Les dades a la factoria d'informació corporativa

PID_00270636

Juan Vidal Gil
Carles Llorach Rius

Temps mínim de dedicació recomanat: 3 hores

Universitat Oberta de Catalunya

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.

Carles Llorach Rius

Màster en Gestió d'Empreses - MBA per la Universitat Rovira i Virgili i enginyer en Informàtica per la Universitat Politècnica de Catalunya. Professor col·laborador a la Universitat Oberta de Catalunya.

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

Segona edició: febrer 2020

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

Índex

Introducció

En altres mòduls d'aquesta assignatura hem vist els magatzems de dades des d'una perspectiva general de l'organització. Atesa la importància que tenen les dades per si mateixes, en aquest mòdul ens centrarem a estudiar-los des del punt de vista de la gestió/governança de les dades.

Amb això, ens estem referint a la integració de les dades, a la gestió de les dades mestres, a la qualitat de les dades, a la gestió de les metadades i a tots aquells aspectes legals i ètics que hem de contemplar en el tractament de les dades.

Objectius

En aquest mòdul es pretén oferir una visió global dels processos de gestió de la dada en l'organització, posant l'accent en aquells estretament relacionats amb els magatzems de dades.

Mitjançant l'estudi, s'aconseguiran els objectius següents:

1. Prendre consciència del valor i importància de les dades com a actiu de l'organització.

2. Entendre la funció i l'àmbit de les activitats de govern de la dada.

3. Conèixer la importància i la gestió dels processos de qualitat de la dada.

4. Aprofundir en el concepte de metadada coneixent les seves diferents funcions i usos.

5. Conèixer els aspectes legals i ètics per al tractament adequat de les dades.

1. La importància de les dades

1.1. Dades

Les organitzacions desenvolupen sistemes informàtics on resideixen les seves dades: en el cas de la base de dades operacional, l'important són les dades actuals, mentre que, en el cas del magatzem de dades, la importància rau en les dades històriques.

En els dos entorns, la dada és molt important. Aquelles organitzacions que ho entenen així i que actuen d'acord amb això solen rebre el nom d'organitzacions orientades a la dada (en anglès, *data-driven*).

En aquest mateix context, sorgeix el concepte de governança de dades (en anglès, *data governance*, DG), que és una disciplina de control de qualitat per a l'avaluació, la gestió, l'ús, la millora, la supervisió, el manteniment i la protecció d'informació/dades de l'organització.

Alguna de les activitats que típicament s'emmarquen en la governança de dades és la gestió de dades mestres (en anglès, *master data management*, MDM) i la neteja de dades (en anglès, *data cleaning* o *data scrubbing*), entre d'altres.

La integritat de les dades és un problema important en la majoria de les organitzacions, i el desenvolupament d'un magatzem de dades s'utilitza amb freqüència com un vehicle per millorar la qualitat de les dades de manera significativa. L'exactitud de les dades pot significar estalvis considerables en àrees com màrqueting, atenció al client, etc. Existeixen estudis duts a terme per organitzacions com ara Gartner Group (una de les principals empreses de prospecció de mercat) que estimen en un 4% els estalvis obtinguts a partir de la millora de la integritat de les dades en les organitzacions.

Així apareix el concepte de *data warehouse governance*, que recull aquelles pràctiques centrades en com es creen les dades, com són recollides, tractades i manipulades, emmagatzemades, posades a disposició per al seu ús o retirades.

Denominarem programa al conjunt de pràctiques que, podent variar significativament, depenent del seu enfocament: en el compliment (*compliance*), en la integració de dades, en la gestió de dades mestres (MDM), etc., estan alineades amb les polítiques corporatives: en un àmbit de lògica de negoci, estratègia tecnològica, seguretat, etc.

Les activitats de les organitzacions són generalment horitzontals i afecten diversos departaments o funcions (comercial, trànsit, administració, etc.). L'organització horitzontal també rep el nom de «per activitats o processos» i és totalment contrària a l'organització tradicional vertical, per departaments o funcions. L'organització «vertical» es visualitza com una agregació de departaments independents els uns dels altres i que funcionen autònomament. Un bon desplegament dels nostres programes requereix una concepció àmplia (horitzontal) de la nostra organització.

Les organitzacions necessiten passar del govern informal al govern de dades formal, quan es produeix alguna de les situacions següents:

- L'organització arriba a ser tan gran que la gestió tradicional no és capaç de lliurar les dades relatives a activitats multifuncionals/transversals.

- Els sistemes de dades de l'organització es fan tan complicats que la gestió tradicional no és capaç de lliurar les dades relatives a activitats multifuncionals/transversals.

- Els arquitectes de dades de l'organització, els equips de SOA (*service-oriented architecture*) o altres grups enfocats horitzontalment, necessiten una visió corporativa (en lloc de fragmentada en sitges) de les preocupacions i les opcions relatives a les dades.

- La regulació: el compliment legal o l'existència de requisits contractuals que ho exigeixen.

Un *data warehouse* interactua, per definició, amb gran part de l'organització. Les polítiques, processos i procediments del programa s'han de comunicar clarament a tots els afectats per assegurar que l'esforç requerit generarà benefici.

