Qualitat de les dades

PID_00251622

David Cabanillas Barbacil

Temps mínim de dedicació recomanat: 4 hores

Universitat Oberta de Catalunya

Índex

Introducció

El potencial valor de la dada no ha parat de créixer en els últims anys, com a resultat de la transformació digital progressiva. Actualment, les organitzacions tenen a la seva disposició múltiples estratègies per assolir aquest valor, com per exemple *big data** (dades massives), *business analytics* (analítica empresarial) o *business intelligence* (intel·ligència empresarial). No obstant això, tal com apunta McKinsey, les organitzacions, en la gran majoria dels sectors, encara no han aconseguit capturar el valor de la dada.

Lectura complementària

N. Henke; J. Bughin; M. Chui; J. Manyika; T. Saleh; B. Wiseman; G. Sethupathy (2016). *The age of analytics: Competing in a data-driven world*. McKinsey Global.

Tal com apunten Ransbotham i Kiron, el govern de la dada és clau per desbloquejar l'oportunitat que proporcionen les dades i els algorismes.

*** Més informació a: https://goo.gl/7nHbyp**

Els motius darrere de la impossibilitat de capturar el valor de les dades són múltiples i diversos, i inclouen aspectes com la falta de talent, l'existència de diverses àrees d'informació o fins i tot l'escepticisme per part de la direcció.

Lectura complementària

S. Ransbotham; D. Kiron (2017). *Analytics as a Source of Business Innovation*. MITSloan.

Dins de les àrees d'aplicació del govern de la dada (encara que pot trobar-se de forma independent), destaca la **qualitat de la dada**. Sense certs nivells en la qualitat de les dades, la presa de decisions i l'eficiència dels processos es poden veure afectades per una pèrdua d'integritat, per falta de completesa o fins i tot per l'aparició d'inconsistència. És el que, en definitiva, moltes vegades anomenem *garbage in, garbage out*.

En aquest mòdul estudiarem la necessitat i la importància de la qualitat de la dada, en què consisteix, què aporta, com implementar-la, què hem de tenir en compte com a millors pràctiques i quines tecnologies suporten la qualitat de la dada.

Objectius

Aquest material didàctic està adreçat a:

1) Desenvolupadors i consultors que volen saber què significa qualitat de la dada o *data quality*.

2) Desenvolupadors i consultors que volen ajudar en el camp del desenvolupament d'estratègies de negoci que incloguin qualitat de dades.

3) Gestors que estan interessats en la transformació digital de la seva organització i en la inclusió de la qualitat de la dada com un dels seus pilars fonamentals.

En els materials didàctics d'aquest mòdul, trobarem les eines indispensables per assimilar els objectius següents:

1) Entendre el concepte de *data quality*, les situacions en què és necessari desplegar una solució d'aquest tipus i els avantatges que proporciona.

2) Saber en què consisteix un programa de qualitat de dades.

3) Enumerar i donar a conèixer millors pràctiques de qualitat de les dades.

4) Conèixer tècniques i tecnologies per a la gestió de la qualitat de les dades.

Si bé l'obra és autocontinguda en la mesura del possible, els coneixements previs necessaris són:

1) Coneixements bàsics sobre *business intelligence* i *big data*.

2) Coneixements sobre estratègia i gestió de les tecnologies de la informació (TI).

S'introduiran els conceptes necessaris per al seguiment d'aquest material.

1. Qualitat de les dades

1.1. La necessitat de la qualitat de les dades

Perquè la dada pugui considerar-se com un actiu de valor i l'organització pugui prendre millors decisions, millorar processos, reduir costos i crear noves fonts d'ingressos, cal que, al llarg del seu cicle de vida, la dada no tingui problemes que afectin la seva qualitat.

Aquest tipus de problemes comporten greus costos associats, derivats de prendre decisions errònies, incrementar els costos operatius, generar insatisfacció entre els clients, deteriorar la imatge corporativa i una pèrdua de confiança entre els clients, els empleats i els proveïdors. Sense la confiança de les persones que prenen decisions, es pot retardar o fins i tot aturar l'explotació de les dades, fet que suposa una barrera per convertir-se en una empresa orientada a la dada, o *data driven*.

Confiança

La confiança és la seguretat o l'esperança ferma que algú té d'un altre individu o d'alguna cosa. També es tracta de la presumpció d'un mateix i de l'ànim o el vigor per obrar. En el nostre cas, és la confiança sobre les dades que tenen les persones en l'organització.

Data driven

Data driven fa referència a un procés o a una activitat que és guiada per les dades, en lloc de ser impulsada per la simple intuïció o experiència personal. En el nostre context, les decisions es prenen sobre les dades i no sobre especulacions o sensacions.

Els problemes de qualitat de dades són sistèmics en les organitzacions i, de fet, diferents estudis porten temps mostrant la magnitud del problema a què ens enfrontem:

- El 2002, la baixa qualitat en les dades dels seus clients va significar pèrdues de 611 bilions de dòlars a les companyies dels Estats Units.*
- El 2004, es va estimar que la qualitat de la dada comportava pèrdues d'almenys el 10%; possiblement estava més a prop del 20%.**
- El 2011, la baixa qualitat de la dada es considerava la principal raó per la qual el 40% de les iniciatives de negoci fracassaven en aconseguir els objectius definits.***
- El 2014, el 59% de les empreses citaven la qualitat de les dades com una barrera per a l'adopció de *business intelligence*.****
- El 2016, les organitzacions van perdre de mitjana 9,7 milions de dòlars anualment, a causa de la mala qualitat de dades.*****

* W. Eckerson (2002). *Data Quality and the Bottom line.* TDWI.

** T. Redman C. (2004). *Data: An Unfolding Quality Disaster. DM Review Magazine.*

*** T. Friedman; M. Smith (2011). *Measuring the Business Value of Data Quality.* Gartner.

**** Diversos autors (2011). *2014 Analytics, BI, and Information Management Survey.* Information Week.

***** A. D. Duncan ; M. Y. Selvage ; S. Judah (2016). *How a Chief Data Officer Should Drive a Data Quality Program.* Gartner.

L'impacte de la qualitat de dades va molt més enllà de la presa de decisions, i és possible trobar-lo al llarg de tots els processos d'una organització, per exemple:

- Dades de clients errònies incideixen en l'efectivitat de les campanyes de màrqueting.
- Adreces incorrectes produeixen enviaments fallits de productes i un increment en els costos operatius.
- Els mesuraments incorrectes de productes poden conduir a problemes significatius de fabricació o transport; per exemple, que el producte no encaixi en el camió que l'ha de transportar.
- Les males dades a la cadena de subministrament de queviures costen a la indústria australiana més de mil milions de dòlars australians. IBM, juntament amb GS1 a Austràlia,* va comparar les dades dels productes de supermercat de tres grans superfícies amb les dades dels quatre principals proveïdors. Es va revelar que els supermercats estaven treballant amb dades amb inconsistències d'entorn del 80%: des d'errors o falta de dimensions, fins al nombre de palets que es poden carregar per producte o les condicions d'emmagatzematge.

