Administració de les dades

PID_00275591

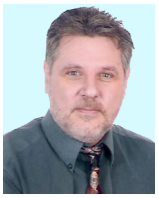
Manel Mendoza Flores
Miquel Colobran Huguet
Javier Panadero Martínez

Temps mínim de dedicació recomanat: 3 hores

Universitat Oberta de Catalunya

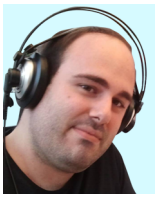
Manel Mendoza Flores

Enginyer de telecomunicacions, especialista en seguretat informàtica amb experiència en l'àmbit de l'administració pública i del sector privat. Diplomat en Ciències Empresarials per la Universitat Oberta de Catalunya (UOC) i en Gestió de Projectes (PMP). Durant la seva trajectòria formativa s'ha complementat amb diverses certificacions del món IT (Cisco, Microsoft, CISSP, etc.). Des de l'any 2011, col·labora amb la UOC en diversos àmbits de la docència, laboratoris en línia i direcció del TFG. Apassionat per les noves tecnologies, que combina amb les obligacions de la seva família i l'estima del Delta de l'Ebre, d'on és natural. Actualment, desenvolupa la seva ocupació com a expert de seguretat al sector privat en un entorn internacional.



Miquel Colobran Huguet

Doctor en Informàtica per la Universitat Autònoma de Barcelona (UAB). Consultor a la Universitat Oberta de Catalunya (UOC) d'assignatures sobre administració de sistemes i seguretat, i també d'informàtica i legislació en el grau i màster d'Informàtica i Multimèdia. Ha elaborat diversos materials i llibres sobre administració de sistemes, seguretat, informàtica forense i legislació aplicada a les tecnologies de la informació. La seva recerca s'emmarca dins de la seguretat, la influència de les TIC a la societat i l'enginyeria del coneixement.



Javier Panadero Martínez

Enginyer informàtic i doctor en Computació d'Altes Prestacions per la Universitat Autònoma de Barcelona (UAB). Des de 2019, és professor dels Estudis d'Informàtica, Multimèdia i Telecomunicació de la Universitat Oberta de Catalunya (UOC). Director del màster universitari en Enginyeria Computacional i Matemàtica. Ha elaborat diversos materials sobre administració de sistemes i programació. Els seus interessos de recerca inclouen la computació paral·lela i distribuïda, l'optimització i simulació de sistemes complexos i els algorismes intel·ligents.

Primera edició: setembre 2020

Av. Tibidabo, 39-43, 08035 Barcelona
Autoria: Manel Mendoza Flores, Miquel Colobran Huguet, Javier Panadero Martínez
Producció: FUOC

Índex

Introducció

Actualment **un dels grans valors de totes les organitzacions és la informació** (també en podríem dir dades, tot i que ja veurem quina diferència hi ha). Aquesta és la matèria primera del sistema informàtic. Per tant, el que s'ha de conèixer més bé és on està, perquè com que tots els ordinadors estan connectats per xarxes informàtiques, correm el perill que les dades estiguin disperses per totes les estacions de treball de l'organització. Si això passa, no sabrem ni quines dades tenim ni on les hem d'anar a buscar.

Necessitarem saber on està la informació per fer-ne còpies de seguretat, ja que, en cas que hi hagi un desastre, el programari es pot reinstal·lar, però les dades les ha creat l'organització, no es poden «comprar» enlloc. Com que actualment la informació és un dels actius més importants, cal protegir-la.

També és important saber on estan aquestes dades perquè les puguem combinar per obtenir dades noves sobre la nostra organització.

Per a la direcció (i, en general, per a tota l'organització), disposar de la informació adient a temps permet prendre decisions correctes en cada situació i en el moment oportú.

Objectius

Els materials didàctics d'aquest mòdul contenen les eines necessàries perquè l'estudiant assoleixi els objectius següents:

1. Entendre que la informació de l'organització és molt valuosa.

2. Comprendre que cal saber on està tota la informació de l'organització.

3. Conèixer diferents maneres de mantenir la integritat de la informació.

4. Aprendre per què s'han fet populars els servidors de bases de dades i altres arquitectures de dades.

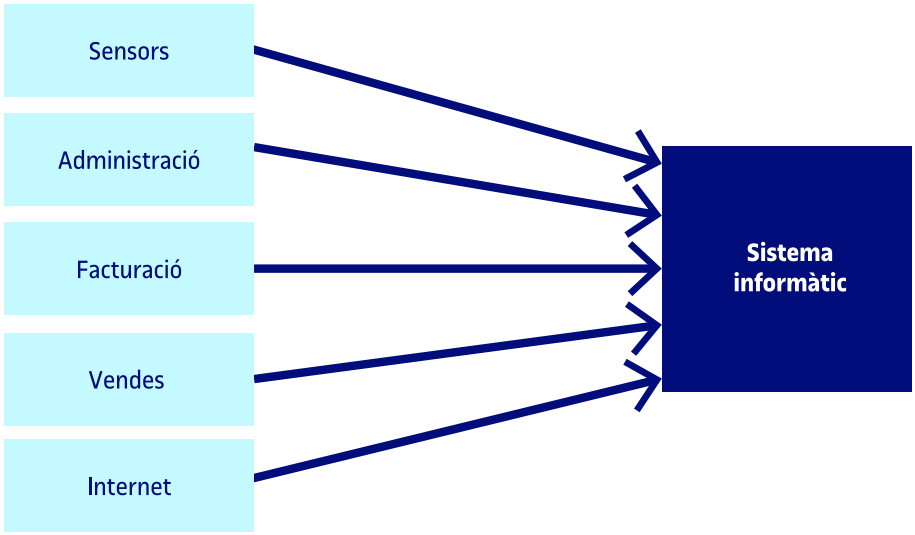
5. Saber a quins llocs hi pot haver les dades de l'organització i quins són els millors llocs on podrien estar.

6. Conèixer els fonaments de seguretat de les bases de dades.

1. Les dades i l'organització

L'organització crea dades contínuament, per la qual cosa una manera de considerar el sistema informàtic és com si només es tractés d'un magatzem de dades. Les fonts de dades són variades, com es pot observar a la figura 1.

Figura 1. Un sistema informàtic pot recopilar dades de diverses fonts



Aquest «gran dipòsit» de dades en brut no és útil d'aquesta manera, perquè si les dades no tenen una organització i una coherència, amb la gran quantitat que n'hi ha, no tindria cap mena de sentit.

Abans de continuar endavant, hem de dir la diferència que hi ha entre informació i dades. La definició formal de cadascun d'aquests conceptes és la següent:

Les **dades** són els registres dels successos.

La **informació** és el processament de les dades perquè tinguin sentit.

Per exemple, 70293 pot voler dir 7 de febrer de 1993 o embalatge 7029 per al camió número 3. Un altre exemple, 1813 pot ser un número de matrícula de cotxe o pot ser una hora, les 18 hores i 13 minuts.

Dades i informació

Malgrat aquesta diferència formal, sovint els informàtics utilitzem aquestes dues paraules com a sinònimes. Com que el processament per donar sentit a les dades es pot fer a molts llocs, nosaltres també farem servir indiferentment els termes *dades* i *informació*.

Generalment es guarden dades i es presenta a l'usuari informació, és a dir, que el processament d'aquestes dades perquè tingui sentit es fa en el programari en el moment de presentar-la en el dispositiu de sortida. En principi això és el que dona més flexibilitat al sistema, ja que permet convertir les dades a qualsevol format de sortida. Tanmateix, sovint això queda lluny de la realitat,

perquè diferents programaris guarden les dades amb diferents formats. A més a més, la majoria de programaris són incompatibles entre ells, per la qual cosa la recuperació creuada d'informació no és gens senzilla.