La informació prové de fonts internes (sistemes de producció) i externes (fins a un 20%) i comporta problemes com la saturació d'informació, la dificultat d'accés, no ser selectiva, etc. Tot això haurà de ser contemplat a l'hora de dissenyar els nostres programes.

2. Integració de dades

En el mòdul «Construcció de la FIC» vam veure la importància de la integració de dades en els processos d'actualització dels magatzems de dades, concretament en el component d'integració i transformació de dades. La correcta integració de dades en un magatzem és una qüestió crítica per a la seva explotació correcta. En aquest apartat veurem les diferents disciplines associades a la integració de dades.

2.1. Disciplines que intervenen en la integració de dades

Sabem que el magatzem de dades té un paper fonamental pel que fa a la consolidació i integració de la informació: permet passar del que anomenem teranyina d'entorn operacional a un entorn centralitzat i integrat. També sabem que la integració de les dades suposa un autèntic repte, si tenim en compte la disparitat d'orígens, formats, eines i sistemes que habitualment tenen en les companyies.

La **integració de les dades** no és un problema exclusiu dels magatzems de dades, sinó que s'aplica a tots els sistemes que gestionen informació i és una activitat que té tot un conjunt de disciplines associades.

Algunes d'aquestes disciplines poden ser: la qualitat de les dades, la gestió de dades mestres, la definició de mètriques homogènies, el cicle de vida de la dada, etc. Aquestes disciplines només són part de les disciplines que intervenen en els processos de gestió de dades en les companyies i es poden englobar dins d'una altra disciplina denominada *govern de la dada*.

2.2. Govern de la dada

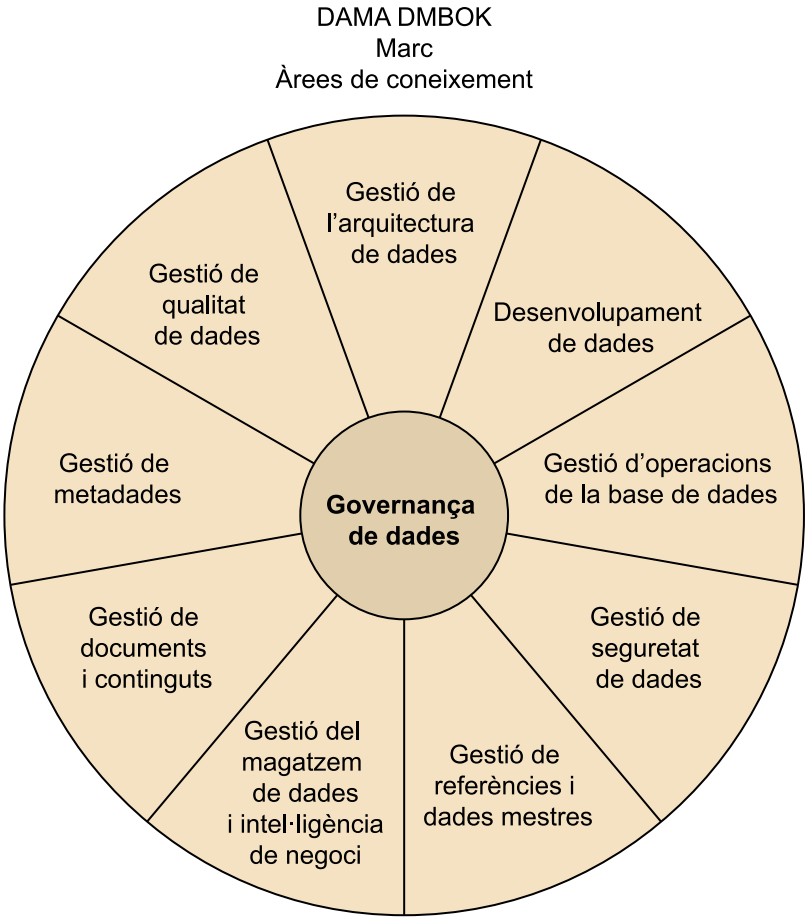
El **govern de la dada** és una disciplina empresarial important l'objectiu de la qual és proporcionar més control sobre la creació, maneig, manteniment, emmagatzematge, ús i intercanvi d'informació vital per al negoci.

Segons el diccionari de la DAMA (Data Management Agency), el govern de les dades o la governança de dades són els exercicis de control i autoritat (planificació, monitoratge i millora) sobre la gestió de les dades.

La governança de dades es compon del conjunt d'àrees següents:

- Arquitectura de dades: anàlisi i disseny.
- Gestió de bases de dades.
- Gestió de seguretat de la dada.
- Gestió de qualitat de la dada.
- Gestió de dades mestres.
- Gestió de sistemes d'intel·ligència de negoci i emmagatzematge de la dada.
- Gestió de documents i continguts.
- Gestió de metadades.

Figura 1. Àrees que considera la governança de dades



Font: www.dama.org.

En aquest mòdul estudiarem aquelles disciplines del govern de la dada més relacionades directament amb els magatzems de dades, com poden ser:

- Qualitat de la dada.
- Gestió de metadades.

3. Qualitat de la dada

Tal com ja s'ha explicat en altres mòduls, la qualitat de la dada és una qüestió crucial per als magatzems de dades i, en general, per a l'organització. La falta de qualitat de les dades és un dels problemes principals als quals s'enfronten els responsables de sistemes d'informació i les empreses, perquè representa clarament un dels problemes «ocults» més greus i persistents en qualsevol organització.