  *** Més informació a: https://goo.gl/nhqCVE**
- La recollida errònia de dades de pressió i temperatura en la producció de peces de plàstic pot fer que surtin al mercat vehicles que tinguin problemes d'incendi per una mala fabricació dels seus components.

Per contra, una alta qualitat de les dades pot millorar l'avantatge competitiu i la capacitat de les organitzacions. Entre els impactes positius, podem destacar:

- Més efectivitat en l'adquisició i la retenció de clients.
- Optimització en totes les àrees de l'organització.
- Execució de processos eficients en la cadena de subministrament i de producció.
- Eliminació de costosos errors operatius.
- Penetració ràpida en nous mercats.
- Presa de decisions de negocis intel·ligent i oportuna.

Cal comentar que, en general, les organitzacions no inicien projectes de govern de la dada, sinó que identifiquen problemàtiques específiques i, al final, les combinen sota el mateix programa. De fet, freqüentment molts dels pro-

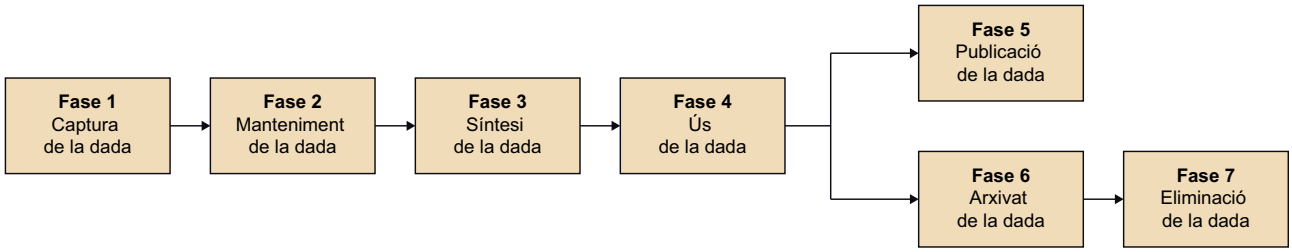
jectes de govern de la dada tenen els seus orígens en el seu intent de regirar problemàtiques associades a la qualitat de dades.

1.2. Motius de la baixa qualitat de les dades

La gestió de les dades amb què compta una organització ha anat guanyant pes en els últims anys, i avui dia s'ha convertit en una operació clau per assegurar el futur de qualsevol organització. L'èxit o el fracàs en la gestió de la qualitat de dades està íntimament vinculat amb els riscos que s'assumeixen en aquest tipus d'operacions de gestió, i especialment amb el pla i les mesures adoptats per afrontar-los; però, i malgrat l'evident importància que les organitzacions es dotin d'un correcte pla de gestió de qualitat de dades, en moltes ocasions o no s'aplica cap pla en la gestió, o només se centren en l'entrada de dades.

És important recordar que la dada té un cicle de vida, com s'il·lustra a la figura 1, i que al llarg d'aquest cicle podem trobar exemples de baixa qualitat, com es recull en la taula 1.

Figura 1. Fases del *data governance*



Font: Marcos Pérez Rodríguez

Taula 1. Fases del cicle de vida de la dada i exemple de baixa qualitat de dades

Fase	Exemple de baixa qualitat de dada
1...7	Qualsevol intervenció manual en el procés de flux de dades. Sistemes d'informació no sincronitzats que han de compartir informació.
2...3	Aplicar una fórmula o un algorisme sobre dades que és incorrecte.
5	Enviament d'informes desfasats.
6	Metadades que descriuen les dades de manera incorrecta.

Metadada

La metadada fa referència a dades sobre les dades. Per exemple: temps i creació de la dada, creador/font de la dada...

La captura de la dada, **fase 1**, és el punt d'entrada de les dades i el punt més freqüent d'errors de la baixa qualitat de les dades. Es tracta de problemes a les correccions ortogràfiques, transposicions de nombres, codis incorrectes o parts de la dada que no han estat incloses, dades que han estat col·locades en els camps incorrectes i noms, sobrenoms, abreviatures, acrònims irreconeixibles, tipus de dades incoherents o el fet que la captura de les dades es faci de manera incorrecta (per exemple, sensors no calibrats correctament). Aquests tipus d'errors estan augmentant a mesura que les empreses traslladen els seus negocis a l'entorn digital i a la quarta revolució, i es permet als clients/proveïdors i tercers d'introduir o fer servir dades sobre ells directament en els mateixos sistemes. Es poden evitar molts errors d'entrada de dades mitjançant l'ús de rutines de validació que comproven les dades a mesura que s'introdueixen

en els sistemes comprovant sintaxi, formats, dades estranyes o estructures no coincidents.

A la **fase 2** entren en joc les eines ETL (*extract, transform, load*), que permeten extreure, transformar i carregar la dada. Pensem en un petit exemple d'ETL: suposem que tenim dades procedents de dos sistemes diferents, que els desitgem emmagatzemar en la mateixa destinació, però podria haver-hi algunes diferències entre tots dos. Per exemple, un pot denotar el sexe com M i F, l'altre el designa com 0 i 1. Ara, si es desitja emmagatzemar en un únic fitxer i el sexe es vol emmagatzemar com M i F, s'ha de transformar el 0 i 1 a M i F. S'ha de fer un procés ETL de traduir els $0 \Rightarrow$ M i els $1 \Rightarrow$ F.

ETL

ETL és l'acrònim d'*extract, transform and load*, i fa referència als processos que permeten extreure, transformar i carregar dades per habilitar el seu consum eficient.

A la **fase 3**, trobem les regles i els algoritmes aplicats a la dada. Per exemple, es pot crear una regla que indiqui si donar crèdit a un client o no. En aquest cas, la regla/procés, a diferència de la fase 2, és part de la lògica de negoci de l'organització. Una regla podria ser un camp booleà que indica si donar o no el crèdit seguint la regla: si el salari $\geq$1.000 i les despeses $\leq$200, donem el crèdit, si no, no el donem. En tots dos casos, si la regla o el procés fossin incorrectes, s'introduirien errors en les dades.

A la **fase 5**, el temps en què s'utilitza la dada pot ser un factor d'error. Per exemple, si estem enviant dades de vendes correctes, però vam indicar que són d'un període i en realitat són d'un altre.

Finalment, a la **fase 6** un error pot venir si les metadades que descriuen les dades emmagatzemades són obsoletes o incorrectes.

