de suport a l'ETL d'aquest tipus han evolucionat molt en els darrers anys.

És convenient utilitzar una eina com a suport al desenvolupament del component d'integració i transformació, ja que es requereix que sigui molt flexible i s'adapti ràpidament als canvis produïts en la resta dels components de la FIC.

3.5.3. Altres eines de suport

A més de les eines anteriors, hi ha altres eines especialitzades per oferir suport a les operacions de depuració i integració d'alguns tipus de dades, com, per exemple, les que fan referència a noms i domicilis. Aquestes eines estan basades en diccionaris específics per a cada idioma o entorn en el qual emmagatzemen noms de persones, cognoms, noms d'empreses, noms de carrers, etc., així com diferents maneres de representar-los i abreujar-los o maneres errònies que a vegades s'utilitzen. Generalment, usen patrons per reconèixer els diferents tipus d'ocurrència que es produeixen i d'aquesta manera identificar els sinònims.

3.6. Rendiment del component de transformació i integració

Un aspecte important en el desenvolupament del component de transformació i integració és el relatiu al rendiment dels processos. Generalment, aquest component tractarà un volum de dades alt i les necessitats d'actualització i disponibilitat de la informació són exigents, de manera que optimitzar la finestra d'execució d'aquest component serà un aspecte a considerar des del moment del seu disseny.

Alguns aspectes relatius al disseny que ens poden ajudar a millorar els temps de procés són els següents:

- Filtres de registres i columnes: quedar-nos en els primers passos amb les dades estrictament necessàries, evitant arrossegar a la resta del procés columnes o registres innecessaris.

- Transformacions en memòria: realitzar la major part de les transformacions en memòria, si és possible. Evitar processos de lectura i escriptura en disc que són costosos en temps.

- Transformacions en base de dades: en el cas de disposar d'una base de dades amb alta capacitat de procés i arquitectura optimitzada per a un rendiment òptim, ens podem plantejar dur les dades a la base de dades des de l'extracció i realitzar les transformacions en base de dades. Aquest tipus d'arquitectura es denomina ELT (sigles del terme en anglès *Extract Load Transform*).

- Ús d'objectes de la base de dades per optimitzar els processos d'extracció: definir índexs i mantenir estadístiques de base de dades actualitzades sobre les taules origen sobre les que s'executin extraccions pesades.

- Revisió d'operacions pesades: algunes operacions com les agregacions i ordenacions són costoses en recursos i temps. Minimitzar-ne l'ús a l'estrictament necessari.

- Paral·lelització de processos: en el cas que les operacions a realitzar siguin paral·lelitzables i disposem d'una arquitectura que afavoreixi el paral·lelisme.

- Cerca del mètode òptim d'obtenció d'actualitzacions (apartat 3.1.2).

4. Construcció del magatzem de dades: departamental, corporatiu i operacional

4.1. Construcció del magatzem de dades corporatiu

El magatzem de dades corporatiu ofereix una visió integrada i historiada de les dades de l'organització. La seva missió és emmagatzemar les dades per subministrar-les als diferents magatzems de dades departamentals.

En aquest apartat, estudiarem diferents aspectes relacionats amb el desenvolupament del magatzem de dades corporatiu. Repassarem breument el que ja hem estudiat relacionat amb aquest punt i aprofundirem en un element que encara no hem estudiat: el model de dades del magatzem de dades corporatiu.

4.1.1. Revisió del procés de desenvolupament

Segons l'estratègia presentada en aquest mòdul, el magatzem de dades corporatiu es desenvoluparà de manera gradual. En un primer moment, mitjançant el projecte global de desenvolupament tindrem una visió general de les dades que contindrà el magatzem de dades corporatiu. Per tant, dissenyarem el seu esquema a un alt nivell i definirem les entitats que hi apareixeran, sense entrar en el detall dels atributs concrets. Així mateix, podrem fer una previsió del volum de dades que arribarà a tenir i determinar quina màquina i quin SGBD el podran suportar.

Mitjançant el projecte de desenvolupament d'infraestructura obtindrem la màquina i instal·larem i configurarem l'SGBD.

El disseny inicial de l'esquema del magatzem de dades corporatiu guiarà el seu desenvolupament. Ho refinarem amb el desenvolupament dels successius projectes autònoms en els quals dividim la construcció de la FIC. El treball fet en aquests projectes estarà supervisat per l'arquitecte de la FIC, juntament amb l'administrador de bases de dades responsable de l'entorn. A mesura que anem posant en marxa aquests projectes, el magatzem de dades corporatiu s'anirà poblant de dades.

Construirem el magatzem de dades corporatiu de manera gradual. Partim d'un disseny inicial que refinarem amb els diferents projectes de desenvolupament de magatzems de dades departamentals que també el poblaran de dades.

4.1.2. El model de dades del magatzem de dades corporatiu

Els magatzems de dades departamentals estan orientats a l'accés de les seves dades per part dels analistes, i per aquest motiu estan dissenyats segons el model de dades multidimensional, que ha estat definit per permetre fer consultes complexes de manera simple. Tot i així, la missió principal del magatzem de dades corporatiu és emmagatzemar i subministrar les dades necessàries als magatzems de dades departamentals, és a dir, està orientat a l'emmagatzematge de les dades. Segons quin model de dades el dissenyarem?