La **gestió de dades** constitueix un recurs estratègic en l'organització i la seva qualitat, un punt crucial en aquesta gestió.

Una correcta gestió de la qualitat de les dades ens aportarà els beneficis següents:

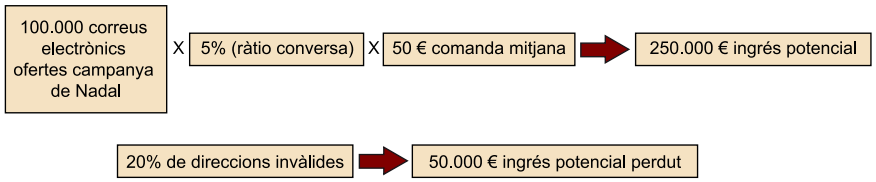
- Visió única del client-usuari/producte-servei/proveïdors.
- Millora de la comunicació client-usuari/proveïdors.
- Estalvi de temps de conciliació de la informació.
- Garantia de la correcta unificació de bases de dades (fusions d'empreses).
- Confiança en la dada, millora de processos de *reporting* i analítica.

Exemple de processos empresarials que es beneficien d'una gestió correcta de la qualitat de la dada

- Campanyes de màrqueting.
- Anàlisi de geomàrqueting.
- Compliment de les normatives. Protecció de dades (LOPD).
- Estalvi de costos d'errors de facturació, enviaments, comunicació amb clients.
- Processos de detecció del frau.

A la figura 2 podem veure el cost que pot tenir una qualitat baixa de la dada en una campanya de màrqueting per correu electrònic.

Figura 2. Exemple de l'impacte econòmic d'una baixa qualitat de la dada



3.1. Objectius de la qualitat de la dada

Amb la finalitat de garantir la qualitat de les dades, els processos de qualitat d'aquestes dades busquen els objectius següents:

- **Precisió de les dades**: que cada dada sigui fidel representant del que la funció que se li atribueix requereix, fent-ho de la manera establerta.

- **Confiabilitat de les dades**: que la dada que representa la informació sigui coherent i estable.

- **Completitut de les dades**: que garanteixi que ni en les mateixes dades, ni en els registres o taules on s'emmagatzemen faltin camps o valors, que tot estigui complet.

- **Conformitat de les dades**: que es respectin les condicions de format establertes a l'hora de donar d'alta la dada.

- **Consistència de les dades**: que, a més de garantir que la dada és correcta quant als seus atributs, no vulneri cap regla de negoci.

- **Unicitat de les dades**: que no existeixin duplicitats.

El compliment d'aquests objectius garanteix una qualitat adequada en la dada, i per revisar el seu compliment hem d'establir tot un conjunt de processos i comprovacions.

3.2. Etapes principals en la gestió de la qualitat de la dada

Aconseguir els objectius de qualitat indicats en la secció anterior suposa la realització d'una sèrie d'etapes ben definides:

- **Perfilat de la dada**: processos encaminats a explorar les fonts d'origen, obtenint informació estadística sobre la font (rangs de valors, distribució, nuls, valors únics, patrons).

- **Validació de la dada**: bateria de comprovacions encaminades a assegurar la correcció, consistència, conformitat i completitut de les dades.

- **Neteja de la dada**: detecció i correcció d'errors en les dades (falta de completitut, inconsistències, incorreccions, etc.) ja sigui modificant, o bé eliminant els registres afectats.

- **Enriquiment de la dada**: processos encaminats a millorar, depurar, normalitzar o completar la informació d'una font, utilitzant altres fonts complementàries.

3.2.1. Perfilat de la dada

Els processos de perfilat de la dada ens permeten realitzar una exploració prèvia per obtenir informació estadística de les dades que també ens poden ajudar a conèixer la qualitat de la informació d'origen. Hi ha processos de perfilat d'estructura i de contingut.

Exemples d'operacions en un perfilat de dades

Alguns exemples d'operacions de perfilat de dades que es poden realitzar sobre una entitat de dades són les següents:

a) A la taula:

- Calcular el volum de registres.
- Verificar el compliment de regles de negoci.
- Verificar la integritat referencial (pare-fill). Detectar valors orfes.
- Detectar dependències entre columnes (correlacions).

b) A la columna:

- Obtenir els valors únics columna, duplicats, freqüències.
- Calcular la columna: mitjana, màxim, mínim, desviació típica, variància.
- Comprovar el tipus de dada, longitud, distribució de longituds.
- Revisar el nombre de nuls, blancs.
- Comprovar la distribució de patrons (dd/mm/aaaa, XX-XXX).
- Ajustar a patrons predefinits (adreces, codi postal, correu electrònic, telèfon, etc.).
- Ajustar a patrons definits per l'usuari (expressions regulars).

3.2.2. Validació de la dada

Els processos de validació de la dada realitzen les comprovacions necessàries per assegurar la correcció, consistència, conformitat i completitut de les dades. Existeixen dos tipus de validacions:

a) Validacions tècniques: totes les que ens garanteixen la consistència tècnica de les dades, evitant duplicitats, camps nuls, *outliers*, falta d'integritat referencial, etc.

b) Validacions de negoci: totes les que ens garanteixen la consistència de les dades en funció de regles de negoci.