Com és possible d'imaginar, els errors d'entrada de dades s'agreugen quan les organitzacions intenten integrar dades de múltiples sistemes o porten a terme canvis estructurals en els sistemes d'origen. De vegades aquests canvis són deliberats, com quan un administrador afegeix un nou camp o un valor de codi i no es notifica a la resta de l'organització. En altres casos, es generen noves dades a partir de dades existents (per exemple, el sexe es pot mirar d'obtenir a partir del nom de l'usuari). A causa de la complexitat dels sistemes actuals, els canvis en els sistemes font es propaguen de manera fàcil i ràpida a altres sistemes per mitjà del cicle de vida.

1.3. Motius pels quals no s'inverteix en la qualitat de les dades

Per tot el que hem comentat fins al moment, hauria de ser prioritari per a una organització invertir en la qualitat de la dada i, no obstant això, és una cosa que no forma part de les seves prioritats. Com apunten els últims informes de la Society for Information Management (SIM), anomenats IT Trends Study, les principals prioritats del departament de TI són la seguretat dels actius digitals, la manca de talent, l'alineació amb el negoci, la credibilitat i la continuïtat de negoci, de manera que aquesta tendència està canviant.

Les principals raons que justifiquen no invertir en la qualitat de dades són:

1) Es considera que les dades són correctes: es tracta del motiu més estès. No obstant això, la majoria dels estudis defensen el contrari. Això se sol argumentar per dues raons principals:

a) Por d'allò desconegut: dins de l'empresa, no se sap molt bé com afrontar una iniciativa de qualitat de dades, per desconeixement de les tècniques existents, de què suposa aquesta iniciativa, el que implica, etc.

b) Por del que es pot trobar: tradicionalment, la qualitat de les dades s'havia assignat a les àrees de TI, un departament que no necessàriament hauria de conèixer el significat de les dades. Per això, és important que la responsabilitat d'una dada romangui en l'àrea de negoci que la gestiona.

2) Descobriment d'errors: amb més qualitat de les dades, es produeix un increment de la comprensió del rendiment de l'organització. Per exemple, es podria descobrir que es tenen menys clients dels que es creia o que no s'ha estat capaç de detectar fraus (entre d'altres).

3) No es veu un valor directe en la inversió: la majoria dels estudis defensen que disposar de dades de qualitat evita que hi hagi retards en els projectes i que s'entri en sobrecostos. Així doncs, la inversió en aquest tipus de projectes genera un gran retorn en el futur.

4) No se'n percep la necessitat: es tracta d'un motiu poc real perquè qui no necessita comprovar que els seus informes són correctes? Qui no vol saber si les seves decisions són consistents? O qui no necessita detectar un frau? Hi ha molts motius per apostar per implementar un projecte de qualitat de dades.

5) Excessivament costós: encara que pugui semblar-ho, realment és una percepció errònia. Actualment, hi ha moltes eines assequibles o fins i tot gratuïtes que permeten implementar solucions, totals o parcials, de qualitat de dades en temps curts i amb una gran efectivitat.

1.4. Què és la qualitat de dades?

Hem estat parlant de la necessitat i de les barreres de la qualitat de la dada, però no n'hem introduït encara una definició formal. És el moment de fer-ho, segons la norma ISO 9000: 2015.*

*** Més informació a: https://goo.gl/kEwyR6**

ISO 9000

ISO 9000 és un conjunt de normes sobre qualitat i gestió de qualitat, establertes per l'Organització Internacional de Normalització (ISO), i especifica la manera en què una organització opera els seus estàndards de qualitat, temps de lliurament i nivells de servei.

S'entén per **qualitat de dades** el grau en què les dades compleixen un conjunt de característiques o dimensions.

Per comprendre aquesta definició, cal entrar en el detall de les característiques o les dimensions que han de complir les dades. No existeix un consens en

la indústria sobre quines són aquestes dimensions. Per exemple, segons l'OECD (Organization for Economic Co-operation and Development), són **rellevància**, **exactitud**, **credibilitat**, **oportunitat**, **accessibilitat**, **interpretabilitat** i **coherència**. Segons EUROSTAT (Statistical Office of the European Communities), cal afegir: **puntualitat**, **transparència**, **comparabilitat** i **exhaustivitat** (que en el cas de l'OECD, estan incloses en les seves dimensions).

En el present material, i amb l'objectiu de ser el més imparcials que resulti possible, considerarem l'enfocament de DAMA.* Per a aquesta organització, les dimensions de la qualitat de les dades són:

*** Més informació a: *The six primary dimensions for data quality assessment.* (2013). Dama UK Chapter.**

DAMA

DAMA és una associació sense ànim de lucre. Té com a objectiu ajudar els professionals de les dades mitjançant la creació de taxonomies i documents de referència i la seva divulgació.

- **Completesa**, que consisteix en la proporció de dades emmagatzemades respecte al conjunt total.
- **Unicitat**, que consisteix que la dada ha de guardar-se de forma única per evitar inconsistències.
- **Atemporalitat**, que consisteix en el grau en què la dada representa la realitat en un moment temporal específic.
- **Validesa**, que consisteix que la dada presenta conformitat (format, tipus, rang) respecte a la seva definició.
- **Precisió/exactitud**, que consisteix en el grau en què la dada descriu la realitat (amb independència del temps).
- **Consistència**, que consisteix en l'absència de diferències en comparar dues representacions de la mateixa dada, evitant informació contradictòria.

Altres dimensions que es poden incloure són usabilitat, duplicació, disponibilitat, confiança o valor.

A la pràctica, la qualitat de les dades és una preocupació per part dels professionals que participen en els sistemes d'informació, que van des de l'emmagatzematge de dades i la intel·ligència empresarial a la gestió de la relació amb els clients i la gestió de la cadena de subministrament. És a dir, qualsevol dada que estigui d'alguna manera relacionada amb l'empresa o l'organització i la qualitat de la dada. La preocupació de les empreses ha fet incorporar la qualitat de les dades com a part fonamental del *data governance*.

Per atacar el problema de la qualitat de les dades, les organitzacions necessiten invertir en les persones, els processos i les tecnologies necessàries per transformar dades defectuoses en informació fiable i processable, disponible per a totes les parts en qualsevol moment que la necessitin. Les millors iniciatives de qualitat de dades tenen aquestes quatre característiques llistades a la taula 2.