Així doncs, les organitzacions creen constantment grans quantitats de dades i el que fan els sistemes informàtics és processar-les i distribuir-les entre tots els elements de l'organització per augmentar l'eficàcia del conjunt.

El **sistema informàtic** pretén guardar la informació de l'organització perquè sigui fàcil recuperar-la posteriorment i de la mateixa manera que s'havia guardat.

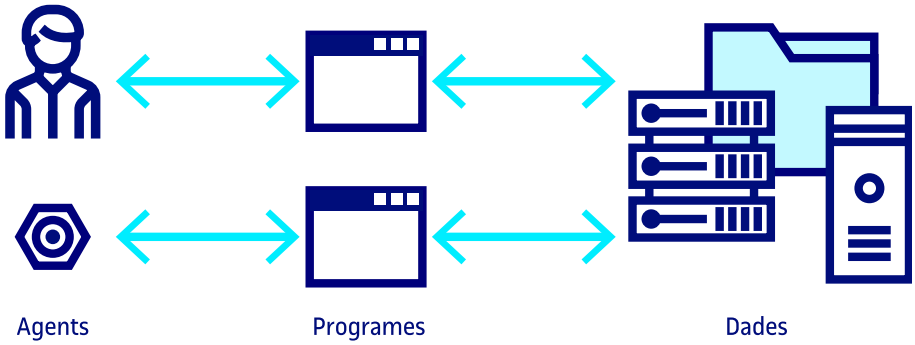
El primer objectiu del sistema informàtic és guardar la informació però fent-ne possible la recuperació igual que s'havia guardat. Ara bé, amb això desaprofitem el gran potencial dels ordinadors: la capacitat de manipular la informació.

En aquest context, «manipular informació» té un sentit força ampli. Aquests són alguns dels significats possibles:

- Poder afegir informació nova a la que ja teníem.
- Poder esborrar informació obsoleta de la que ja teníem.
- Poder modificar informació incorrecta de la que ja teníem.
- Poder relacionar diferents informacions per construir informació nova, contrastar la coherència o la veracitat de la informació que es vol entrar, reduir la redundància de la informació o buscar-hi noves relacions.

La interacció de les dades amb l'organització la podem veure d'aquesta manera:

Figura 2. Cadena d'interacció entre els usuaris, els programes i les dades



Tal com mostra la figura 2, la cadena d'interacció de les dades està composta per tres components principals: agents, programes i dades. A continuació, es defineix cadascun d'aquests.

- **Agents**. Poden ser, per exemple, persones o sensors que, a partir de programes, accedeixen a les dades de l'organització. També podríem considerar aquests agents com a entitats que estan fora de l'organització, a internet, però que interaccionen amb el sistema informàtic i modifiquen d'alguna manera les dades del sistema.

- **Programes**. Són el que s'utilitza per accedir a les dades de l'organització d'una manera ordenada i correcta. Sembla lògic que l'accés a les dades, atès que són un bé valuós de l'organització, s'ha de protegir. Els programes s'encarreguen (juntament amb el sistema operatiu) de fer tots els controls necessaris per evitar els accessos no desitjats a les dades de l'organització.

Amb aquesta visió de les dades de l'organització, sembla senzill, viable i potser fins i tot fàcil resoldre el problema d'un usuari que digui que necessita extreure una determinada informació del sistema informàtic. Pel que hem vist aquí, fins ara, aquest problema es reduiria a crear un nou programa que accedís a les dades. De moment, per exemple, no hem tingut en compte com es manegen els permisos i els grups que incorporen els diferents fitxers que integren la informació de l'organització per protegir el contingut, ja que la informació s'ha de protegir d'accessos indiscriminats.

El sistema informàtic recull informació, la guarda, la manipula i la presenta a l'usuari.

2. On està la informació

Desgraciadament, la realitat no és tan simple. A l'apartat anterior tot era ideal: guardàvem les dades i, quan les necessitàvem, les agafàvem i les presentàvem (imprimíem). Si era el que ens calia, ja havíem acabat. La realitat involucra molts més processos que fan que el sistema sigui més complex.

Aquests són alguns dels problemes:

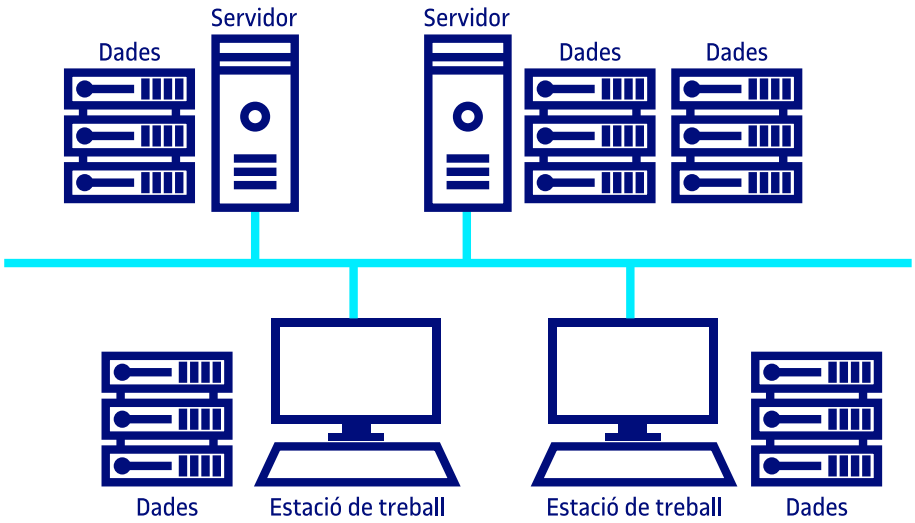
- Les dades estan guardades en fitxers.
- Les dades estan guardades en un format determinat.
- Les dades no sempre estan guardades on cal.
- Les fonts de dades, és a dir, els sensors, les persones, els formularis web, etc., que generen dades, les generen en un format determinat.
- Cada font de dades pot fer servir un format diferent i es pot guardar en fitxers potencialment diferents i en ubicacions diferents.

  Formularis web diferents poden generar fitxers de text diferents, i un sol usuari, depenent de l'aplicació que utilitzi, crearà informació diferent en formats diferents que anirà a parar a fitxers diferents. Si, per exemple, fa un document amb processadors de textos del balanç anual comptable per al banc i ho fa en diferents fitxers, segons els comptes, crearà informació relacionada (del mateix tema) que és molt difícil de poder relacionar mai l'una amb l'altra.

Per tant, el que ens passa a la pràctica és que les dades de l'organització estaran disperses des del punt de vista del programari. No tots els programes podran accedir a totes les dades, independentment dels permisos de l'usuari.

El problema és encara més greu si ens fixem en l'estructura de la xarxa. Vegem els llocs on hi pot haver potencialment informació. Les dades poden estar en aquests llocs perquè s'ha decidit així o perquè els usuaris, per desconeixement, generen informació i la guarden en aquests llocs, sense consultar-ho.

Figura 3. Distribució de les dades a diversos llocs dins l'organització



Tal com mostra la figura 3, podem tenir informació en qualsevol directori de qualsevol partició de qualsevol disc dels servidors o de les estacions de treball, fins i tot, en ordinadors portàtils.

És a dir, anant a l'extrem, podem tenir informació duplicada en màquines diferents, amb sistemes operatius diferents, que tinguin sistemes de fitxers diferents.

El pitjor és que una part d'aquesta informació pot ser crítica i molta pot ser desconeguda pel departament d'informàtica, o potser sap que existeix i no la té controlada i, per tant, no disposa de cap mecanisme de recuperació davant d'un desastre.