Exemples de validacions tècniques

- Duplicats: detecció de duplicats per repetició clau exacta i mitjançant algoritmes de comparació de cadenes.
- Camps obligatoris: validar que estiguin informats tots els camps obligatoris.
- Tipus de dades: tipus de dada esperada (numèric, alfanumèric).
- Consistència claus foranes: validar integritat referencial entre les claus de taules referenciades (pare-fill).
- Conciliació entre fonts: conciliació de registres entre font origen i font destí (agregats, comptatge de registres).

Exemples de validacions de negoci

- Rang de valors: per a determinades variables podem tenir un rang de valors possible, exemple: l'edat no pot ser negativa.
- Patró del camp: el camp ha de tenir un patró establert (exemple: correu electrònic xxxx@yyyy.zzz).
- Codificació interna: variables que compleixen en la seva codificació interna (exemple: DNI).
- Compliment de regles de negoci: validacions de negoci particulars de cada font i negoci concret.

Una problemàtica molt habitual en els processos de validació de la dada és la relacionada amb la deduplicació de la dada.

La **deduplicació** és un procés que persegueix la identificació de duplicats per diferents criteris. Els processos de deduplicació són imprescindibles no només per eliminar registres i dades redundants, sinó també per a projectes de consolidació de fonts d'informació i enriquiment de dades.

Existeixen diferents tècniques per identificar registres duplicats. El cas més senzill és quan els nostres registres coincideixen per clau; no obstant això, hi ha casos de duplicats en els quals la clau no coincideix exactament, fins i tot encara que es tracti del mateix registre. Sovint solen ser casos en els quals hi ha algun camp de la clau amb nuls, camps de tipus cadena de caràcters en els quals hi ha errors tipogràfics, camps amb el mateix valor i diferent format (per exemple, camps de tipus data). Per a aquests casos en els quals no busquem una clau exacta hem de realitzar creus i podem fer-ho amb diferents tipus d'encreuament o *matching*:

a) *Matching* determinista: es comparen els diferents atributs associats a l'entitat i s'obté un resultat positiu o negatiu. Abans de la comparació se solen realitzar transformacions, normalitzacions, codificacions i neteges prèvies a la comparació. A la figura 3 es pot veure un exemple d'aquest cas.

b) *Matching* probabilístic: mitjançant algoritmes específics es comparen diferents atributs. Aquests algoritmes retornen un percentatge que indica el grau de similitud entre els atributs comparats. Els algoritmes de comparació hauran de ser adequats per a al tipus de dades, ja que no és el mateix comparar una cadena de text lliure (com un nom, raó social, descripció de producte, etc.) que un codi (telèfon, CIF, codi postal, número de ref., etc.). Igual que amb el *matching* determinista, és convenient realitzar transformacions, codificacions i neteges prèvies. Finalment, s'agafen tots els percentatges obtinguts de les diferents comparacions i es realitza una mitjana ponderada. Determinats

atributs poden tenir més pes que d'altres, per exemple, en comparar empreses tindrà més pes la raó social que el telèfon. Un exemple seria el que es presenta en el cas b) de la figura 6.

Figura 3. Exemples de *matching* determinista (a) i probabilístic (b)

a)

Raó social	CIF	Adreça	Codi postal	Telèfon	Municipi	Població
Laboratorios Roma S.A.	A11112222	Polígono industrial Los Sauces	30140	9611112233	Santomera	Murcia
✗	✗	✗	✓	✓	✗	✓
Raó social	CIF	Adreça	Codi postal	Telèfon	Municipi	Població
Roma S.A.		Polígono Los Sauces, C/Enebro, 10	30140	9611112233	Santmera	Murcia

b)

Raó social	CIF	Adreça	Codi postal	Telèfon	Municipi	Població
Laboratorios Roma S.A.	A11112222	Polígono industrial Los Sauces	30140	9611112233	Santomera	Murcia
40%		70%	100%	100%	90%	100%
Raó social	CIF	Adreça	Codi postal	Telèfon	Municipi	Població
Roma S.A.		Polígono Los Sauces, C/Enebro, 10	30140	9611112233	Santmera	Murcia

3.2.3. Neteja de la dada

Els processos de neteja de la dada corregeixen els errors detectats en les dades (falta de completitut, inconsistències, incorreccions, etc.) modificant o eliminant els registres afectats. Existeixen diferents alternatives per corregir la dada:

- **Eliminació de registres**: pot venir obligada per possibles errors (registres duplicats).
- **Valor indeterminat**: en casos de valor no informat d'una columna, una alternativa pot ser definir un codi indeterminat per a aquesta columna.
- **Estimació del valor**: en casos de valor no informat, una alternativa pot ser estimar el valor (extrapolacions, mitjana de valors de columna).
- **Correcció de paraules**: correcció d'errors ortogràfics o tipogràfics, eliminació de blancs innecessaris.
- **Normalització i estandarditzación de dades**: correcció de camps que han de seguir valors estàndards (carrers, municipis, ciutats, noms de persones).

3.2.4. Enriquiment de la dada

Els processos d'enriquiment de la dada permeten millorar les dades existents complementant la informació de les fonts amb altres fonts, ja siguin internes o externes. És molt comú l'ús de bases de dades estàndards per completar informació geogràfica (ciutats, municipis, codis postals, carrers, coordenades, etc.) o la informació sociodemogràfica dels clients (edat, estat civil, nombre de fills, etc.), bases de dades d'empreses. En altres casos els camps informats seran modificats pel seu valor normalitzat (carrers, municipis, etc.).