Taula 2. Característiques i les seves descripcions

Característica	Descripció
Col·laboratiu	Negoci i TI comparteixen la responsabilitat de la qualitat de les dades, amb funcions i tecnologia clarament definides i adaptades a les habilitats i perspectives úniques dels analistes de negoci, administradors de dades i desenvolupadors i administradors de TI.
Proactiu	Negoci i TI reconeixen que totes les organitzacions pateixen algun grau de mala qualitat de les dades i treballen conjuntament per identificar i corregir els problemes abans que afectin el rendiment del negoci.
Reutilitzable	El perfil de dades i les regles de negoci de neteja poden reutilitzar-se en qualsevol nombre d'aplicacions, per agilitzar i accelerar els processos i ajudar a garantir alts estàndards de qualitat.
Pervasiu	L'entorn de qualitat de dades s'estendrà a totes les parts interessades, dominis de dades, projectes i aplicacions, independentment d'on resideixin les dades.

Perquè la qualitat de les dades sigui més efectiva, ha de ser impulsada per una metodologia que incorpori les característiques definides anteriorment. Idealment, la metodologia serà supervisada i implementada per un òrgan del govern de la dada.

1.5. Com resoldre els reptes?

Per gestionar la qualitat de les dades en una organització, s'han de definir una sèrie de tasques i rols ja que, com hem comentat, la resolució de problemes a l'entrada de dades no és suficient:

- **Els rols tècnics i no tècnics o de negocis han de comunicar-se**. La manca de col·laboració entre aquestes dues parts d'una organització és una de les principals raons per les quals molts projectes de qualitat de dades no compleixen les seves expectatives inicials. Tradicionalment, negoci i TI s'han basat en fulls de càlcul, documents, correus electrònics i altres mecanismes tediosos i imprecisos per comunicar els requisits de qualitat de les dades. Inevitablement, sota aquestes condicions és difícil per als analistes de negoci i els administradors de dades, és a dir, les persones que han de seguir el pla de qualitat de dades, esbossar requisits de negoci de qualitat de dades en termes clars perquè la part de la TI les pugui entendre i executar. La mala interpretació, els retards, els alts costos i els resultats no òptims són comuns simplement perquè negoci i TI estan parlant dos idiomes diferents, sense un marc comú. Els detalls crítics es poden perdre en la traducció. Per tant, la col·laboració entre negoci i TI és essencial per a la qualitat de les dades i les iniciatives de gestió de dades relacionades.

- **Dotar-se d'un sistema de supervisió contínua de les dades recentment incorporades** (a més dels ja existents). Un sistema que ha d'integrar eines de detecció que permetin filtrar i categoritzar les dades segons la seva qualitat, i detectar —abans que siguin requerides pel sistema de gestió (que les transformarà primer en informació rellevant i, després, en coneixement sensible per a la presa de decisions)— el seu grau de coherència, oportunitat i fiabilitat, entre altres factors determinants. Les empreses que estan admi-

nistrant la qualitat de les dades estan operant normalment en un entorn per lots o *batch*; és a dir, executen comprovacions de dades, creen conjunts de dades revisades periòdicament i les introdueixen en el sistema de dades. Si bé el processament diari és certament útil, en alguns entorns comprovar la qualitat de les dades tan bon punt es crea la dada, per exemple, quan s'introdueix una nova transacció o al moment en què un nou conjunt de dades estigui disponible, és obligatori o, com a mínim, avantatjós.

En resum, per arribar a obtenir una qualitat de les dades efectiva, hem de ser capaços de dur a terme les activitats següents:

- **Objectius**: determinació de les millors oportunitats per optimitzar les dades.
- **Conceptualització**: definició de dimensions en l'organització.
- **Avaluació**: avaluació dels nivells de qualitat actual de les dades.
- **Eliminació**: eliminació de les fonts de problemes.
- **Automatització**: automatització del control de la qualitat de dades.
- **Seguiment**: seguiment de la gestió de processos i les seves dades.
- **Mesura**: mesurament de la qualitat de dades.

I la part més important, unir totes aquestes tasques en un marc o en un programa entre els tres agents que han d'intervenir per assegurar una bona qualitat de les dades: **persones**, **processos** i la **tecnologia**.

2. Programa de qualitat de dades

2.1. Introducció

Per gestionar la qualitat de dades en una organització, hem de desenvolupar el que es coneix com a programa de qualitat de dades. Per tant, hem d'introduir aquest concepte.

S'entén per **programa de qualitat de dades** la metodologia estratègica i sistemàtica per a l'avaluació de la qualitat de les dades dins d'una organització, la qual permet identificar atributs de qualitat de dades, analitzar aquests atributs en el seu context actual o futur i proporcionar una guia per millorar la qualitat de les dades.

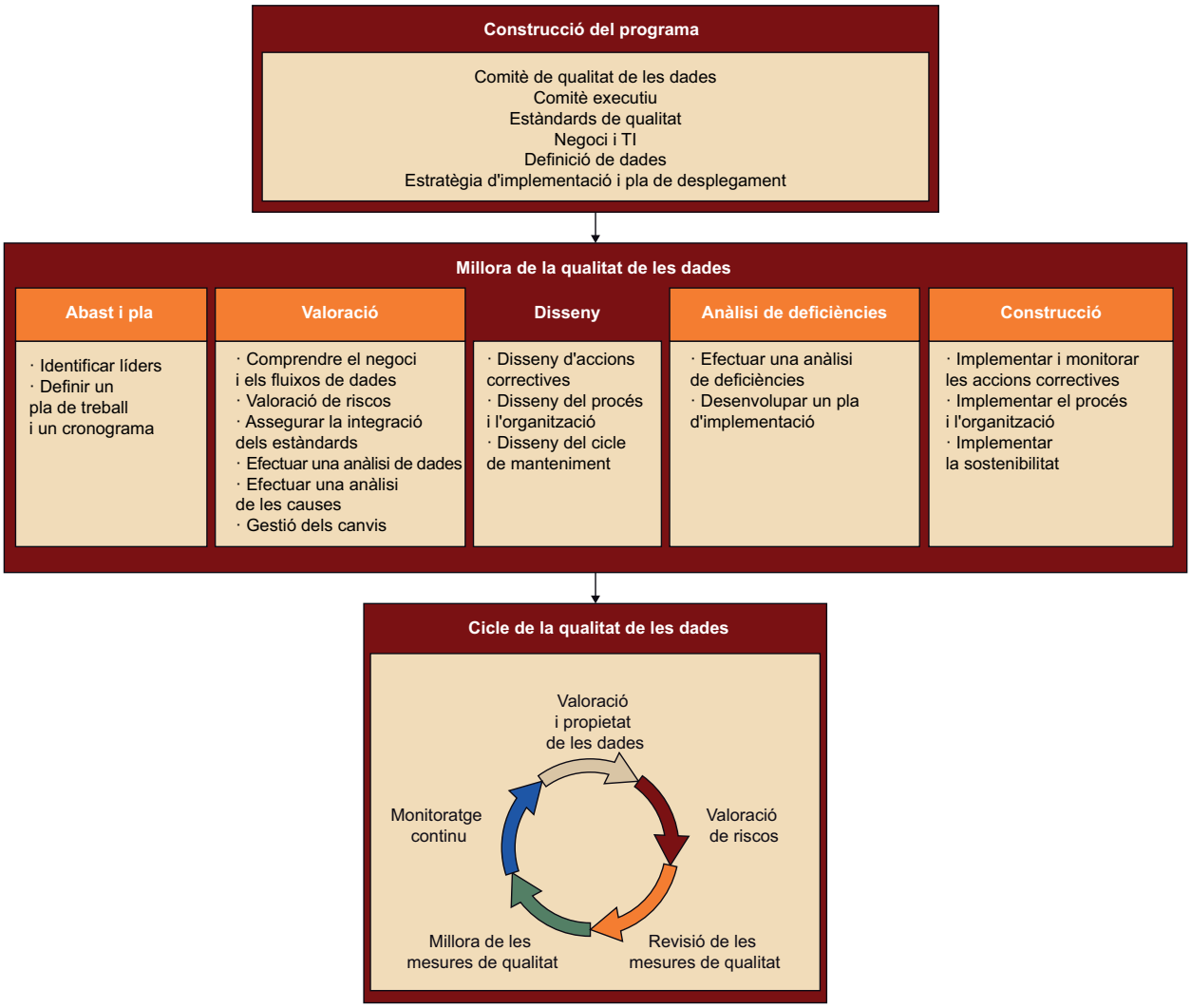
El programa permet orquestrar persones, processos i tecnologia, amb l'objectiu d'aconseguir el màxim retorn d'una iniciativa de qualitat de dades.