Una possibilitat consisteix a utilitzar el model de dades multidimensional de la mateixa manera que ho fem per als magatzems de dades departamentals. Aquest model s'ha pensat de manera exclusiva per fer consultes centrades en els fets (els fets són el focus d'atenció i el model permet emmagatzemar-ne dades històriques molt fàcilment); tot i així, no presenta aquesta facilitat per emmagatzemar dades històriques sobre les dimensions que qualifiquen els fets.

Exemple de fets i dimensions

En un magatzem de dades departamental de vendes, les dades relatives a les transaccions de vendes seran els fets (esdeveniments de negoci) i les dades de productes, clients o centre de vendes seran dimensions (eixos per classificar els fets).

Encara que puguem fer consultes sobre el magatzem de dades corporatiu, aquest no és el seu principal objectiu. Haurem d'utilitzar un model que sigui més adequat per a l'emmagatzematge de dades, especialment de dades històriques.

Una altra possibilitat és utilitzar un model conceptual dels que fem servir per construir sistemes operacionals: entitat/relació, orientat a objectes, etc. Aquests models ens permetran definir les entitats que apareixen i les relacions que hi ha entre elles. Amb aquests models, durem a terme dissenys semblants als que fem per als sistemes operacionals.

La diferència substancial apareixerà a l'hora d'implementar els esquemes dissenyats, és a dir, a l'hora de situar-nos en l'àmbit lògic. Per exemple, si emprem el model relacional, per construir un sistema operacional aplicaríem les regles de normalització per obtenir un esquema normalitzat. Un dels objectius de normalitzar és evitar tenir redundància de dades. D'aquesta manera, si hem de modificar les dades corresponents a una entitat, n'hi haurà prou amb fer-ho en un sol lloc (ja que només estan emmagatzemades en una taula i referenciades per d'altres). En el magatzem de dades corporatiu no farem modificacions sobre les dades emmagatzemades: només emmagatzemarem imatges successives de les dades. Per tant, normalitzar per evitar problemes amb les modificacions no té raó de ser. D'aquesta manera, podrem fer dissenys no normalitzats, si s'adapten millor a l'estructura de les dades que cal emmagatzemar.

Per implementar el magatzem de dades corporatiu, podem utilitzar el model relacional. Tot i així, no caldrà normalitzar el disseny de l'esquema, ja que no farem modificacions sobre les dades emmagatzemades.

També podrem fer la implementació utilitzant el model orientat a objectes; en aquest cas, la implementació estarà més a prop del disseny conceptual realitzat. Un dels avantatges principals de les bases de dades orientades a objectes és que permeten associar directament les dades amb el codi que les tracta, particularment per fer modificacions sobre aquestes. D'altra banda, ofereixen més facilitat que les bases de dades relacionals per implementar esquemes molt complexos. En el magatzem de dades corporatiu, els esquemes dissenyats no seran especialment complexos. A més, no tindrem operacions de modificació associades a les dades. Així doncs, sembla que en implementar el magatzem de dades corporatiu en una base de dades orientada a objectes no s'aprofiten els avantatges que ofereix aquest tipus de base de dades.

No hi ha res en contra d'implementar el magatzem de dades corporatiu sobre una base de dades orientada a l'objecte. Malgrat això, no podem aprofitar els avantatges principals que ofereix aquest tipus de base de dades ateses les característiques del sistema que s'ha de construir.

4.1.3. Transformacions per construir l'esquema del magatzem de dades corporatiu

En aquest apartat, estudiarem el conjunt de transformacions que aplicarem als esquemes de les fonts de dades operacionals per implementar l'esquema del magatzem de dades corporatiu. Construirem l'esquema en qualsevol dels models lògics esmentats a l'apartat anterior.

L'objectiu d'aquestes transformacions és doble:

- D'una banda, necessitem que l'esquema dissenyat permeti reflectir l'evolució produïda en les dades i les seves relacions, és a dir, que emmagatzemi la història de les dades.

- D'altra banda, com que només farem consultes sobre les dades, l'objectiu és que el disseny estigui optimitzat perquè consumeixin el menor temps possible. Tot això tenint present que el disseny de l'esquema pot estar desnormalitzat.

Model del magatzem de dades corporatiu

A Silverston, Inmon i Graziano (1997), *The Data Model Resource Book*, la primera tasca que es proposa per definir el model del magatzem de dades corporatiu a partir d'un model de dades corporatiu operacional és la d'eliminar les dades operacionals que es cregui que no seran utilitzades pels analistes. És així perquè proposen construir el model del magatzem de dades corporatiu en un sol pas, centrant-se en l'emmagatzematge de les dades.

Nosaltres el construirem en diferents passos, un per a cada projecte autònom que desenvolupem, i en cada pas definirem de manera exclusiva al magatzem de dades corporatiu les dades que necessitin els analistes del magatzem de dades departamental associat al projecte autònom. Per tant, no hem d'eliminar dades operacionals de manera explícita, ja que inclourem exclusivament les dades que necessitem per al projecte.

Addició d'un element de temps

Sempre haurem d'afegir com a mínim un element de temps a cada unitat de dades al magatzem de dades corporatiu. L'objectiu principal és expressar d'alguna manera el període de validesa de les dades en l'entorn operacional, ja que aquest és el temps que requereixen els analistes per estudiar l'evolució de les dades.