3.3. Implementació dels processos de gestió de la qualitat de la dada

Les etapes definides en l'apartat anterior donen com a resultat la identificació d'una sèrie de regles de validació, correcció, neteja i enriquiment de la dada. Aquestes regles han de ser implementades en diferents processos de gestió de la dada a fi de garantir la qualitat de la dada en l'organització.

Alguns dels processos més crítics són els següents:

- **Component de transformació i integració de la FIC**: ha de garantir que la informació que es consolida en els magatzems de dades estigui depurada.

- ***Hub* MDM**: s'implementaran aquestes regles per garantir la qualitat en la integració de dades sobre les dades mestres.

- **Sistemes operacionals que tractin informació crítica de la companyia**: així s'aconsegueix que la informació en origen sigui com més fiable millor.

- **Processos de monitoratge de la qualitat de la dada**: sobre algunes entitats es dissenyaran i aplicaran processos per mesurar la qualitat de les seves dades.

Respecte a l'últim procés esmentat, el de monitoratge, es definiran una sèrie de mètriques per mesurar la qualitat de les dades d'entitats determinades. Dels resultats obtinguts en els monitoratges podrem obtenir una sèrie d'indicadors o ràtios sobre la qualitat de la dada. També podem crear alertes sobre aquestes ràtios. L'objectiu final serà tenir un quadre de comandament sobre qualitat de la dada.

3.4. Tendències en els processos de gestió de la qualitat de la dada

En els últims anys en els processos de gestió de dades en entorns *Big Data* s'ha desenvolupat el concepte de *Data Curation*. L'objectiu d'aquesta tècnica és automatitzar tots els processos de neteja, estandardització i enriquiment de la dada en funció d'algoritmes de *machine learning* i sistemes experts.

Aquests processos apliquen diferents tècniques analítiques per millorar la qualitat de les dades, relacionar diferents fonts i enriquir les dades partint d'un conjunt nombrós i heterogeni de fonts d'origen de dades, que poden contenir un volum molt elevat de registres i informació estructurada i no estructurada.

Dins de les múltiples tècniques que es poden aplicar, n'esmentem algunes:

- Identificació de registres de diferents fonts que fan referència a la mateixa entitat de dades, la qual cosa permet establir relacions (*clustering*, distàncies, etc.).

- Identificació de registres duplicats (*clustering, matching*, etc.).

- Ús de patrons per identificar determinades entitats (noms, telèfons, direccions, etc.).

- Revisió de columnes: obtenció de rellevància de termes en col·leccions o documents, comparació de distribucions per a columnes numèriques.

- Mineria de textos per identificar patrons, relacions i enriquir la in- formació.

- Obtenció de probabilitats en la identificació d'entitats.

Els resultats obtinguts es relacionen amb un nivell de confiança i són visualitzats i presentats de manera que afavoreixen la intervenció manual.

4. Gestió de metadades

En altres mòduls s'ha parlat sobre la importància de les metadades en el context de la FIC. Es van tractar diferents tipus de metadades de la FIC (les metadades de fonts de dades, les del magatzem de dades i les del component d'integració i transformació), que ens ajuden a gestionar la FIC, fer-la evolucionar i, en alguns casos, són consultables per diferents tipus d'usuaris amb la finalitat de conèixer millor l'estructura dels models, com es relacionen les entitats o com es transformen les dades.

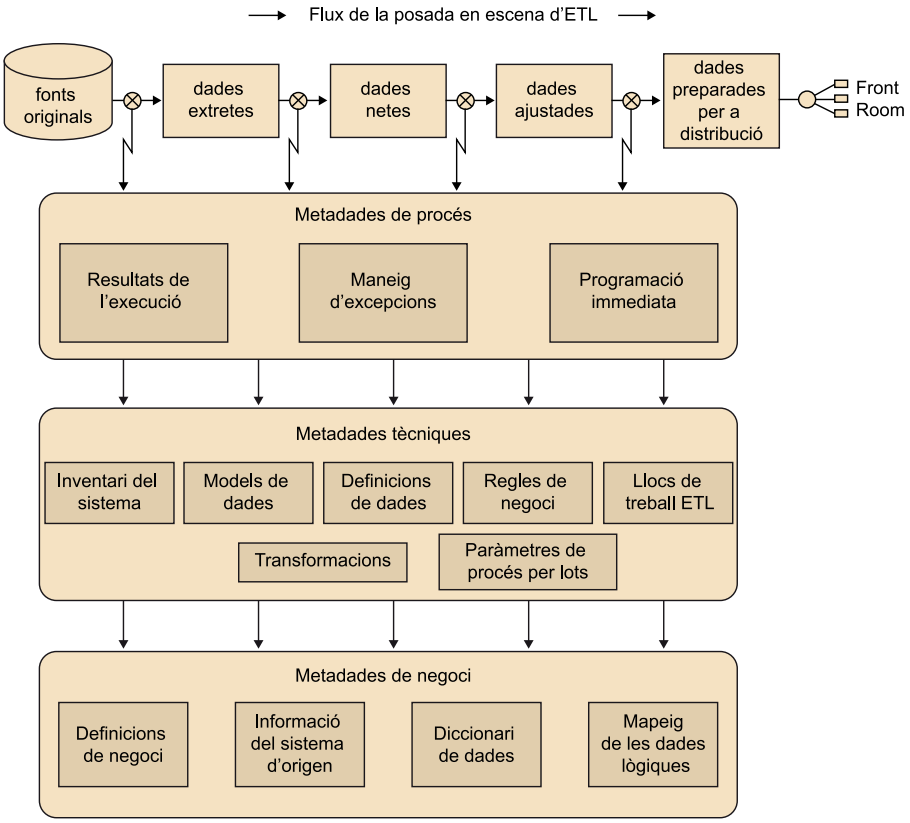
Les metadades no són àmbit exclusiu de la FIC i són un element crític en les activitats de govern de la dada.