Encara que el valor d'una metodologia de qualitat de dades pot semblar evident, massa organitzacions aborden iniciatives de qualitat de dades sense plans o amb plans mal definits que introdueixen riscos, passen detalls per alt i porten a terme esforços redundants. Per contra, una metodologia estratègica i sistemàtica permet avaluar adequadament el seu projecte de qualitat de dades, involucrar les parts interessades de negoci i TI, definint funcions i responsabilitats, i dotar dels processos i de les eines adequades per abordar el repte de la qualitat de les dades. Com en tota tasca que cal abordar, si des del primer moment estan ben definides les metes, els rols i les activitats que cal fer, les probabilitats d'èxit augmenten.

Instituir un programa de qualitat de les dades dins d'una organització va molt més enllà de l'adquisició d'eines de neteja de dades, la creació d'un consell d'administració de dades o el fet de documentar correctament una sèrie de processos. El programa de qualitat de dades consisteix en un **cicle iteratiu d'avaluació, planificació, execució, gestió** i **revisió**.

Això vol dir que en el programa es determina la manera d'actuar, es defineixen els objectius, es designa els responsables de cada iniciativa i les accions que cal prendre. Això requereix processos reproduïbles, apalancats en els instruments adequats i en persones amb una formació i unes habilitats adequades, com s'il·lustra a la figura 2.

Figura 2. Procés complet del programa de qualitat de dades



Font: David Cabanillas

El programa de qualitat de dades consisteix en un llistat d'elements:

- **Maduresa**: permet entendre en quin estat es troba la nostra organització respecte a la qualitat.

- **Expectatives**: permeten delimitar què esperem de l'aplicació d'un programa de qualitat de dades. És a dir, es tracta de preveure/quantificar beneficis.

- **Mesura**: permet mesurar l'estat de l'organització abans i després de l'aplicació del programa de qualitat de dades.

- **Polítiques**: defineixen la guia que cal seguir sobre la qualitat de les dades.

- **Processos**: identifiquen i delimiten processos de l'organització que tenen contacte amb les dades.

- **Govern**: defineix els rols dins de l'organització pel que fa a la qualitat de les dades.

- **Estàndards**: estableixen les normes sobre les quals s'ha de regir la qualitat de les dades.

- **Metodologies**: es tracta de les eines i tècniques utilitzades per aconseguir una qualitat de dades adequada.

A continuació, revisarem amb detall cadascun d'aquests elements.

2.2. Maduresa respecte a la qualitat de la dada

Per poder adoptar un programa de qualitat de dades, cal comprendre **en quina situació es troba la nostra organització**. Aquest assessorament es duu a terme per mitjà d'un model de maduresa per a la qualitat de la dada. El model de maduresa de qualitat de dades utilitzat en aquest material es basa en el model de maduresa de capacitats (CMM, que és l'acrònim de *capability maturity model*), desenvolupat per l'Institut d'Enginyeria de Programari de la Universitat Carnegie Mellon.

Lectura complementària

D. Loshin (2010). *The Practitioner 's Guide to Data Quality Improvement*. Morgan Kaufmann.

Per adoptar un enfocament de gestió del rendiment per a la qualitat de les dades, cal anar més enllà de solucions *ad hoc*, és útil visualitzar com la gestió de la qualitat de les dades encaixa amb totes les activitats dependents de la informació dins de l'organització. Les diferències en la maduresa de les organitzacions es mesuren pels processos i les persones encarregades de la identificació de dades defectuoses.

Solució *ad hoc*

Una solució *ad hoc* és aquella que està específicament elaborada per un problema o un fi precisos i, per tant, no és generalitzable ni utilitzable per a altres propòsits.

L'estat de maduresa de les organitzacions es pot dividir en cinc estats (que van des d'inicial, en el qual no hi ha qualitat de la dada i s'apliquen solucions puntuals, fins a optimitzat, en què hi ha un programa integrat en tots els processos, com es mostra a la taula 3).

Tot i que els models de maduresa fonamentats en CMM són els més estesos en l'àmbit de la qualitat de la dada, també és possible trobar altres models fonamentats en quatre etapes, com il·lustra la taula 4.

Bones pràctiques

Es refereix a tota experiència que es guia per principis, objectius i procediments apropiats que s'adeqüen a una determinada perspectiva normativa o a tota experiència que ha donat resultats positius.

2.3. Expectatives respecte a la millora de les dades

Abans d'aplicar un programa, és útil especificar per endavant les seves expectatives pel que fa a la qualitat de les dades i els mètodes que s'utilitzaran per avaluar-lo. Per exemple, es considerarà acceptable una taxa d'error en la variable x de menys de l'1%?