2.1. Possibles solucions

Com acabem de veure, tenim un problema que pot arribar a ser molt greu i que s'ha repetit històricament a totes les organitzacions. Vegem-ne possibles solucions.

1) A partir de la taula d'aplicacions i del disseny dels usuaris en general, sabem quin programari utilitzen els usuaris. També sabem, a partir de la taula d'aplicacions, on està la informació. El manteniment de la que estigui als servidors és responsabilitat de l'administrador d'usuaris o de l'administrador de servidors. En qualsevol cas, sabem on està i disposem dels mecanismes de seguretat, protecció i recuperació adients en cas d'emergència.

Vegeu també

Vegeu com es fa en el mòdul «Administració d'usuaris».

A partir de la taula d'aplicacions i del disseny dels usuaris en general, sabem quin és el programari que genera informació en local. De tot aquest programari, s'ha d'implementar el mecanisme de còpies de seguretat que faci falta, i

habilitar els permisos i la seguretat que calguin per garantir-ne la confidencialitat i poder-les incorporar al mecanisme general de còpia per estar previnguts en cas de fallada.

2) Informació «no controlable». Dins d'aquesta categoria hi acostuma a haver el programari d'ofimàtica (fulls de càlcul, processadors de textos, petites bases de dades, agendes, etc.). S'ha de posar especial atenció en el fet que les dades estiguin en la unitat de xarxa privada de l'usuari (de la qual es fa una còpia de seguretat cada dia).

3) Igualment, s'ha de formar els usuaris en l'ús de les eines informàtiques amb què treballen perquè aprenguin a posar la informació en la seva unitat personal de xarxa en lloc de fer-ho en la unitat local (o gravar-ho on proposa l'aplicació per defecte, ja que aleshores moltes vegades ni el mateix usuari no sap on s'han guardat realment les dades). Aquesta «cultura informàtica» pot evitar la dispersió de la informació per tots els discos de l'organització, que és un dels grans problemes d'administració.

Reflexió

Sabeu on està la informació a la vostra organització? Ho podeu comentar en el fòrum de l'assignatura.

Informàtica portàtil amb mobilitat

Actualment, amb el creixement a les organitzacions de la informàtica amb mobilitat, hi ha un gran problema afegit al de la dispersió, ja que aquests equips necessiten autonomia i disponibilitat en tot moment. La primera qüestió que se'ns planteja és: què passa amb les dades necessàries per treballar? Poden ser sensibles si surten fora de l'organització sense cap tipus de protecció. La segona qüestió és que aquests equips modifiquen o afegeixen dades al sistema informàtic (estan desconnectats de la xarxa). Com sincronitzen aquesta informació amb el sistema de l'organització? Una darrera qüestió seria que de vegades necessiten funcionar connectats a la xarxa de l'organització com si fossin una estació de treball més i, de vegades, els cal connectar-se a la xarxa de l'organització des de fora. Com es pot fer això?

S'han de buscar maneres de saber on està tota la informació de l'organització. Ens servirà, per exemple, per garantir-ne la seguretat.

3. La consulta de la informació

De la mateixa manera que l'organització genera una gran quantitat de dades constantment, els usuaris també necessiten nova informació de l'organització molt sovint. Això pot voler dir noves consultes a les dades o generar noves fonts de dades en el sistema.

Per tant, l'administrador ha de conèixer totes les bases de dades de l'organització.

Bases de dades

En aquest cas fem servir el terme *base de dades* en el seu sentit més ampli, ja que engloba des d'un full de càlcul fins a una base de dades completa, incloent, fins i tot, un document d'un processador de textos. Tot és informació de l'organització.

Ja que la seguretat de **tota** la informació de l'organització és responsabilitat de l'administrador, cal:

- Conèixer-ne la localització.
- Disposar d'un mètode de restauració en cas de problemes. Això implica còpies de seguretat, etc.
- Tenir una idea general del contingut. Ha de tenir informació de la informació (en podríem dir metainformació).

La localització és important perquè té una gran rellevància en les polítiques de les còpies de seguretat i en els temps d'accés per als usuaris, com també la facilitat d'accés, d'ampliacions futures, d'ampliació del grup d'usuaris que accedeixen a les dades, etc.

La idea general del contingut té utilitat per a l'optimització del sistema (no s'han de repetir bases de dades ni dades que ja en formen part) i per a les peticions de la direcció, que generalment van encaminades a extreure informació del sistema.

Mineria de dades

Els mètodes de mineria de dades pretenen extreure informació coherent a partir de quantitats d'informació no coherent i probablement dispersa. Hi ha una assignatura que tracta d'aquest «problema».

Quan aconseguim tenir totes les dades als servidors (algunes particions, d'alguns discos, d'alguns servidors) el problema no està resolt, però almenys hem avançat molt. Ja podem fer algunes coses:

- Que diferents grups/usuaris accedeixin a les mateixes dades (les comparteixin). Això els permet afegir-hi elements i consultar, modificar i esborrar aquestes dades.
- Com que les dades estan als servidors, la seguretat és molt més alta.
  - Còpies de la informació.
  - Pèrdua per fallada de l'equip.

- Robatori.
- Permisos i, per tant, restriccions d'accés.

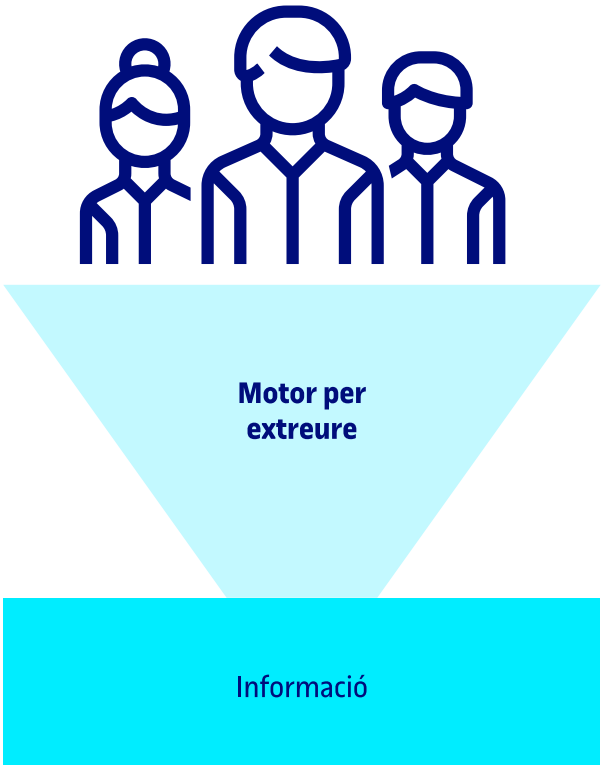
- Més facilitat d'accés i de difusió, perquè en aquestes condicions la informació és potencialment accessible per a tota l'organització (està controlat per la seguretat).

3.1. Les consultes de la direcció

Sembla que ens podem començar a plantejar un problema del qual fins ara no havíem parlat: el problema de la direcció.

La direcció veu el sistema informàtic com un magatzem d'informació i fa consultes a aquesta informació per tenir una ajuda en la presa de decisions.

Figura 4. Les organitzacions fan diverses cerques amb les dades



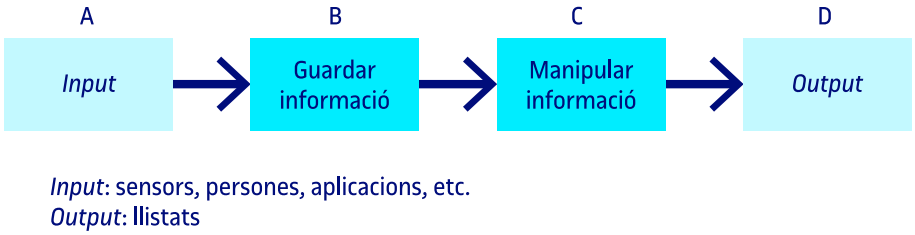
Per tant, com mostra la figura 4, la direcció veu el sistema informàtic com una caixa negra abstracta, a la qual, a l'hora de prendre decisions, li fa peticions del tipus «quin és el producte que ha donat menys hores de producció i més benefici de vendes?», o «quin producte ha costat menys de vendre als comercials?», o «quins productes tenen menys cost de transport?», o «de quins components hem tingut més pèrdues?», o, finalment, per exemple, «quins treballadors han rendit menys?».