A més del període de validesa, podrem disposar d'altres dades relacionades amb el temps; per exemple, el moment de l'extracció de les dades, de la seva càrrega, de la disponibilitat per a l'analista, etc.

En alguns casos, podrem obtenir aquest temps de l'entorn operacional, ja que haurà quedat registrat el moment en el qual s'han produït els canvis (és així en tots els mètodes d'extracció de dades immediates; en el d'empremta de temps també, malgrat que en aquest cas podem haver perdut moviments). En d'altres, l'haurem d'afegir amb el component d'integració i transformació.

Podem expressar el període de validesa de les dades de diferents maneres, entre les quals les més freqüents són les següents:

- Utilitzant dos camps, un per indicar el moment d'inici i un altre per al moment final.

- Mitjançant un sol camp que indiqui el moment d'inici. El moment final de validesa correspondrà al moment d'inici d'un altre valor per a les mateixes dades, en el cas que n'hi hagi.

- Utilitzant un sol camp que indiqui el període de validesa (per exemple, les dades del mes de gener d'enguany).

Haurem d'afegir elements de temps a diferents graus de granularitat de les dades, si ho requereixen.

- Des del punt de vista del fitxer: totes les dades d'un fitxer corresponen al període de validesa indicat. Per exemple, emmagatzemem les dades de cada mes en un fitxer diferent.

- Des del punt de vista del registre: per a cada registre d'un fitxer podem indicar el període de validesa com a part de la clau del registre.

- Des del punt de vista del camp: cada camp d'un registre té associats els camps necessaris per expressar el període de validesa.

Així mateix, els camps que afegirem per indicar el temps tindran granularitats diferents, segons es requereixi. En alguns casos n'hi haurà prou amb indicar l'any i en d'altres s'haurà d'indicar el temps en segons.

Cada dada del magatzem de dades corporatiu que ho requereixi ha de tenir associat un període de validesa. D'aquesta manera, els analistes poden estudiar-ne l'evolució.

Organització de dades segons la seva estabilitat

Hem de tenir present que emmagatzemarem l'evolució de totes les dades de l'organització en una base de dades, el magatzem de dades corporatiu. Si no prestem especial atenció a l'espai requerit per les solucions que proposem, la mida total del resultat es pot disparar. Concretament, en relació amb l'apartat anterior, hem de prestar especial atenció a quin grau de granularitat afegim a un element de temps per emmagatzemar la història.

Estabilitat i grau de granularitat de les dades

En un banc, hem dissenyat l'entitat client al magatzem de dades corporatiu mitjançant una taula en la qual tenim les dades personals del client (adreça, telèfon, etc.) juntament amb la suma del saldo total dels seus comptes i dos camps de temps que indiquen el període de validesa de les dades (granularitat de les dades des del punt de vista del registre). D'aquesta manera, cada vegada que canviï alguna de les dades del client, hem d'emmagatzemar un nou registre amb les noves dades. Si el saldo canvia de mitjana una vegada al dia per a cada client, cada dia emmagatzemarem una nova versió de cada registre complet, incloses les dades personals que canvien una vegada cada cinc anys de mitjana. És a dir, malgastem molt espai, perquè hem emmagatzemat 1.825 vegades (365 × 5) la mateixa adreça.

Aquest mateix disseny en un sistema operacional pot presentar altres problemes, tot i que no d'espai. En aquest cas, si un atribut d'un registre canvia, es perd el valor anterior i no s'emmagatzema una versió completa del registre.

En l'exemple anterior, les dades del registre de client tenen una estabilitat diferent. Tot i així, hem definit la granularitat d'emmagatzematge de canvis des del punt de vista del registre, sense tenir en compte aquest fet. Per optimitzar la quantitat de dades emmagatzemades, hauríem d'haver considerat, d'una banda, el camp saldo i, de l'altra, les dades personals, ja que tenen estabilitats diferents. Si implementéssim aquest exemple en una base de dades relacional, hauríem de tenir en una taula les dades personals i en una altra el saldo, i definir en els dos casos la granularitat des del punt de vista del registre, perquè no sabem el nombre màxim de canvis que pot tenir el saldo i, per aquest motiu, no podríem tenir la granularitat des del punt de vista del camp.

Per optimitzar l'espai ocupat al magatzem de dades corporatiu, dins de cada entitat hem de definir la granularitat de les dades a les que afegim un element de temps segons el seu grau d'estabilitat davant dels canvis.

Addició de dades derivades

Els models operacionals generalment no inclouen dades derivades; és a dir, aquelles obtingudes a partir d'altres dades emmagatzemades a la base de dades. No té gaire sentit emmagatzemar dades derivades en els casos en què els valors base dels càlculs poden canviar amb freqüència, per la qual cosa s'haurien de recalcular cada vegada que es produís un canvi.

Els principals **avantatges** d'emmagatzemar dades derivades a la base de dades són els següents:

- Es té més velocitat d'accés a aquestes dades, ja que no cal calcular-les cada vegada que es necessiten.

- S'eviten errors. Si directament proporcionem les dades, evitem que algú les calculi de manera errònia.

Els **inconvenients** són els següents:

- Si les dades base canvien, s'han de calcular les dades derivades a partir d'aquestes.

- S'ocupa més espai d'emmagatzematge.