Taula 3. Estat de maduresa de les organitzacions

Estat	Solucions	Col·laboració
Inicial	Les solucions per als problemes són *ad hoc*.	Les solucions no són compartides, cosa que impedeix la replicació de la solució.
Repetible	S'identifiquen i avaluen les fonts de la baixa qualitat de la dada.	Es comparteixen les bones pràctiques entre els membres de l'organització.
Definit	Hi ha un entorn per monitorar la qualitat de la dada.	L'equip de qualitat de dades en documenta problemes i solucions.
Gestionat	S'avalua i mesura l'impacte de la qualitat de dades.	La informació és compartida i es generen informes d'impactes sobre possibles problemes.
Optimitzat	Les millores estratègiques i la supervisió contínua del procés del cicle de vida de les dades mitjançant panells s'apliquen en tota l'organització.	L'entorn de qualitat de la dada inclou oportunitats de millora.

Taula 4. Estat de maduresa de les organitzacions

Estat	Importància de la qualitat de la dada per a l'organització	Enfocament
Desconegut	Limitada.	Es prenen solucions quan la informació sol ser inferior a l'estàndard.
Reactiu	Es comença a reaccionar davant els problemes de qualitat de les dades, ja que impacten en el rendiment del negoci.	Inversió en resposta a un esdeveniment que ha causat problemes.
Proactiu	Es comença a definir funcions i a crear figures de gestió.	Es comença a comprendre el valor dels actius de dades més clarament, i a tenir un procés més estructurat per a la seva anàlisi.
Optimitzat	*Business as usual*, les decisions es prenen sobre les dades.	Vincle entre la qualitat de les dades i el rendiment financer.

En un àmbit de negoci, és possible fer-se preguntes com les següents:

- Com ha disminuït el rendiment a causa dels errors?
- Quin percentatge de temps es gasta en la reelaboració de processos fallits?
- Quina és la pèrdua de valor de les transaccions que van fallar a causa de la manca de dades?
- Amb quina rapidesa podem respondre a les oportunitats emergents si disposem de dades de qualitat?

No obstant això, en diferents fases de maduresa, es tindran diferents expectatives. La taula 5 relaciona les expectatives amb la maduresa de les organitzacions.

És a dir, les expectatives a l'hora de millorar la qualitat de les dades no només han de ser d'un àmbit tècnic; el valor afegit de la millora de la qualitat de les dades ha d'estar vinculat a la satisfacció de les expectatives de negoci. Això implica identificar els impactes empresarials, els seus problemes i causes, i després quantificar els costos per eliminar tots els problemes relacionats amb les dades i els seus beneficis associats.

Taula 5. Expectatives per a cada estat de maduresa de les organitzacions

Estat	Caracterització
Inicial	L'activitat de qualitat de les dades és reactiva. No hi ha capacitat per identificar ni documentar les expectatives de qualitat de les dades.
Repetible	Anticipació limitada de certs problemes de dades, i s'identifiquen i notifiquen errors senzills.
Definit	Les dimensions de la qualitat de les dades s'identifiquen i es documenten. Existeix la capacitat de validar les dades utilitzant regles de qualitat de dades definides, i també mètodes per avaluar l'impacte en la part de negoci.
Gestionat	La validesa de les dades s'inspecciona i se supervisa. L'anàlisi de l'impacte a la part de negoci es fa de manera global. I els resultats de l'anàlisi s'han tingut en compte en la priorització de les expectatives.
Optimitzat	Punts de referència de qualitat de dades definits. Les expectatives de qualitat de les dades estan vinculades a objectius de negoci. És possible anticipar els nivells de qualitat i fixar metes de millora. Finalment, s'afegeixen controls per validació de dades integrades en els processos de negoci.

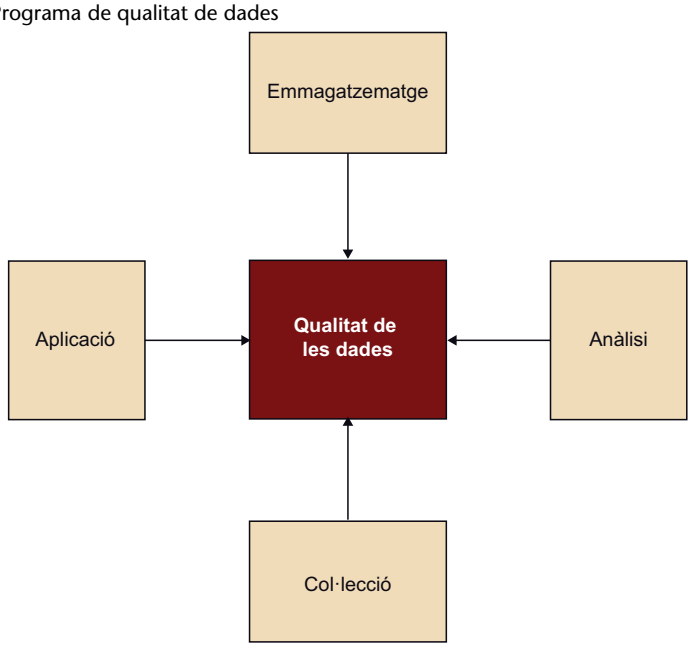
2.4. Mesura

Trobem quatre grans blocs que relacionen les dades amb l'organització, tal com s'il·lustra a la figura 3:*

*** Més informació a: https://goo.gl/q6RQ8u**

- **Aplicació**: el propòsit per al qual es recullen les dades.
- **Col·lecció**: els processos que acumulen les dades.
- **Emmagatzematge**: processos i sistemes utilitzats per arxivar les dades.
- **Anàlisi**: el procés de comprendre les dades per respondre a l'aplicació.

Figura 3. Programa de qualitat de dades



Font: David Cabanillas

Com és possible d'imaginar, en cadascun d'aquests blocs la qualitat es pot veure compromesa i cal controlar-la i mesurar-la.

Què és important mesurar? Hem de tenir diferents dimensions per a la qualitat de la dada:*

*** Més informació a: https://goo.gl/Y22vhD**

- **Coherència**: es defineixen i s'entenen els elements de les dades de forma coherent?
- **Integritat**: l'estructura de les dades i les relacions entre entitats i atributs es manté de manera consistent?
- **Completes**: es presenten totes les dades necessàries?
- **Oportunes**: es disposa de les dades quan és necessari?
- **Accessibles**: les dades són fàcilment accessibles, comprensibles i utilitzables?
- **Validesa**: els valors de les dades estan dins dels rangs acceptables definits pel negoci?
- **Precisió:** les dades representen amb exactitud la realitat o una font verificable?

Els cinc primers atributs pertanyen al contingut i a l'estructura de les dades, i cobreixen una multitud d'aspectes que comunament s'associen amb dades de mala qualitat: errors d'entrada de dades, regles empresarials errònies, registres duplicats i valors de dades que falten o són incorrectes. No obstant això, les dades sense defectes no tenen valor si no és possible entendre-les o accedir-hi. Per aquest motiu, les dues últimes dimensions s'avaluen millor mitjançant entrevistes i enquestes als usuaris de les dades, o per mitjà de mètodes estadístics.