Totes són preguntes sovint força difícils de contestar, malgrat que l'administrador pot arribar a saber que la informació està dins el sistema.

Abans, amb les dades distribuïdes per tota l'organització, era impossible contestar cap d'aquestes consultes. Ara, com que les dades estan als servidors, són totes accessibles. **Quin és el problema?**

Per contestar aquesta pregunta cal tornar una mica enrere i veure com s'arriba al punt actual sobre la informació de l'organització. La figura 5 representa el procés de guardar/recuperar informació.

Figura 5. Cicle de transformació de les dades



De moment només hem aconseguit solucionar el pas B de l'esquema de la figura (guardar informació), ja que tenim tota la informació accessible dins els servidors. Col·lateralment, hem millorat el problema de seguretat i de compartició de la informació, però el que no hem millorat gens és el punt A: la informació continua entrant amb formats completament heterogenis, en una gran quantitat de fitxers diferents i, per tant, els programes per manipular aquesta informació poden ser extremadament complexos de desenvolupar (per no dir impossible). Si manipular informació en formats heterogenis per obtenir informació coherent és complex, i la direcció fa consultes sovint, el problema esdevé intractable.

Si, finalment, obtenim algun resultat d'alguns d'aquests fitxers amb formats heterogenis, aquest resultat també estarà en un format heterogeni (la mateixa informació guardada de dues maneres). Què s'ha de fer per desenvolupar la llista final (pas D)?

No n'hi ha prou amb tenir tota la informació als servidors per poder dir que ja la podem manipular com vulguem.

3.2. Servidors de bases de dades

Amb l'objectiu de poder manipular tota la informació d'una corporació de forma eficient i que sigui útil a l'hora de prendre decisions, van sorgir els servidors de bases de dades.

Aquests servidors van aparèixer per resoldre la necessitat de les empreses de manejar grans volums heterogenis de dades, al mateix temps que requerien compartir la informació amb un conjunt de persones d'una manera ràpida, segura i senzilla.

Actualment, podem trobar dos tipus de sistemes de base de dades al mercat, amb diverses implementacions en cadascuna de les versions.

1) Bases de dades relacionals

Una base de dades relacional és un col·lecció d'objectes amb relacions definides entre si, organitzades amb taules d'informació on cada entrada representa una fila amb diferents camps o columnes. Les taules s'utilitzen per guardar informació sobre els objectes que representaran a la base de dades. Cada columna d'una taula guarda un determinat tipus de dades i cada camp emmagatzema el valor de l'atribut. D'igual forma, cada fila pot contenir la clau d'indexació que ha de servir per fer les cerques.

Cal destacar les característiques següents de les bases de dades relacionals:

- Integritat de les dades, segons les relaciones establertes entre els diferents camps.

- Conformitat ACID. Així, cada transacció ha de ser conforme als criteris de:
  - atomicitat,
  - coherència,
  - aïllament, i
  - durabilitat.

- Es pot modelar com un sistema transaccional, en què cada petició és independent de l'anterior i forma una única unitat lògica de treball.

- Es pot fer cerques amb el llenguatge SQL.

2) Bases de dades no relacionals o no SQL

Les bases de dades no SQL corresponen a una estructuració de dades que no suporta el llenguatge de consulta SQL i que, a la vegada, no compleixen les propietats ACID anteriorment mencionades. Van sorgir al mateix temps que el *big data*, davant la necessitat de guardar i manejar grans volum de dades de forma eficient. Típicament, aquestes bases de dades no SQL fan servir altres estructures de dades més fàcils d'escalar.

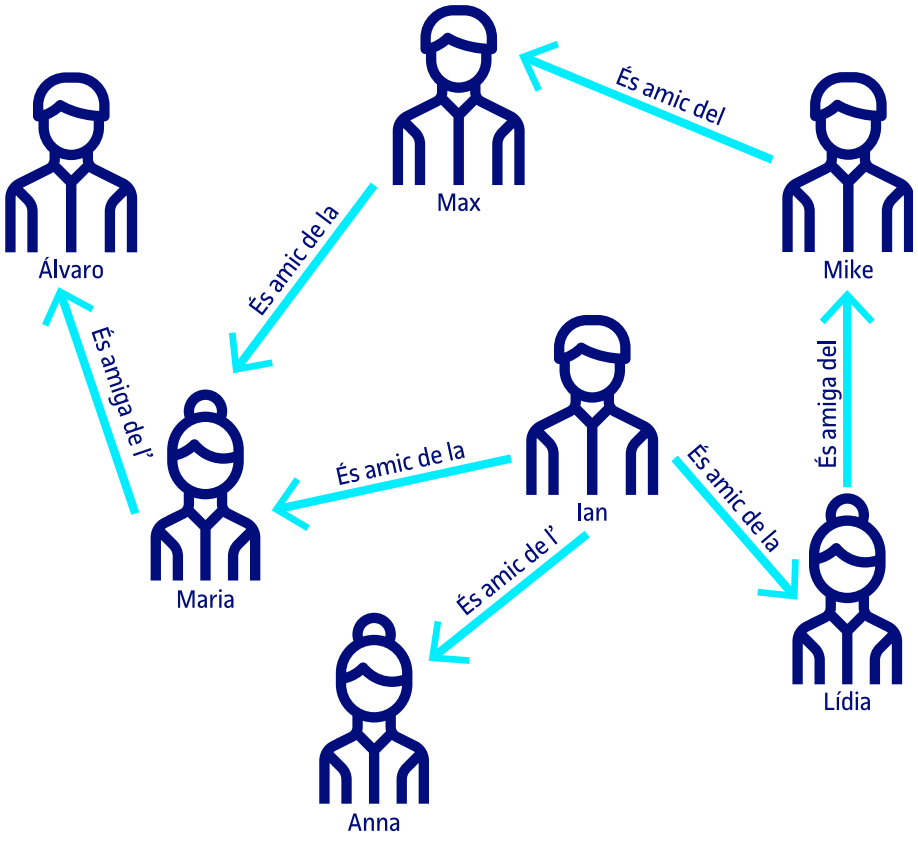
Podem trobar diversos tipus de bases de dades en funció del tipus de dada:

a) Clau-valor. Una base de dades de clau-valor és una base de dades simple que usa un arranjament associatiu com a model de dades, on cada clau està associada a un únic valor en una col·lecció. Aquesta relació es coneix com un parell clau-valor. Aquest és un mètode simple d'emmagatzemar dades, el qual és fàcil d'escalar. En aquesta base de dades, les dades són altament divisibles i es poden escalar de forma horitzontal fàcilment.

b) Documentals. Els documents són moltes vegades l'objecte més eficient per emmagatzemar i indexar la informació. Aquest tipus de bases de dades han estat dissenyades per emmagatzemar i consultar dades com ara documents de tipus JSON. La naturalesa flexible, semiestructurada i jeràrquica dels documents, i les de dades dels documents, permet que evolucionin segons les necessitats de l'organització. El model de documents funciona bé amb casos d'ús com ara catàlegs, perfils d'usuari i sistemes d'administració de contingut, en els quals cada document és únic i evoluciona amb el temps.

c) Gràfiques. El propòsit és crear bases de dades amb dades que estan altament connectades entre elles. Són molt útils per capturar relacions complexes en grans xarxes d'informació. Es representen mitjançant un graf on els nodes i les arestes contenen la informació de la consulta. L'objectiu és representar la informació d'una manera visual, perquè pugui ser llegida de manera més natural. La figura 6 mostra un exemple del resultat d'una consulta a una base de dades gràfica, on es representa les relacions entre un conjunt de persones.