Les metadades ens donen informació sobre com emmagatzemar les dades, com obtenir-les, com explotar-les i com es relacionen les entitats i els processos.

4.1. Tipus de metadades

Existeixen diferents tipus de metadades segons el seu ús i diferents possibles classificacions, una de les més esteses és la que proposa Ralph Kimball i que es mostra a la figura 4.

Figura 4. Tipus de metadades



Font: www.kimballgroup.com.

Aquesta classificació proposa els següents tipus de metadades:

- **Metadada de negoci**: descriu les dades des d'un punt de vista de negoci, i mostra un diccionari de dades que tradueix el model de dades a termes de negoci.

- **Metadada tècnica**: descriu els aspectes tècnics com ara tipus de dades, longituds, relacions entre taules, passos de transformació, composició i dependències dels *jobs*, etc.

- **Metadada de processos**: mostra dades sobre les execucions (registres llegits, escrits, rebutjats, temps de procés, errors, *logs* de processos, etc.).

4.2. Reptes en la gestió de les metadades

El creixement exponencial de la informació dels últims anys obliga a una millora en els processos de govern de dades que passa per una gestió més òptima de les metadades. A continuació mostrarem els reptes principals que s'han d'afrontar en la seva gestió:

4.2.1. Gestió integral de les metadades

El concepte qualitat de metadades sorgeix en grans corporacions que compten amb milers d'atributs i indicadors. Es tracta d'una problemàtica d'integració i/o d'eines de gestió de metadades, no de qualitat de dades en si.

Existeixen solucions que permeten integrar les metadades generades en diferents components a fi de tenir una visió comuna. Hem d'integrar la gestió de metadades que realitzem en el component de transformació i integració de la FIC amb la gestió de les metadades que es duu a terme en el *hub* MDM i les orientades a l'explotació de la dada.

Aquesta gestió integral ha de marcar un llenguatge de negoci comú per unificar les definicions i els criteris que cal aplicar dels indicadors, atributs i càlculs comuns. Per això són necessaris estàndards de metadades.

4.2.2. Estàndards de metadades

Per compartir les metadades entre components, aquests han de «parlar» el mateix idioma en aquest aspecte. Un estàndard de definició de metadades representa aquest idioma comú.

Pel que fa a estàndards, tenim el *common warehouse metadata* (CWM), que ens ajuda a definir i compartir metadades entre components de la nostra arquitectura i solucions de programari.

És necessari potenciar l'ús d'aquest tipus d'estàndards per millorar la gestió de les metadades, definir-les de manera eficient, estàndard i senzilla de compartir.

4.2.3. Informació semiestructurada o no estructurada

És un fet que la informació semiestructurada o no estructurada suposa una font d'informació cada vegada més rellevant en les companyies. Tot i que per la seva naturalesa no es tracti d'una informació emmagatzemable en estructures ben definides, sí que és necessària una capa de metadades que serà més lleugera que en el cas d'informació estructurada, però que ens ajudarà a emmagatzemar-la i gestionar-la.

5. Aspectes legals i ètics de les dades

5.1. Normativa legal de les dades

El primer aspecte que hem de considerar quan treballem amb dades és si estan subjectes a alguna normativa legal vigent on es duu a terme l'activitat organitzativa o contractual que hàgim pogut subscriure amb els nostres proveïdors, clients o socis.

Ens trobem amb dos principis generals del dret que ens són aplicables: la protecció de dades, que està configurada en un àmbit europeu com un dret fonamental dels ciutadans; i la transparència, que regula país a país l'accés a la informació en poder de les administracions públiques, premissa indispensable per a la rendició de comptes.

5.1.1. Protecció de dades

La Llei orgànica 3/2018, de 5 de desembre, de protecció de dades i garantia dels drets digitals, desplega la norma principal de la Unió Europea en la matèria, que és el Reglament 2016/679, general de protecció de dades (RGPD). Però, amb anterioritat, ja el 1981, el Consell d'Europa havia adoptat el Conveni núm. 108, sobre la protecció de les persones, que és l'únic instrument internacional vinculant sobre protecció de dades.

Nota

El Reglament general de protecció de dades va entrar en vigor el 25 de maig del 2016.

Relacionem, a continuació i de manera molt resumida, els aspectes principals que cal considerar:

- L'interessat té el **dret a ser informat** quan s'hagin de recollir les seves dades personals.