A més de les mesures quantitatives, també s'han de considerar les mesures qualitatives. Alguns exemples inclouen:

- **Mesures de satisfacció del negoci**: mesuren l'augment/disminució en la satisfacció del negoci amb la millora en les dades.
- **Mesures de productivitat**: consisteixen en el percentatge de vegades que el consell de governança de dades va detectar i va eliminar projectes redundants intra o interdepartamentals.
- **Oportunitat de negoci / mesures de risc**: mesuren el benefici i l'augment de la competitivitat vinculats a la qualitat de la dada.

- **Mesures de compliment**: permeten entendre el comportament dels usuaris respecte als seus nivells d'accés i l'actualització de les dades.

És molt important establir les mesures de qualitat de dades més importants per a l'organització. Les mètriques poden ser generals, com les que hem discutit, o bé vinculades a un dels àmbits del govern de la dada. Això és necessari per establir una línia base per a la qualitat de les seves dades i per monitorar el progrés de les iniciatives pel que fa a la qualitat de dades.

Lectura complementària

R. K. Pandey (2014). *Data Quality in Data warehouse: problems and solution*. PhD Scholar, Universitat de Surguja (Chhattisgarh), Índia.

2.5. Polítiques o regles sobre les dades

Quins principis o regles s'han d'incloure en la seva política de qualitat de dades? El punt de partida és saber què és una política (en el context d'una organització).

S'entén per **política** un principi o una regla per guiar les decisions i aconseguir resultats.

Cal comentar que el terme no s'utilitza normalment per denotar el que realment es fa. Això es coneix normalment com a procediment o protocol.

La resposta a la pregunta anterior és oberta, tot i que tradicionalment inclou respondre a les preguntes següents:

- **Propòsit**: per què necessitem la política de qualitat de dades? Els seus propòsits han de ser clars i directes.

- **Antecedents**: per què ara és important el programa de qualitat de dades? Com s'alinea amb altres polítiques i objectius estratègics en l'organització? Aquesta secció és important per proporcionar un context a la política i deixar clar qui es beneficiarà i per què, a més de ser útil per explicar la història de la qualitat de les dades en l'organització.

- **Abast**: en quines circumstàncies s'ha d'habilitar la política? Per exemple, s'aplica la política a totes les dades de l'organització? Què passa amb les dades de tercers?

- **Rols i responsabilitats**: quins grups (o rols individuals) seran responsables d'assegurar que s'executi la política? La política serà governada de manera centralitzada o federada entre TI i negoci? Què s'espera de cada rol?

- **Declaració**: com es tractaran situacions i conflictes específics? Aquí és on les mesures de la política de la qualitat de les dades entren en joc.

- **Definicions**: què s'entén per qualitat de dades? Cal incloure una secció de definicions perquè tothom pugui entendre qualsevol acrònim o termes poc comuns, i posar en comú l'argot entre TI i negoci.

- **Legislació**: quines lleis i directives s'han de complir? És una bona idea incloure-les en aquesta secció i proporcionar una major profunditat en els procediments i marcs individuals dictats per aquests actes, però fer una supervisió executiva de fins a quin punt són exactament crítiques les dades des d'un punt de vista legal és sempre beneficiós.

- **Documents de referència**: quines altres polítiques i normes estan vinculades a aquesta política? Es van fer servir altres documents per formular aquestes polítiques?

A la taula 6, es relacionen les polítiques amb l'estat de maduresa de les organitzacions.

Taula 6. Polítiques per a cada estat de maduresa de les organitzacions

Estat	Caracterització
Inicial	Les polítiques són informals i no estan documentades. Les accions repetitives són dutes a terme per diferents membres del personal sense coordinació.
Repetible	L'organització intenta consolidar conjunts de dades d'una font única. Hi ha polítiques inicials.
Definit	S'estableixen directrius personalitzades per establir els objectius de gestió de la qualitat de la dada, i les millors pràctiques estan definides.
Gestionat	Polítiques establertes i coordinades en tota l'empresa.
Optimitzat	Notificació automatitzada de l'incompliment de les polítiques de qualitat de dades.

2.6. Processos o tasques sobre les dades

Tots els sistemes i processos existents per a la recopilació, el registre, l'anàlisi i el report de dades han d'assegurar que siguin exactes, vàlides, fiables, oportunes, rellevants i completes dins de l'organització. Les quatre principals famílies de processos són:

- Identificar el problema.

- Valorar el problema i la seva afectació i importància.

- Acció que cal dur a terme i persones involucrades.

- Avaluació dels resultats i mesures de millores.

Aplicar un conjunt de regles de validació de dades a un conjunt de dades una sola vegada proporciona informació sobre l'estat actual de les dades, però no reflecteix necessàriament com les modificacions i actualitzacions del sistema han millorat la qualitat general de les dades en l'organització.

No obstant això, el seguiment dels nivells de qualitat de les dades al llarg del temps com a part d'un procés de monitoratge continu proporciona una visió històrica de quan i quant va millorar la qualitat de les dades.

Els nivells de qualitat de les dades poden ser rastrejats periòdicament (per exemple, cada dia) per mostrar si el nivell mesurat en la qualitat de les dades està dins d'un rang acceptable, comparat amb els límits històrics de control.

Com controlar-ne l'evolució? Mitjançant gràfics i eines de visualització de mètriques.

- Els gràfics de control estadístic poden ajudar a notificar als administradors de dades quan un esdeveniment d'excepció està afectant la qualitat de les dades i on buscar per rastrejar el procés d'informació incorrecta.

- Aquestes mètriques es consoliden com a eines de visualització de mètriques, ja sigui per mitjà d'un quadre de comandament (sistema d'indicadors visual) o *scorecards* (sistema de mètriques que busca mostrar el progrés cap als objectius).

- Aquests sistemes avaluen l'impacte empresarial dels defectes de dades, i determinen les dimensions de la qualitat de les dades que es poden utilitzar per definir les mètriques de qualitat de les dades en un format visual, perquè l'organització prengui les decisions que consideri oportunes.

Al programa de qualitat de dades, s'ha d'incloure:

- Plantilles d'inspecció de dades estandarditzades.

- Qualitat de les dades operacionals.

- Seguiment de temes i solucions.

- Intervenció manual quan sigui necessari.

- Integritat de l'intercanvi de dades.

- Planificació de contingències.

- Validació de dades.

En la mesura del possible, els processos han d'operar sobre una base de «la primera vegada», en comptes d'emprar la neteja o la manipulació de dades per obtenir la informació requerida. És a dir, es tracta de fer un esforç en el fet que les dades siguin correctes des del principi, perquè causin el menor impacte negatiu en l'organització.