Per definició, les dades emmagatzemades al magatzem de dades corporatiu no canvien. Per aquest motiu, si incloem dades derivades, excepte situacions d'error, no s'hauran de recalcular i no tindrem aquest inconvenient. D'aquesta manera, el principal inconvenient d'incloure-les és el problema de l'espai d'emmagatzematge que requereixen.

Els avantatges que aporta incloure dades derivades a les quals accedeixen els analistes són més elevats que el cost d'emmagatzematge, especialment si calcular-los és complex i costós.

D'altra banda, també podem emmagatzemar aquestes dades derivades en els diferents magatzems de dades departamentals, la qual cosa serà el mitjà d'accés per a la majoria dels usuaris de la FIC.

Calcularem les dades derivades que els analistes necessitin i les emmagatzemarem al magatzem de dades corporatiu o bé directament als magatzems de dades departamentals.

El concepte denominat per Inmon **índex creatiu** està associat a les dades derivades. El plantejament és el següent: atès que per passar les dades des de les fonts de dades operacionals al magatzem de dades corporatiu hem de treballar amb les dades al nivell més baix. Amb una mica de treball addicional podríem calcular a partir d'aquestes una sèrie de dades interessants per als analistes. És a dir, precalculem requeriments dels analistes que podem anticipar i els emmagatzemem al magatzem de dades corporatiu.

A diferència de les dades derivades comentades anteriorment, les dades calculades per a l'índex creatiu són aquelles que no es consideren en l'entorn de les aplicacions operacionals: les obtingudes de manera exclusiva per als analistes.

Exemple d'índex creatiu

Per exemple, en un banc, els comptes que menys activitat tenen, els ingressos més alts, etc.

Canvi de granularitat en les dades

A l'hora de dissenyar el magatzem de dades corporatiu, podem tenir tendència a incloure totes les dades disponibles al nivell de detall més baix que puguem obtenir de les fonts de dades operacionals. D'aquesta manera, intentem preveure necessitats futures que poden ser molt improbables. El resultat és que l'espai requerit es pot disparar.

El magatzem de dades corporatiu ha de cobrir les necessitats actuals dels usuaris. Abans d'afegir qualsevol dada que no es necessiti, hem de valorar les repercussions de cost que això pot tenir.

Concretament, pel que fa a la granularitat, hem d'adaptar el nivell de detall de les dades que obtenim de les fonts de dades operacionals al requerit pels usuaris.

Adaptació de la granularitat de les dades dels clients

En l'exemple del banc, per a cada client els analistes requereixen estudiar l'evolució del seu saldo mensual. Particularment, utilitzen per a cada mes els valors del saldo màxim, mínim i mitjà. A partir del sistema operacional, podem obtenir tots els valors que pren el saldo al llarg de cada dia. Si els analistes coneixen la possibilitat de tenir les dades amb un nivell més baix, però consideren que només necessiten les dades d'àmbit mensual per emmagatzemar el saldo de cada client, obtindríem els valors requerits pels analistes agregant tots els valors obtinguts per a cada mes, i serien exclusivament aquests valors calculats els que emmagatzemaríem al magatzem de dades corporatiu.

En tot moment hem de tenir present que el nivell de granularitat definit suposarà una restricció en l'anàlisi de dades, en el sentit que per sota d'aquest nivell de granularitat no podrem analitzar dades i que afegir aquest nivell més

baix *a posteriori* és molt costós. Tenint present això, arribarem a un nivell que satisfaci les necessitats dels analistes, però que no té per què ser el mateix que el dels sistemes operacionals.

Amb l'objectiu de minimitzar l'espai ocupat, el magatzem de dades corporatiu ha d'emmagatzemar les dades amb la granularitat que necessiten els analistes, no amb la granularitat més baixa que podem aconseguir de les fonts de dades operacionals.

Fusió d'entitats

Aquesta fusió es pot fer per diferents motius:

- Les entitats provenen de diferents sistemes operacionals que s'integren en una entitat al magatzem de dades corporatiu, ja que conceptualment són la mateixa.

- En els sistemes operacionals, tenim les dades organitzades perquè suportin de manera òptima les operacions de modificació, generalment dividint-les en entitats més petites. Quan passem al magatzem de dades corporatiu podem agrupar aquestes entitats, ja que no tindrem problemes amb les modificacions. Concretament, com que no es produeixen canvis en les dades, sabem el nombre d'ocurrències que tenim i així podem definir atributs per contenir múltiples ocurrències. D'aquesta manera, es facilita l'accés a les dades.

- Pot resultar més fàcil representar l'evolució de les relacions entre dues entitats si les emmagatzemem juntes i definir una sola entitat a partir d'aquestes. En aquest cas, fusionar entitats pot representar replicar dades.

Al magatzem de dades corporatiu podem tenir les dades desnormalitzades i replicades, ja que no es produiran modificacions. Mitjançant la fusió d'entitats operacionals es facilita l'accés a les dades, atès que s'hi pot accedir directament en lloc d'haver de seguir les relacions definides entre les dades per obtenir-les. En el cas de replicar dades, hem d'avaluar el cost que representa en espai addicional requerit.

4.2. Construcció del magatzem de dades departamental

La construcció dels magatzems de dades departamentals difereix notablement de la construcció d'un magatzem de dades corporatiu, atès que els magatzems departamentals cobreixen les necessitats d'un grup d'analistes, no de tota l'organització. D'altra banda, els magatzems departamentals no s'alimenten

directament de les fonts de dades, sinó del magatzem de dades corporatiu. Del conjunt de fonts de dades disponibles només s'utilitzaran les que siguin d'interès per al grup d'analistes del departament.