Figura 6. Exemple del resultat d'una consulta a una base de dades gràfica



3.3. Processament de dades a gran escala

A continuació, descriurem les diferents tècniques que ens permeten treballar amb dades de gran volum.

3.3.1. *Big data*

Podem definir el *big data* com l'anàlisi massiva de dades de diversos orígens que, a causa del volum, les aplicacions de programari de processament de dades que tradicionalment s'usaven no són capaces de capturar, tractar i posar en valor en un temps raonable. Tot això ha comportat que apareguin noves tecnologies que fan possible l'emmagatzematge i processament, a més de l'ús que es fa de la informació obtinguda per mitjà d'aquestes tecnologies.

En general, podem tenir diverses fonts de dades que generen aquest volums:

- Persones que fan ús de recursos IT a gran escala com ara Facebook, Twitter, etc.
- Comunicacions entre sistemes, com poden ser els *logs*, dades M2M, dispositius IOT que envien telemetria, etc.
- Dades de màrqueting, on cada cop es capturen més dades de l'usuari.

Avui en dia podem afirmar que el volum de dades gestionades recentment és molt superior al de fa uns anys i que la tendència és que continuï creixent. Podem destacar les propietats següents de les dades:

1) Velocitat: avui en dia la informació es genera des de diferents fonts i requereix que tant el seu processament com l'emmagatzematge siguin immediats. El nostre concepte d'immediatesa ha canviat en els últims temps i es busca informació que arribi pràcticament a l'instant. Així, la velocitat d'anàlisi requerida per la societat actual és una de les característiques fonamentals que tenen les dades a gran escala, en què les dades en constant moviment i processades a temps real cobren protagonisme, executant algoritmes cada vegada més complexos en menys temps.

2) Varietat: depenent de la font de generació, el volum i la freqüència d'aquestes dades poden ser totalment oposades. Així, tenim dades com ara les piulades que tenen una grandària molt petita, però a l'altre extrem podem trobar els vídeos d'alta resolució que ocupem molt d'espai en comparació amb les piulades. Addicionalment, continua en augment la quantia de dades no estructurades a proporció de les tradicionals. Igual que passava amb el volum, aquesta entrada en escena amb força de les dades no estructurades requereix nous tractaments de la informació, requerint noves metodologies i tecnologies per poder ser analitzades.

3) Valor: aquest és un aspecte clau a l'hora de gestionar el cicle de vida de les dades. El valor és sens dubte una qualitat fonamental en l'anàlisi i la que pot influir de forma més directa en la forma de tractar aquetes dades.

4) Variabilitat: en un entorn tan canviant com el de les macrodades, la informació varia molt i això fa que els models o tractaments que s'apliquen al voltant d'aquesta requereixen un control periòdic.

5) Volum: com hem comentat, la quantitat de dades generades està augmentant. A mesura que creixen les bases de dades, també ho han de fer les aplicacions i l'arquitectura construïda per suportar la recollida i l'emmagatzematge de les dades cada vegada més variades. Tot i que el cost per TB d'emmagatzematge s'ha reduït considerablement, continua sent un dels aspectes clau per poder dimensionar correctament un sistema.

6) Veracitat: la informació recollida ha de ser veraç. Això és important per obtenir unes dades de qualitat i esdevé fonamental per poder fer una anàlisi acurada, fins i tot, depenent de les aplicacions que se li vulguin donar.

7) Visualització: la representació en un format que es pugui entendre fàcilment esdevé clau per poder posar en valor les dades amb vista al negoci.

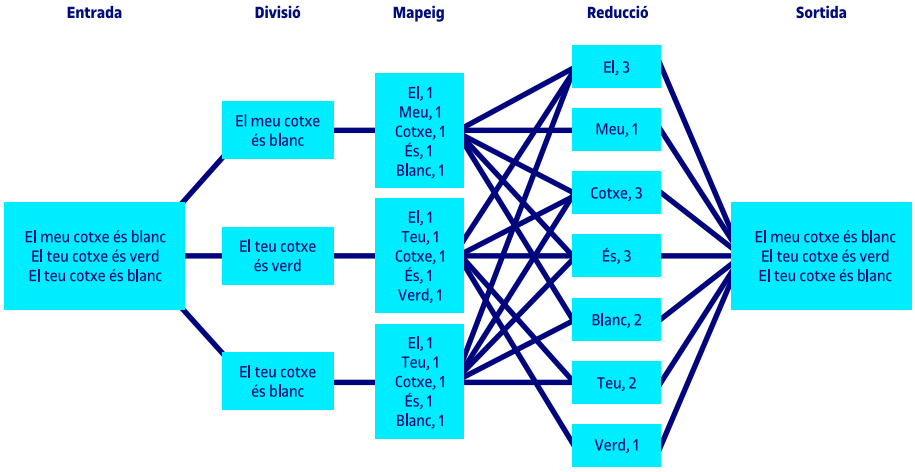
Per tal de treballar amb aquest gran volum de dades, es requereixen tècniques de processament i emmagatzematge eficients. A continuació, presentarem una tècnica àmpliament usada.

3.3.2. Model *map reduce*

Amb l'aparició del gran volum de dades, han aparegut tècniques per processar aquests volums tan elevats de dades. Una d'aquestes tècniques és la tècnica *map reduce*, que consisteix en un procés inicial de mapeig (*map*), filtratge i ordenació de les dades, típicament en servidors paral·lels. Seguidament, un cop tenim les dades formatades, es procedeix a operacions de reducció.

La figura 7 mostra un exemple de *map reduce*. Com podem observar, tenim un *dataset* d'entrada, on després de dividir-lo i mapejar-lo (clau, valor), podem ordenar els camps per la clau (a l'exemple següent, la clau correspon als noms, junt amb els valors de cada entrada). Seguidament, el pas de reducció agrupa les diferents claus sumant els valors, aconseguint una indexació eficient dels continguts.