- El responsable del tractament ha de proporcionar el seu nom i adreça, la finalitat del tractament, els destinataris de les dades i qualsevol altra informació que sigui necessària per garantir un tractament lleial amb l'interessat (art. 10 i 11).

- **Tractament**: les dades es poden tractar només sota les circumstàncies següents (art. 7):
  - Quan l'interessat ha donat el seu consentiment.

  - Quan és necessari per a l'execució d'un contracte o per mesures precontractuals.

  - Quan és necessari per al compliment d'una obligació jurídica.
  - Quan és necessari per protegir els interessos vitals de l'interessat.
  - Quan és necessari per al compliment d'una missió d'interès públic o inherent a l'exercici del poder públic conferit al responsable del tractament.
  - Quan és necessari per al propòsit de l'interès legítim perseguit pel responsable del tractament, sempre que no prevalguin l'interès o els drets i llibertats fonamentals de l'interessat.

- L'interessat té **el dret d'accés** a totes les seves dades tractades. Fins i tot té el dret de demanar la **rectificació, supressió** o **bloqueig** de dades que siguin incompletes, inexactes o que no siguin tractades d'acord amb les disposicions de la Directiva de Protecció de Dades (art. 12).

- **Legitimitat**: les dades personals només poden ser recollides per a finalitats determinades, explícites i legítimes, i no poden ser tractades posteriorment de manera incompatible amb aquestes finalitats (art. 6b).

- **Proporcionalitat**: les dades personals tractades només poden ser les adequades, pertinents i no excessives en relació amb la finalitat per a la qual van ser recollides. Les dades han de ser exactes i, quan sigui necessari, actualitzades; s'hauran de prendre les mesures raonables per suprimir o rectificar les dades inexactes o incompletes, respecte a la finalitat amb la qual van ser recopilades. Les dades s'han de conservar d'una manera que permeti la identificació dels interessats durant un període no superior al necessari per a la finalitat per a la qual van ser recopilades (art. 6).

- **Tractaments especials**: s'apliquen quan les dades personals són sensibles: origen racial o ètnic, opinions polítiques, conviccions religioses o filosòfiques, afiliació a sindicats, salut o sexualitat (art. 8).

- L'interessat pot **oposar-se** en qualsevol moment al tractament de dades personals, amb la finalitat de prospecció de mercat (art. 14).

5.1.2. Transparència

Podríem considerar dins d'aquest camp tot allò destinat a reforçar la confiança de la societat en les institucions públiques i organitzacions, a través de l'impuls del bon govern, la transparència i la rendició de comptes de les seves activitats.

A diferència de la protecció de dades, no existeix una norma europea o internacional que reguli la transparència de manera universal, i són els països els que legislen per garantir el dret a la informació pública.

En la majoria dels casos, el desplegament d'aquestes lleis està suportat per un **portal de transparència** que proveirà la navegació pel portal i la presentació de dades. S'haurà d'establir l'arquitectura tecnològica que s'ajusti millor a aquests propòsits.

Amb aquesta mateixa voluntat, fa anys que han sorgit iniciatives que fomenten la transparència/obertura de les organitzacions a través de la publicació de les seves dades. Les **dades obertes** (*open data*) són aquelles que es consideren accessibles i reutilitzables, sense exigència de permisos específics.

D'aquesta manera, estem creant les condicions per al desenvolupament del mercat de la **reutilització de la informació**, així com la **interoperabilitat** entre diferents organitzacions.

5.1.3. Altres principis

A més de la protecció de dades i la transparència, existeixen altres lleis que regeixen el tractament de dades i hem d'actuar d'acord amb aquestes: la «Llei de cookies» i la «Llei de comerç electrònic», recollides en la Llei de serveis de la societat de la informació (LSSI).

També existeix un marc d'actuació definit per un conjunt de principis relacionats amb l'ús de dades. Els més rellevants i que hauríem de considerar amb atenció són: privacitat, seguretat, identitat, confidencialitat, etc.

5.2. Ètica de les dades

L'ètica en la informàtica és una disciplina nova que pretén obrir-se camp dins de les ètiques aplicades, per la qual cosa trobem diverses definicions:

- Mario González Arencibia la defineix com «la disciplina que analitza els problemes ètics que són creats per la tecnologia dels ordinadors o també els que són transformats o agreujats per aquesta». És a dir, per les persones que utilitzen els avenços de les tecnologies de la informació.

- María Bolaño ens diu: «És l'anàlisi de la naturalesa i l'impacte social de la tecnologia informàtica i la corresponent formulació i justificació de polítiques per a un ús ètic d'aquesta tecnologia». Aquesta definició està relacionada amb els problemes conceptuals i els buits en les regulacions que ha ocasionat la tecnologia de la informació.

- Altres autors (J. B. Peña i E. A. Fernández) formulen l'ètica informàtica com «la disciplina que identifica i analitza els impactes de les tecnologies de la

informació en els valors humans i socials». Aquests valors afectats són la salut, la riquesa, el treball, la llibertat, la democràcia, el coneixement, la privacitat, la seguretat o l'autorealització personal.

L'ètica informàtica es planteja diversos objectius:

- Descobrir i articular dilemes ètics clau en informàtica.

- Determinar en quina mesura són agreujats, transformats o creats per la tecnologia informàtica.