Els dos tipus de magatzems gestionen volums de dades d'ordre de magnituds diferents. Un magatzem de dades departamental conté un conjunt de dades relacionades amb un tema o visió concreta de l'organització (departament). A més, per al tipus d'anàlisis que es realitzen, no és habitual tenir la necessitat de carregar el màxim nivell de detall de les dades. Podem entendre que el magatzem de dades departamental conté un subconjunt agregat de les dades disponibles al magatzem de dades corporatiu. Per contra, el magatzem de dades corporatiu conté una versió consolidada de les dades de tots els magatzems de dades departamentals. Moltes vegades es desconeix *a priori* el tipus d'anàlisi que es realitzarà en el futur, per la qual cosa és habitual carregar un major nivell de detall, per cobrir futures necessitats d'anàlisi.

4.2.1. Disseny del model i aprovisionament de dades

El disseny de model de dades sobre el qual crearem el magatzem de dades departamental serà desnormalitzat, ja que el seu objectiu primordial és la consulta, no la modificació.

Vegeu també

El tipus de disseny es basarà en el model dimensional que es veurà detalladament al mòdul «Disseny i implementació multidimensional de dades» d'aquesta assignatura.

Pel que fa a l'aprovisionament de dades, els processos que carreguin el magatzem de dades departamental seran més senzills que el component de transformació i integració del magatzem de dades corporatiu, atès que en aquest cas la càrrega de dades es realitza des del magatzem de dades corporatiu on les dades ja estan depurades i integrades.

En el procés de càrrega des del magatzem corporatiu al departamental els passos que cal fer seran de tipus:

- Selecció de columnes i registres d'interès per al magatzem departamental.
- Agregacions des de les dades del magatzem corporatiu buscant el nivell de granularitat adequat; generalment, el nivell de detall del magatzem corporatiu serà més gran.
- Transformacions necessàries per adequar les dades a l'esquema multidimensional del magatzem departamental.

4.2.2. Enfocament del projecte

Des del punt de vista de l'execució del projecte, l'envergadura del projecte de creació d'un magatzem de dades departamental serà sempre inferior a un magatzem de dades corporatiu i el seu termini d'execució, molt menor.

En el cas de plantejar la construcció d'un magatzem de dades departamental sense existir prèviament el magatzem de dades corporatiu, serà bona pràctica dur-lo a terme com a projecte autònom dins de la FIC.

Vegeu també

El projecte autònom dins de la FIC s'ha tractat a l'apartat 2.4.3 d'aquest mòdul.

Un altre possible enfocament en la creació del magatzem de dades departamental és la creació d'un *virtual data mart*, que consisteix en una capa lògica sobre el magatzem de dades corporatiu que ofereix una visió parcial de les dades necessàries per als usuaris del magatzem departamental. Aquest enfocament evita el moviment de dades, es realitza en menys temps, però té limitacions quan tractem amb grans volums de dades, ja que hi ha moltes dades calculades «al vol» i a vegades les consultes poden ser lentes.

4.3. Construcció del magatzem de dades operacional

De manera tradicional, l'entorn operacional s'ha estructurat en forma d'aplicacions independents. Per adaptar-se als requeriments de l'entorn, les organitzacions necessiten tenir una visió integrada de les seves dades.

Aplicacions no integrades en un banc

Un banc disposa d'aplicacions independents per gestionar les targetes de crèdit, crèdits hipotecaris, crèdits personals, comptes d'estalvi, etc. Cada aplicació funciona a la seva plataforma, tot i que els treballadors poden accedir-hi des del seu terminal. Si un usuari necessita conèixer tota la informació associada a un client (per exemple, el director d'una oficina o l'empleat d'atenció telefònica als clients) ha d'accedir a cadascuna de les aplicacions per fer la consulta corresponent. Si les aplicacions estiguessin integrades, n'hi hauria prou amb una sola consulta per obtenir les dades personals i comercials del client.

El magatzem de dades operacional ofereix una visió integrada de les dades operacionals de l'organització; a diferència del magatzem de dades corporatiu, no emmagatzema la història de les dades, és volàtil i està actualitzat. La seva missió principal consisteix a oferir suport operacional a l'organització. L'utilitzen principalment usuaris operacionals (oficinistes), tot i que també ho fan els analistes.

Exemple d'aplicació dels magatzems de dades operacional i corporatiu

Des d'un punt de vista de les vendes d'una empresa, si un analista realitza una anàlisi històrica de les vendes per producte, accedirà al magatzem de dades corporatiu o departamental, mentre que un usuari més vinculat a l'operativa diària que necessita saber l'estoc de determinat producte en el moment actual obtindrà aquesta dada en el magatzem de dades operacional.

En aquest apartat, estudiarem diferents tipus de magatzems de dades operacionals, com contribueix aquest concepte a l'organització i també alternatives que té l'organització per disposar-ne.

4.3.1. Paquets d'aplicacions i el magatzem de dades operacional

Al mercat, trobem paquets d'aplicacions que ofereixen una visió integrada de l'entorn operacional. Atès que la majoria de les organitzacions en un sector tenen necessitats semblants, la idea general d'aquests productes és oferir un conjunt d'aplicacions integrades que cobreixin les necessitats generals d'una

organització prototip dins d'un determinat sector. En molts casos, les aplicacions es poden adaptar a les característiques particulars del client que les adquireix, generalment a un cost elevat. Amb freqüència, són les organitzacions les que s'han d'adaptar a la manera de treballar imposada per les aplicacions.