Figura 7. Exemple de *map reduce*, en què es poden observar les diferents etapes

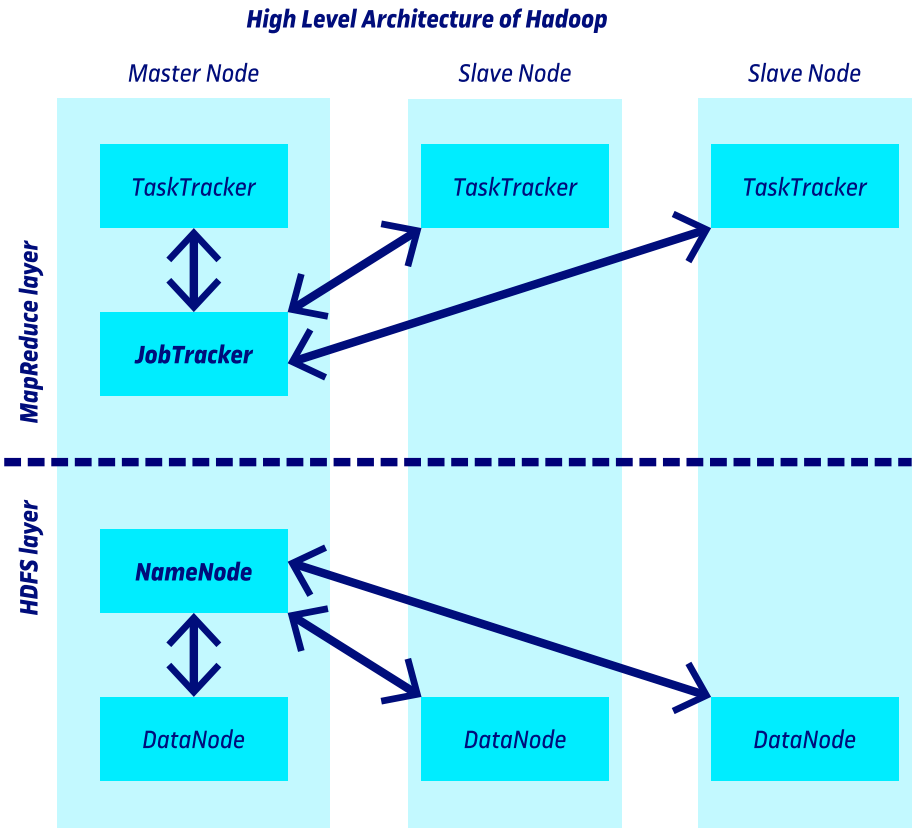


1) Hadoop

Hadoop és un *framework open source* per emmagatzemar, processar i analitzar grans volums de dades, el qual facilita als desenvolupadors les dificultats de programació en paral·lel.

Per a l'emmagatzematge de les dades, Hadoop ha creat l'HDFS, Hadoop Distributed File System, escrit amb Java, en què les dades estan repartides en diferents nodes amb diverses replicacions perquè tinguin una alta disponibilitat. Com podem observar a la figura 8, les dades estan distribuïdes i disponibles al llarg dels diferents nodes que componen el clúster HDFS.

Figura 8. Arquitectura de Hadoop amb HDFS



Font: adaptada d'https://upload.wikimedia.org/wikipedia/commons/8/85/Hadoop-HighLevel_hadoop_architecture-640x460.png.

Per al processament de les dades, Hadoop implementa el sistema *map reduce* que típicament s'implementa conjuntament amb el sistema de dades HDFS.

A partir d'una entrada de dades (*input*) el primer procés que té lloc és la separació de la informació per poder emmagatzemar-la en diferents membres dels clúster i poder aplicar d'igual forma algoritmes de còmput paral·lel.

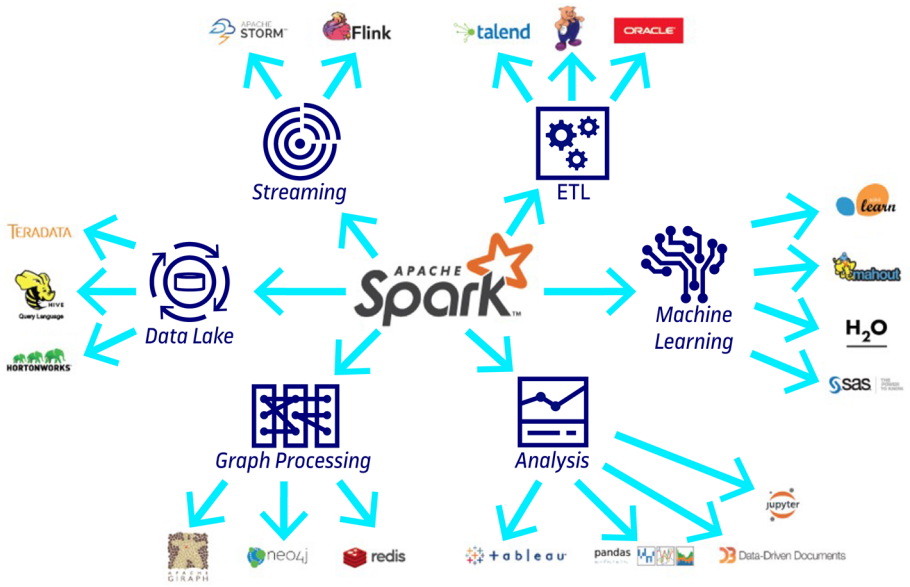
Un cop ja tenim la informació fraccionada, es procedeix al mapeig clau-valor. Posteriorment, ja es pot treballar amb aquestes dades, comminant-les u ordenant-les, per finalment, poder fer una reducció i presentar-les. En podeu veure un exemple a l'apartat anterior on s'explica el *map reduce*.

2) Spark

Spark és un alternativa més focalitzada al processament de dades *big data*. A diferència de Hadoop, fa còmputs de dades molt més ràpidament, principalment perquè treballa amb dades en memòria i, per tant, disminueix les operacions de lectura/escriptura al disc, que típicament són accions lentes.

A diferència de Hadoop, Spark pot llegir dades de diferents tipus de fitxers (HDFS, Amazon S3, Google Storage, etc.), posant molt de focus en la rapidesa de processament de les dades.

Figura 9. Spark permet treballar amb diferents tipus de dades i connectors



Com podem observar a la figura anterior (figura 9), Spark ens ofereix més flexibilitat a l'hora de tirar endavant altes càrregues de treball, tenint algoritmes iteratius, processament en *streaming*, API amb Python, etc. Per aquesta finalitat, Spark té molts connectors i suport per a diferents tipus de repositoris de dades i anàlisi.

3.4. ERP

Actualment encara s'ha dut molt més enllà la idea dels servidors de bases de dades. Les aplicacions, que sabem que utilitzen totes les organitzacions, s'instal·len integrades en un servidor de bases de dades. Després es comercialitza el producte.

El resultat s'anomena *ERP* o ***sistema de planificació de recursos de l'empresa***. És un conjunt de mòduls o paquets (aplicacions), com, per exemple, comptabilitat, facturació i nòmina tot perfectament integrat i funcionant en un servidor de bases de dades.

Aquests sistemes de gestió de la informació, com que integren i automatitzen molts aspectes operacionals de l'organització, homogeneïtzen tot el procés d'introducció d'informació i els informes. Atès que integra tota la informació en un mateix lloc, les consultes que es poden fer són moltes i molt potents.

3.5. *Cloud data management*

Cada cop es més habitual enviar les dades al núvol i, per això, cal considerar diversos aspectes per facilitar la gestió de la migració i, en especial, la seguretat de les dades migrades.

1) Seguretat de les dades. Podem trobar dues variants:

a) Seguretat de les dades en trànsit, durant l'enviament o consulta d'aquestes dades. Sempre és recomanable fer servir protocols segurs, com ara HTTPS.

b) Seguretat en l'emmagatzematge permanent, en què l'aspecte a considerar és la gestió de les claus de xifratge. Hi ha dues opcions:

- Clau de xifrat proporcionada pel mateix client.
- Clau de xifrat gestionada i proveïda pel proveïdor.

2) Govern de l'accés a les dades. Amb el núvol, la compartició dels fitxers és més senzilla, tot i que no s'ha de perdre mai de vista que cada fitxer té uns permisos d'accés que s'han de respectar tot i accedir des del núvol.

3) *Backups* i *disaster recovery*. En aquest aspecte, tot i que el núvol té més opcions de *backup*, cal remarcar que les còpies de seguretat i els mecanismes de continuïtat són un punt clau per a un bon govern de les dades.

4) Govern de les dades i qualitat. Tot i que la compartició de fitxers al núvol és més senzilla, cal tenir en compte que la informació ha de ser coherent i ha de tenir una única font d'origen.

També cal assenyalar que les organitzacions poden implantar sistemes de prevenció de fuga de dades o DLP, on cada fitxer es classifica i es protegeix de forma individual per evitar accessos indeguts. N'és un bon exemple l'Azure Information Protection, eina integrada amb la suite MS Office.

3.6. Repositoris remots

Avui en dia el núvol ofereix diverses formes d'emmagatzemar la informació en línia. A continuació, analitzem alguns dels sistemes que hi ha en l'actualitat.