- Analitzar i proposar un marc conceptual adequat i formular principis d'actuació per determinar què fer en les noves activitats ocasionades per la informàtica en les quals no es perceben amb claredat línies d'actuació.

- Utilitzar la teoria ètica per aclarir els dilemes ètics i detectar errors en el raonament ètic.

- Proposar un marc conceptual adequat per entendre els dilemes ètics que origina la informàtica i, a més, establir una guia quan no existeix reglamentació de donar ús a Internet.

L'ètica informàtica (i, per extensió, la de les seves dades) ha d'estar, almenys, present en les àrees següents:

- La utilització de la informació.
- L'àmbit informàtic com a nova forma de bé o propietat.
- L'àmbit informàtic com a instrument d'actes potencialment nocius.
- Pors i amenaces de la informàtica.
- Dimensions socials de la informàtica.

Així, el professional que treballa amb dades sensibles (per exemple: sobre persones o grups) destinades a prendre decisions ha d'adoptar una forma de conducta que garanteixi:

- Responsabilitat.
- Confidencialitat.
- Qualitat del producte.
- Judici.
- Promoure un enfocament ètic en la gestió.
- Promoure el coneixement.
- Actualització permanent.

No és necessari establir regulacions sobre el que s'ha de fer amb les dades. L'objectiu ha de ser ajudar a **prendre decisions** efectives en l'àmbit dels negocis, a través de mètodes i tècniques que facilitin discussions internes, rigoroses i productives. Aquestes discussions poden expressar posicions coherents i consistents de la perspectiva d'una organització sobre l'ús de les seves dades.

Si ens traslladem a l'àmbit de les xarxes socials i les macrodades (*big data*) que es generen, se'ns ocorre fàcilment que captar aquestes dades i fer-ne mineria de dades (*data mining*) per vendre informació és el que dona valor a les xarxes socials. Aquestes dades, un cop processades i convertides en intel·ligència, són d'un valor monetari incalculable.

Això fa pensar que, un cop aconseguida i processada la informació, podria ser utilitzada per manipular i intentar modificar el comportament humà. Això porta, com a conseqüència lògica, a un problema ètic: qui i com controlarà l'ús de tota aquesta informació? Es fa necessari mantenir pràctiques ètiques, que no estan del tot definides.

Un últim aspecte que cal considerar, relatiu a l'explotació de dades, és que la mineria de dades té moltes aplicacions útils, però també un enfocament merament exploratori que fa discutible la validesa de certes deduccions. L'ús d'informació personal amb finalitats predictives té conseqüències directes sobre la vida de les persones i exigeix, per tant, actuar en un marc de responsabilitat. Aleshores, es fa necessari un codi d'ètica…

No podem acabar aquest bloc sense comentar que és molt habitual que les organitzacions desenvolupin instruments per protegir les seves dades, pensant en el seu ús fraudulent per part d'empleats, clients o proveïdors.

L'objectiu desitjat és garantir la disponibilitat, integritat i confidencialitat de les dades que gestionem, proporcionant els recursos i aplicant els controls necessaris per a aconseguir-ho.

Per a això, trobem: codis de conducta, normes d'ús d'eines electròniques, polítiques de seguretat de la informació, acords de confidencialitat, drets de propietat intel·lectual, etc.

Exemple de continguts comuns d'aquests instruments

- Parts afectades.
- Responsabilitats.
- Definició d'informació confidencial.
- Deure de confidencialitat.
- Protecció de dades: comptes d'usuari, contrasenyes, etc.
- Ús de la informàtica i les comunicacions.
- Control d'accés i privacitat.
- Propietat intel·lectual i industrial.
- Etc.

Resum

En aquest mòdul hem abordat la gestió de dades més enllà de la FIC i hem introduït diferents activitats de govern de la dada. Per a aquelles que estan relacionades més directament amb la FIC, com la gestió de la qualitat de la dada, hem entrat en més detall.

Així mateix, s'ha ressaltat el concepte de metadada com a aspecte crític en la gestió de dades de l'organització, sense oblidar el compliment de la normativa vigent i l'atenció a l'ètica en el tractament de les dades.

Abreviatures

API Sigles en anglès d'interfície de programació d'aplicacions: és el conjunt de subrutines, funcions i procediments que ofereix certa biblioteca per ser utilitzat per un altre programari com una capa d'abstracció.

BPEL Sigles en anglès de llenguatge d'execució de processos de negoci: llenguatge d'orquestració de processos de negoci basada en estàndards, compost per serveis.

CDI Sigles en anglès d'integració de dades de clients. Disciplina de gestió de dades mestres centrada en la integració i normalització de dades de clients.

DAMA Sigles de Data Management Agency: associació internacional dedicada a l'avenç i definició de millors pràctiques en l'entorn de la gestió de dades.

MDM Sigles en anglès de gestió de dades mestres: disciplina que ofereix una visió única de les dades buscant la seva estandardització i fiabilitat.

Bibliografia

Berson, A.; Dubov, L. (2010). *Master Data Management and Data Governance*. McGraw-Hill/ Osborne Media.

English, L. P. (1999). *Improving Data Warehouse and Business Information Quality*. Nova York: John Wiley & Sons, Inc.

Fan, W.; Geerts, F. (2012). *Foundations of Data Quality Management*. Morgan & Claypool Publishers.