Algunes organitzacions intenten «alliberar-se» de les seves antigues aplicacions mitjançant l'adquisició de paquets d'aplicacions integrades. Tot i així, és freqüent que organitzacions que adquireixen un d'aquests paquets hagin de conservar part de les seves aplicacions tradicionals perquè les noves no cobreixen totalment les seves necessitats, o bé s'ha decidit adquirir només una part del paquet d'aplicacions. En aquests casos, han de conviure els dos entorns i generalment les aplicacions del paquet adquirit han d'obtenir dades de les aplicacions que hi havia prèviament.

Quina relació hi ha entre el magatzem de dades operacional i els paquets d'aplicacions integrades?

Tots dos ofereixen una visió integrada de les dades de l'entorn operacional. Les característiques d'aquests també són idèntiques per als dos i en ambdós casos l'objectiu principal és oferir suport operacional a l'organització.

Els paquets d'aplicacions integrades l'àmbit de les quals és tota l'organització són implementacions comercials del magatzem de dades operacional.

Alguns dels paquets comercials que hi ha al mercat estan dissenyats tenint en compte l'arquitectura de la FIC i ofereixen mitjans per integrar-s'hi, especialment per obtenir les dades que construiran al magatzem de dades corporatiu. En altres casos, van més lluny i ofereixen com a ampliació del paquet bàsic una versió genèrica de la FIC per a les organitzacions del sector. Tot i així, pel que fa a la presa de decisions dels analistes, oferir un suport estàndard que sigui adequat per a tots els analistes és molt més complex que el que es refereix al suport al treball més operacional de l'organització, ja que les necessitats són més diverses.

Si construïm el magatzem de dades operacional, una vegada dissenyada la base de dades que el suporta i els mètodes d'obtenció i refrescament de les seves dades a partir de les aplicacions operacionals, construirem aplicacions directament sobre aquest. En alguns casos, les noves aplicacions cobriran necessitats noves i, en d'altres, substituiran aplicacions antigues com passa amb els paquets d'aplicacions integrades.

El magatzem de dades operacional pot ser adquirit en forma de paquet d'aplicacions integrades o construït en l'organització.

4.3.2. Velocitat de refrescament de les dades

En el cas que convisqui el magatzem de dades operacional amb les aplicacions operacionals, cas freqüent a les organitzacions, aquest ha d'obtenir part de les seves dades a partir de les de les aplicacions. Inmon distingeix quatre classes de magatzems de dades operacionals segons la velocitat de refrescament de les dades:

- Classe I: s'actualitza uns segons després que es produeixin les modificacions en les aplicacions operacionals.

- Classe II: s'actualitza unes hores després que es produeixin les modificacions.

- Classe III: el període d'actualització és superior a les vint-i-quatre hores.

- Classe IV: s'actualitza de manera no planificada.

En els magatzems de dades operacionals de classe I no es poden acumular els moviments obtinguts a les fonts de dades operacionals per aplicar-los de manera massiva. S'ha de disposar d'un mitjà per traslladar les modificacions directament (mitjançant disparadors, RPC o altres mitjans). Generalment, no s'apliquen massa transformacions a les dades en els magatzems d'aquesta classe.

Per actualitzar els magatzems de dades operacionals de classe II, III i IV podem aplicar els mètodes d'obtenció d'actualitzacions a partir de les fonts de dades operacionals estudiades en aquest mòdul. Per als de classe II, en la majoria dels casos necessitarem un mètode de captura immediata. Per als de classe III i IV, podrem utilitzar qualsevol dels mètodes estudiats; a més, com que no guardem la història de les dades, no hi ha cap problema per utilitzar un mètode de captura retardada encara que perdem detall dels moviments.

Com més restrictives siguin les condicions d'actualització, més costarà implementar-les. Per aquest motiu, un magatzem de dades de classe I serà molt més car que un de classe IV.

4.3.3. Planificació d'incorporació del magatzem de dades operacional

Plantejar l'adquisició o la construcció del magatzem de dades operacional com un sol projecte presenta el greu inconvenient de la dificultat de justificar el seu cost davant de l'organització.

En situacions normals, és més adequat fer un desenvolupament iteratiu similar al plantejat per desenvolupar la resta de la FIC. Sovint, es pot plantejar el magatzem de dades operacional com una estructura de suport per actualitzar les dades del magatzem de dades corporatiu. En aquests casos es construirà, com a part dels projectes atòmics en els quals s'ha dividit la construcció de la FIC, la part corresponent a l'emmagatzematge de les dades, de manera conjunta amb el magatzem de dades corporatiu. Així mateix, dins de la construcció de la FIC, podem plantejar projectes de desenvolupament o d'ampliació del magatzem de dades operacional, amb les mateixes premisses amb les quals definim la resta dels projectes.

Si se'n planifica la construcció abans de disposar del magatzem de dades corporatiu, pot servir per construir-lo. Tot i així, es corre el risc de pretendre incorporar al magatzem de dades operacional funcionalitats pròpies del magatzem de dades corporatiu.