- Emmagatzematge d'objectes, com ara fitxers ofimàtics.

- Emmagatzematge de fitxers en format NAS, que serveixen de repositori de fitxers. Com a exemple podem trobar l'Amazon Elastic File System o l'Amazon Windows File System.

- *Block Storage*. Àmpliament utilitzat per sistemes de bases de dades o ERP.

4. Protecció de la informació

La protecció de la informació es pot mirar des de dos vessants. En el primer, senzillament es tracta de protegir-la d'accessos no desitjats. En el segon, es tracta de protegir-la per no perdre-la.

La **informació** és un dels béns més valuosos de l'organització. Cal invertir una part important dels esforços a protegir-la.

4.1. Seguretat de la xarxa

La informació està dins la xarxa de l'organització en servidors, estacions de treball, portàtils, etc. El personal hi accedeix i, per mitjà de mecanismes d'accessos, es limita i es controla qui pot accedir a aquesta informació, de quina manera i què hi pot fer. Disposar d'un bon disseny de l'entorn d'usuari, és el primer pas per evitar accessos no desitjats a la informació.

L'altre pas per tenir sistemes segurs és evitar intrusions. És molt difícil tenir sistemes completament segurs, per tant, només podem procurar que el sistema sigui tan segur com sigui possible.

4.2. Còpies de seguretat

Les còpies de seguretat de les dades d'una organització és un dels temes més importants dins de l'administració de les dades, ja que, sense una bona política de recuperació davant de desastres, la informació es pot perdre. Bàsicament, s'ha de tenir present que les dades s'haurien de copiar d'acord amb la política de còpies establerta per aquestes; i tenir un historial de còpies de seguretat, que segons l'organització, pot ser des de pocs mesos fins a anys.

Les polítiques de les còpies de seguretat són un punt clau en la continuïtat d'una empresa davant un incident informàtic. Així, cal remarcar els punts següents a revisar dins la política de còpies de seguretat.

1) Comprovar un inventari de la informació.

2) Control d'accés a la informació, sent accessible únicament a personal autoritzat.

3) Validar les còpies de seguretat de la informació crítica.

4) Periodicitat de les còpies de seguretat. Aquí, cal tenir en compte diversos aspectes:

- Variació de les fonts de dades.
- Cost de l'emmagatzematge de les còpies.
- Obligacions legals que s'hagin de cobrir amb la còpia de dades.
- Tipus de còpia de seguretat:
  - **Completa**: es fa una còpia de totes les dades.
  - **Incremental**: únicament es copien les dades que han canviat des de l'ultima còpia.
  - **Diferencial**: es copien les dades que han canviat des de l'última còpia completa.

5) Caducitat de les còpies de seguretat.

6) Comprovació del bon estat de les còpies.

7) Xifrat de la informació, incloent les còpies de seguretat.

8) Ubicació de les còpies de seguretat. S'ha d'assegurar una bona custòdia d'aquestes còpies.

9) Si es fan les còpies al núvol, addicionalment s'han de validar els aspectes legals i normatius associats a la custòdia de les dades.

10) Finalment, un cop s'hagi considerat que la còpia no és vàlida, s'ha de destruir de forma correcta i segura.

4.3. Seguretat en bases de dades

Podem entendre una base de dades com un «conjunt de dades integrades, adequat a diversos usuaris i a diferents usos». De manera que els problemes de seguretat esdevindran per l'ús simultani d'aquestes dades.

La protecció de les dades s'ha de fer contra fallades físiques, fallades lògiques i errades humanes (siguin o no intencionades), que poden alterar o corrompre les dades, ocasionant que la base dades esdevingui inútil pel motiu que es va crear.

Qualsevol SGBD ha de proporcionar tècniques que permetin que els usuaris tinguin accés únicament a una part de la base de dades i no a la resta, de manera que els SGBD tenen un subsistema de seguretat d'autorització encarregat de garantir la seguretat d'algunes parts de la base de dades contra l'accés no autoritzat.

Les bases de dades porten mecanismes per prevenir fallades (subsistema de control), detectar-les quan s'hagin produït (subsistema de detecció) i corregir-les després que s'hagin detectat (subsistema de recuperació).

Els aspectes fonamentals de la seguretat en les bases de dades són:

- **Confidencialitat**: no proporcionar dades a usuaris no autoritzats. Inclou aspectes de privadesa (protecció de dades personals).
- **Accessibilitat**: la informació ha d'estar disponible.
- **Integritat**: permet assegurar que les dades no han estat falsejades.

4.3.1. Confidencialitat

Per facilitar l'administració, els SGDB incorporen el concepte de perfil, rol o grup d'usuaris que reuneix un conjunt de privilegis, de manera que l'usuari assignat a un grup hereta tots els privilegis del grup.

El subsistema de control d'accés s'encarrega de denegar o concedir l'accés als usuaris. Poden haver-hi diversos tipus d'autorització en un SGDB:

- **Autorització explícita**: usada en els sistemes tradicionals. Es tracta d'emmagatzemar qui pot accedir a quins objectes de la base dades i amb quins privilegis. S'acostuma a utilitzar una matriu de control d'accessos.
- **Autorització implícita**: l'autorització sobre un objecte es pot deduir a partir d'altres. Per exemple, si es pot accedir a una classe en un SGDB, es pot accedir a totes les instàncies de la classe.
- **Autorització positiva**: si hi és, indica l'existència de l'autorització.
- **Autorització negativa**: és la negació explícita d'una autorització.

4.3.2. Accessibilitat

Els sistemes de bases de dades han d'assegurar la disponibilitat de les dades a aquells usuaris que necessiten accedir-hi. Així doncs, hi ha mecanismes que permeten recuperar la base dades contra fallades lògiques o físiques que destrueixen totalment o parcialment les dades.

Uitlitats de seguretat dels SGDB

Els SGDB contenen utilitats pròpies, però des d'un punt de vista de seguretat, s'haurien de considerar alternatives com, per exemple, servidors tolerants a fallades.

El principi bàsic que sustenta la recuperació de la base de dades davant de qualsevol fallada és la redundància física. Les més habituals són les provocades per fallades elèctriques, fallades del maquinari i fallades en els dispositius d'emmagatzematge (discs).

Aquesta situació va motivar l'aparició del concepte de **transacció**. Davant de qualsevol fallada, s'ha de poder assegurar que, després d'una actualització, la base de dades quedi en un estat consistent. Per aconseguir-ho, es creen unes unitats d'execució anomenades *transaccions* que es poden definir com a seqüències d'operacions que s'han d'executar de forma atòmica. O es realitzen totes les operacions de la transacció globalment o no se'n fa cap.

4.3.3. Integritat

En aquest context, s'entén per integritat la correcció, la validesa o la precisió de les dades de la base de dades. L'objectiu és protegir la base de dades contra operacions que puguin introduir inconsistències a les dades. El subsistema d'integritat d'un SGBD ha de detectar i corregir, en la mesura que es pugui, les operacions incorrectes.

Hi ha dos tipus d'operacions que poden violar la integritat de les dades: les operacions semànticament inconsistents i les interferències a causa d'accessos concurrents.

1) Integritat semàntica. Hi ha operacions que poden vulnerar les restriccions definides en el disseny de la base de dades (per exemple, restriccions en els dominis o en els atributs). Aquestes restriccions poden ser estàtiques (també anomenades d'estat o situació) o dinàmiques (anomenades de transició).

2) Integritat operacional. En sistemes multiusuari, és imprescindible un mecanisme de control de la concurrència per conservar la integritat de la base de dades. Altrament, es podrien produir importants inconsistències derivades de l'accés concurrent.

Resum

Les dades són la base del sistema informàtic i també tenen un gran valor per a l'organització. El gran perill és que poden estar a molts llocs. Hem d'intentar educar els usuaris perquè les concentrin als servidors d'una manera natural. Això dona uniformitat i seguretat a les dades.

Malgrat tot, això no és suficient per satisfer una de les grans demandes de l'organització: extreure noves conclusions a partir de la informació. L'única manera és concentrant la informació en una única aplicació, un servidor de bases de dades.

Si optem per aquesta solució, hem d'instal·lar una interfície homogènia en aquest servidor de bases de dades, una arquitectura de dades, un ERP o un Data Warehouse.

Atès que s'ha convertit en un actiu molt valuós, la informació cal que sigui protegida. Per tant, calen tant polítiques de còpies de seguretat com proteccions d'accés a les dades.

Activitats

1. Observeu organitzacions del vostre entorn (botigues, empreses, bancs, etc.) i intenteu trobar les fonts de dades que hi pot haver.

2. Fixeu-vos en l'entorn. Veureu la mateixa informació amb molts formats diferents, per exemple, l'hora (digital, analògica, etc.). Com ho guardaríeu en una base de dades? Mireu i trobareu altres exemples de dades amb multitud de formats diferents.

Exercicis d'autoavaluació

1. Enumereu algunes de les fonts de dades dels llocs següents:

a) Un hospital.
b) Una discoteca.

2. Un dels comercials de l'organització us diu que no vol fer servir una base de dades per fer els càlculs de venda dels productes o per guardar els productes i els increments, perquè mai no ha treballat així: ell negocia directament amb el client i, com a molt, pot introduir les dades en un full de càlcul al seu portàtil (i encara us farà un gran favor, si ho fa). Què li diríeu?

3. Concentraríeu les dades als servidors?

a) Sí, perquè així es poden compartir més fàcilment.
b) No, és més pràctic distribuir-les als llocs on calen.
c) No, perquè podríem col·lapsar els servidors i la xarxa.
d) Sí, perquè és més senzill fer consultes a la informació de l'organització.
e) Sí, perquè eliminem completament la redundància.
f) a i d.

4. Quina d'aquestes frases sobre els servidors de bases de dades és falsa?

a) Els servidors de bases de dades concentren la informació dispersa en un sol lloc.
b) Els ERP tenen com una de les seves bases un servidor de bases de dades.
c) Un servidor de bases de dades amb la informació de l'organització facilita les consultes de la direcció.
d) Amb un servidor de bases de dades, el trànsit de la xarxa disminueix perquè els formats de sortida són homogenis.
e) Un servidor de bases de dades homogeneïtza la manera de guardar la informació.

5. Una d'aquestes tasques no és responsabilitat de l'administrador de dades.

a) Evitar al màxim la duplicitat d'informació dins l'organització.
b) Assegurar la disponibilitat de les dades als usuaris.
c) Configurar bases de dades corporatives.
d) Connectar les bases de dades amb el servidor web.
e) Vetllar pel funcionament correcte de les bases de dades.

Solucionari

Exercicis d'autoavaluació

1. a) Es genera informació a molts llocs. Alguns són:

Recepció de pacients.
Prescripció de receptes.
Visites a totes les consultes.
Tot el departament de comptabilitat.
Tot el departament de comandes a proveïdors.
Departament de nòmines.

b) També es genera informació a molts llocs. Alguns són:

Màrqueting.
Contractació d'espectacles.
Comptabilitat.
Facturació.
Nòmines.
Comandes a proveïdors.
Seguretat.

2. Si una persona us exposa aquest problema, se l'ha de convèncer amb arguments com, per exemple, els següents:

- L'organització treballa com una unitat i, per tant, tots els preus estan a un sol lloc, dins d'una base de dades en un servidor.
- Pot agafar les dades i manipular-les, però si negocia canvis de preu els ha de reflectir en la base de dades per tal de saber:
  - Quin és el venedor més competitiu.
  - Quin és el venedor que treballa millor.
  - Quin és el venedor que factura més.

D'aquesta manera, l'organització pot assignar incentius (econòmics) o d'un altre tipus als comercials amb millors índexs de venda.

Si només té les dades en local i perd l'ordinador, li roben, etc., ni ell ni ningú de l'organització no les podrà recuperar, per la qual cosa l'organització tindrà una pèrdua d'informació i ell pot perdre els incentius que hem comentat.

La informació de l'organització és propietat de l'organització. Per tant, és l'organització qui estableix les directrius de funcionament i no les persones que la manipulen.

3. f

4. d

5. d

Glossari

base de dades *f* Terme genèric que indica un lloc on guardar dades.

dada *f* Registre dels successos.

data warehouse Base de dades que emmagatzema una gran quantitat de dades transaccionals integrades per ser usades per a l'anàlisi.

data mart Conjunt de fets i dades organitzades per a suport decisional basades en la necessitat d'una àrea o departament específic. Les dades estan orientades a satisfer les necessitats particulars d'un departament i només tenen sentit per al personal d'aquest departament.

data mining Anàlisi de les dades per descobrir relacions, patrons o associacions desconegudes.

diccionari de dades *m* Compendi de definicions i especificacions per a les categories de dades i les seves relacions.

dimensió *f* Entitat independent, dins del model multidimensional d'una organització, que serveix com a clau de recerca (actuant com a índex) o com a mecanisme de selecció de dades.

DSS *m* Vegeu **sistema de suport de decisions**.

enterprise resource planning Vegeu **planejament de recursos de l'empresa**.

ERP *m* Vegeu **planejament de recursos de l'empresa**.

font de dades *f* Qualsevol element (que pertany o no a l'organització) que crea dades que fa servir el sistema informàtic.

informació *f* Processament de les dades perquè tinguin sentit.

llenguatge de consultes estructurat *m* Llenguatge que es fa servir per interrogar els gestors de bases de dades.
en.: *structured query language*
sigla: SQL

planejament de recursos de l'empresa *m* Sistema integrat de gestió de la informació que pretén donar una solució completa al problema de la informació dins d'una organització.
en.: *enterprise resource planning*
sigla: ERP

OLAP (*online analytical processing*) *m* Conjunt de principis que proveeixen un entorn de treball dimensional per a suport decisional.

OLTP (*online transaction processing*) *m* Sistema transaccional que manté les dades operacionals de l'organització.

servidor de bases de dades *m* Sistema que guarda dades i que, per algun mecanisme, generalment mitjançant peticions per mitjà de la xarxa, se li demana informació o se n'hi introdueix.

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

sistema de gestió de bases de dades *m* Programari que gestiona dades d'una manera ordenada per guardar-les i recuperar-les.
sigla: SGBD

sistema de suport de decisions *m* Sistema d'aplicacions automatitzades que assisteix a l'organització en la presa de decisions mitjançant una anàlisi estratègica de la informació històrica.

SQL *m* Vegeu **llenguatge de consultes estructurat**.

structured query language *m* Vegeu **llenguatge de consultes estructurat**.

Bibliografia

Barcelo García, M.; Pastor i Collado, J. (1999). *Gestió d'una organització informàtica*. Barcelona: Universitat Oberta de Catalunya.

Date, C. J. (2000). *An Introduction to Database Systems*. Estats Units: Addison Wesley.

Elmasri, R.; Navothe, S. (2000). *Fundamentals of Database Systems*. Estats Units: Addison Wesley.

Inmon, W. (2005). *Building de Data Warehouse* (4a. edició). Estats Units: Wiley.

Kimball, R.; Ross, M. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling* (2a. edició). Estats Units: Wiley.

Pfleeger, C. (1997). *Security in Computing*. Estats Units: Prentice Hall.

Prague, C.; Irwin, M. (1996). *El libro de Access para Windows 95*. Madrid: Anaya Multimedia.