Com que el cost resulta difícilment justificable, és recomanable adquirir o desenvolupar el magatzem de dades operacional de manera iterativa, mitjançant projectes autònoms.

Resum

Conèixer l'arquitectura de la FIC representa un avenç per a les organitzacions que volen oferir eines de suport a la presa de decisions per als analistes. Tot i que es poden considerar arquitectures alternatives, presenten greus deficiències que les fan inviables en la majoria de les situacions.

L'arquitectura de la FIC s'ha de traçar des d'un nivell alt, tenint en compte la perspectiva de tota l'empresa. Tot i així, és un error plantejar el desenvolupament de la FIC com un sol projecte: s'ha de fer de manera iterativa, per mitjà de projectes independents amb objectius i beneficis clars. És a dir, els successius projectes de desenvolupament de la FIC fan que la seva funcionalitat augmenti amb el temps i, amb aquesta, el benefici que aporten a l'organització.

En aquest mòdul, a més d'estudiar diferents alternatives en l'arquitectura de la FIC, així com diferents maneres de planificar la seva construcció, hem estudiat aspectes concrets de la construcció dels seus components. Especialment, hem prestat especial atenció al component d'integració i transformació per la seva dificultat; així mateix, hem estudiat detalladament les peculiaritats de la construcció del magatzem de dades corporatiu i del magatzem de dades de dades operacional.

Exercicis d'autoavaluació

1. Podem adquirir la FIC?

2. En què consisteix l'entorn operacional de teranyina?

3. Podem construir el magatzem corporatiu després d'haver construït diferents magatzems de dades departamentals?

4. És adequat combinar el magatzem de dades operacional i el corporatiu en una sola estructura?

5. La FIC ha d'incloure tots els components estudiats en el mòdul «La factoria d'informació corporativa»?

6. Podem construir la FIC amb un sol projecte?

7. Quines característiques han de tenir els projectes de construcció de la FIC?

8. Quina estructura han de tenir els projectes de construcció de la FIC?

9. Segons quina metodologia desenvoluparem els projectes de construcció de la FIC?

10. Què passa amb l'entorn operacional quan desenvolupem la FIC?

11. Qui és el patrocinador de la FIC?

12. Quin és el principal inconvenient dels mètodes de captura retardada per obtenir les modificacions produïdes a les dades de les fonts de dades operacionals?

13. Quines característiques té el model de dades del magatzem de dades corporatiu?

14. Quina relació hi ha entre el magatzem de dades operacional i els paquets d'aplicacions integrades?

15. Quin model de dades té el magatzem de dades operacional?

16. Des del punt de vista de les metadades del component de transformació i integració, quina opció és més òptima: desenvolupar el component amb codi manual o utilitzar una eina de suport?

Solucionari

Exercicis d'autoavaluació

1. Generalment, no. La FIC s'ha de construir a cada organització. Alguns paquets d'aplicacions tenen com a àmbit tota l'organització, inclouen funcionalitats de la FIC i estan pensats per cobrir les necessitats d'una organització estàndard. Si el sistema d'informació de l'organització està implementat mitjançant un paquet estàndard que inclou la FIC, en aquest cas la FIC hauria estat adquirida. Aquesta és una situació límit; generalment, la FIC s'ha de construir a mida per a les diferents organitzacions.

2. És el resultat de l'evolució incontrolada de l'entorn operacional. Mitjançant la construcció de programes d'extracció d'informació i bases de dades temporals per elaborar informes puntuals, arribem a un entorn d'alta complexitat amb una estructura de relacions entre els diferents components, que s'assembla a una teranyina.

3. La construcció de magatzems de dades departamentals ha d'estar basada en el magatzem de dades corporatiu. Si intentem construir el magatzem de dades corporatiu a partir dels diferents magatzems de dades departamentals que hi havia prèviament, la integració de les dades d'aquests serà molt complexa i segurament també requerirà la modificació posterior dels magatzems de dades departamentals.

4. Tot i que tenen algunes característiques semblants, els seus objectius són diferents i el tipus d'operacions a les quals ofereixen suport són incompatibles. El magatzem de dades operacional ha d'estar configurat per fer operacions de modificació. Tot i així, el magatzem de dades corporatiu necessita estar optimitzat per fer de manera exclusiva operacions de consulta.

5. No necessàriament. El magatzem de dades operacional és una estructura opcional, tot i que convé incloure-la per la funcionalitat que aporta. La resta dels components sí que són necessaris.

6. Si plantegem la construcció de la FIC mitjançant un sol projecte, serà massa complex i tindrem dificultats per justificar-ne el cost. Per tant, aquesta estratègia de construcció no serà adequada.

7. El primer projecte ha de tenir com a objectiu planificar la construcció de la FIC. Posteriorment, tindrem projectes de desenvolupament d'infraestructura. Dividirem la construcció de la FIC en projectes autònoms que aportin un valor clar a l'organització, que tinguin un responsable a dins i que es desenvolupin en un termini raonable.

8. Els projectes han de ser complets, és a dir, han de preveure l'obtenció i l'emmagatzematge de les dades i el seu accés.

9. La metodologia de desenvolupament més adequada és una metodologia iterativa (CLDS), ja que generalment no es coneixen detalladament els requeriments dels analistes d'informació i aquesta metodologia ens permet descobrir-los mitjançant refinaments successius.

10. Quan planifiquem la construcció de la FIC, també hem de planificar el desmantellament de l'entorn operacional en teranyina, de manera que a mesura que construïm els projectes que implementa la FIC també eliminem els programes d'extracció de dades i les bases de dades temporals que es deixen d'utilitzar.

11. És una figura política. Sol ser un directiu d'alt nivell de l'organització que està convençut dels beneficis que pot aportar la FIC i la seva missió és obtenir els recursos per als projectes el benefici dels quals no és directament justificable i intentar solucionar l'oposició que pugui sorgir en l'organització a la construcció de la FIC.

12. El problema principal que presenten aquests mètodes és que poden perdre el detall dels moviments produïts. Mitjançant aquests mètodes, obtenim un resum de totes les operacions fetes en una única operació de modificació.

13. Es tracta d'un model de dades orientat a emmagatzemar-les. Els esquemes des del punt de vista lògic estan dissenyats de manera que l'emmagatzematge i les consultes es facin de manera òptima.

14. Els paquets d'aplicacions integrades l'àmbit de les quals és tota l'organització són implementacions comercials del magatzem de dades operacional.

15. El magatzem de dades operacional està construït utilitzant un model de dades com el de les aplicacions operacionals. Els esquemes des del punt de vista lògic estan dissenyats de manera que les operacions de modificació es facin de manera òptima.

16. És més òptim utilitzar una eina de suport que ens generarà la metadada de manera automàtica.

Glossari

ASCII Sigles en anglès d'*American Standard Code for Information Interchange* (codi estàndard estatunidenc per a l'intercanvi d'informació), és un codi de caràcters basat en l'alfabet llatí.

CASE Sigles en anglès de *Computer aided system engineering*.

CLDS Metodologia iterativa de desenvolupament de projectes. Correspon a les sigles de SDLC al revés, com a contraposició als plantejaments d'aquesta metodologia.

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

EBCDIC Sigles en anglès d'*Extended Binary Coded Decimal Interchange Code*, és un codi estàndard de 8 bits usat per computadores *mainframe* IBM.

Enterprise datawarehouse bus matrix Arquitectura creada per Kimball que proposa una construcció basada en magatzems departamentals interconnectats.

Extract Transform Load Terme en anglès per denominar els processos d'extracció, transformació i càrrega que alimenten els magatzems de dades.

ETL Vegeu *Extract Transform Load*.

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

índex creatiu *m* Conjunt de dades interessants per als analistes, calculades en el moment de passar les dades de les fonts de dades operacionals fins al magatzem de dades corporatiu.

main frame Ordinador central o corporatiu.

massively parallel processors Arquitectura de computació en la qual diferents plataformes compostes per processador i memòria s'interconnecten mitjançant una línia d'alta velocitat.

MPP Vegeu *massively parallel processors*.

RPC *Remote Procedure Call*.

SDLC Metodologia de desenvolupament de projectes segons un cicle de vida en cascada en *systems development life cycle*.

SGML Sigles d'*Standard Generalized Markup Language* (llenguatge de marcatge generalitzat estàndard). Sistema per a l'organització i etiquetatge de documents.

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

sistema decisional *m* Aquell que dona suport als processos de presa de decisions per part dels analistes en l'organització.

sistema informacional *m* Vegeu sistema decisional.

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

SMP Vegeu *symmetric multi processing*.

symmetric multi processing Arquitectura de computació en la qual un conjunt de processadors comparteix memòria comuna com a mitjà de comunicació entre aquests.

XML Sigles en anglès d'*eXtensible Markup Language* (llenguatge de marques extensible), és un llenguatge de marques utilitzat per emmagatzemar dades de forma interpretable.

Bibliografia

Devlin, B. (1997). *Data Warehouse from Architecture to Implementation*. Reading, Mass.: Addison Wesley Longman, Inc.

Inmon, W. H. (1996). *Building the Data Warehouse* (2a. ed.). Nova York: John Wiley & Sons, Inc.

Inmon, W. H. (1999). *Building the Operational Data Store*. Nova York: John Wiley & Sons, Inc.

Inmon, W. H. (2005). *Building the Data Warehouse* (4a. ed.). Nova York: John Wiley & Sons, Inc.

Inmon, W. H.; Imhoff, C.; Sousa, R. (1998). *Corporate Information Factory*. Nova York: John Wiley & Sons, Inc.

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).

Inmon, W. H.; Welch, J. D.; Glassey, K. L. (1997). *Managing the Data Warehouse*. Nova York: John Wiley & Sons, Inc.

Jarque, M.; Lenzerini, M.; Vassiliou, Y.; Vassiliadis, P. (2000). *Fundamentals of Data Warehouses*. Berlín: Springer Verlag.

Kelly, S. (1997). *Data Warehousing in Action*. Nova York: John Wiley & Sons, Inc.

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

Kimball, R. (2009). *Data Warehouse Toolkit Classics: The Data Warehouse Toolkit* (2a. ed.); *The Data Warehouse Lifecycle Toolkit* (2a. ed.); *The Data Warehouse ETL Toolkit*. Hoboken: John Wiley & Sons.

Mattison, R. (1996). *Data Warehousing: Strategies, Technologies and Techniques*. Computing McGraw-Hill.

Silverston, L.; Inmon, W. H.; Graziano, K. (1997). *The Data Model Resource Book*. Nova York: John Wiley & Sons, Inc.