En els casos en què no sigui possible, qualsevol ajust de dades ha de seguir un procés clar i documentat, que es pugui verificar fàcilment. Cal incorporar controls apropiats per reduir la probabilitat d'error. Quan s'obtinguin dades de tercers o d'altres departaments, s'acordarà un protocol per garantir que aquestes dades compleixin els mateixos nivells que s'han definit en el programa.

2.7. Govern

El govern de la dada i la qualitat de les dades estan molt relacionats. Si les dades fossin aigua:

- ***Data governance*** seria l'aixeta, encarregada que les persones tinguin les eines i el coneixement adequats, a més de la distribució de l'aigua en les quantitats adequades i a les persones correctes.

- ***Data quality*** seria la depuradora, encarregada que l'aigua sigui bona, no estigui contaminada i mantingui un nivell acceptable de qualitat de forma contínua.

En aquesta relació entre el govern i la qualitat, cal recordar els rols que estan involucrats en la gestió de qualitat de dades:

- **Cap de projecte**: responsable de supervisar el programa d'intel·ligència de negoci o projectes individuals i d'administrar les activitats diàries basades en l'abast, el pressupost i les restriccions d'horari. També és necessari que sàpiga el nivell de qualitat de les dades; per això, interactua amb els representants de negocis per a establir els requisits de qualitat de les dades.

- **Administrador**: ajuda l'organització a comprendre el valor i l'impacte de l'entorn de *business intelligence* i a abordar els problemes que hi sorgeixen. Sovint, els problemes de qualitat de dades es detecten durant els projectes d'intel·ligència de negoci, i l'agent de canvi d'organització pot tenir un paper instrumental: ajudar l'organització a entendre la importància de tractar els problemes.

- **Analista de negoci o dades**: transmet els requisits del negoci, i aquests inclouen requisits detallats de qualitat de les dades. L'analista de dades reflecteix aquests requisits en el model de dades i en els requisits per als processos d'adquisició i lliurament de dades. Junts, asseguren que els requisits

de qualitat es defineixen, es reflecteixen en el disseny i es transmeten a l'equip de desenvolupament.

- **Administrador de dades**: l'administrador de dades és responsable en última instància de la gestió de dades com un actiu corporatiu.

2.8. Estàndards

En el context de la qualitat de la dada, és important disposar d'estàndards o normes sobre les dades.

S'entén com **estàndard** un procés, protocol o tècnica utilitzats per fer una tasca concreta.

Aquestes normes estan destinades a ser utilitzades amb flexibilitat per promoure una millor qualitat de les dades, en lloc de constituir un conjunt rígid de requisits. També poden ser apropiats els enfocaments alternatius per aconseguir aquests objectius, sempre que aconsegueixin el resultat d'obtenir dades fiables que donin suport a una presa de decisions informada i que sigui possible de dur a terme.

Les normes són essencials per assegurar que:

- La recollida de dades és exacta i coherent en tota l'organització.
- Els registres es completen i processen amb precisió.
- Les dades es mantenen segures i confidencials.
- Les sortides de dades poden comparar-se internament i externament.

A la taula 7, es relacionen els estàndards amb l'estat de maduresa de les organitzacions.

Taula 7. Tecnologies per a cada estat de maduresa de les organitzacions

Estat	Caracterització
Inicial	No s'han definit estàndards ni existeixen definicions de les dades.
Repetible	Definicions d'elements de dades i ús de metadades.
Definit	Estàndards de dades de negoci i gestió de metadades.
Gestionat	Normes per a l'intercanvi gestionades per mitjà del procés de supervisió d'estàndards de dades.
Optimitzat	Conformitat amb estàndards per mitjà d'una estructura orientada a les polítiques.

2.9. Tecnologia

La tecnologia ha de donar suport als punts anteriors. Entre el que s'espera de la tecnologia, llistem:

- Procediments estandarditzats per a l'ús d'eines de qualitat de dades, per a l'avaluació i la millora de la qualitat de les dades.

- Ús de tècniques basades en regles de negoci per a la validació de dades.

- Correcció automàtica de dades guiada per polítiques i regles de negoci definides. Quan parlem d'automatització, pot fonamentar-se en regles senzilles (*if... then... else*) o fins i tot en patrons complexos fonamentats en *machine learning*.

- Anàlisi d'impacte i escenaris hipotètics compatibles amb el panell de control i les eines de generació d'informes.

A la taula 8, es relaciona la tecnologia amb l'estat de maduresa de les organitzacions.

Taula 8. Tecnologies per a cada estat de maduresa de les organitzacions

Estat	Caracterització
Inicial	Es desenvolupen internament rutines *ad hoc*.
Repetible	Es disposa d'eines per avaluar la qualitat de les dades. Per a l'anàlisi de dades, l'estandardització i la neteja.
Definit	Procediments estandarditzats per a l'ús d'eines de qualitat de dades, per a l'avaluació i la millora de la qualitat de les dades.
Gestionat	Correcció automàtica de dades guiada per polítiques de governança i regles de negoci.
Optimitzat	Els usuaris no tècnics poden definir i modificar dinàmicament les regles i les dimensions de la qualitat de les dades.

2.10. Metodologies

Com detectar que no hi ha qualitat de dades? Trobem principalment dues metodologies:

- De dins cap a fora.
- De fora cap a dins.

2.10.1. Metodologia de dins cap a fora

El mètode de dins cap a fora comença amb l'anàlisi de les dades. Es duu a terme un examen de les dades (sobre les fonts de dades existents), utilitzant la tecnologia de perfilat de dades. Les imprecisions de dades es revelen a partir del procés, i després s'analitzen juntes per generar un conjunt de qüestions sobre dades, per a la seva posterior resolució.

L'anàlisi s'ha de dur a terme per un analista de dades. La metodologia comença amb un conjunt complet i correcte de regles que defineixen l'exactitud de les dades. Es tracta de treballar amb les metadades (és a dir, dades sobre les dades).

Per exemple:

- La dada *temperatura* pot tenir la metadada *rang* que, per dades a Espanya, podria ser de -20 a 45 graus centígrads.

- Un altre metadada en aquest context seria *temperatura_ciutat*, que relacionaria la dada *temperatura* amb la dada *ciutat*.

El procés de determinar les metadades correctes implica inevitablement relacionar TI i negoci. L'analista ha de detectar el comportament en les dades, i requerirà una consulta per determinar per què és així. Això, sovint, condueix a modificacions en les metadades. Aquestes consultes són sempre productives, perquè la pregunta sempre està recolzada per informació de les dades.

2.10.2. Metodologia de fora cap a dins

Aquest mètode busca problemes en el negoci, no en les dades. Identifica fets que suggereixen que els problemes de qualitat de les dades estan tenint un impacte en el negoci.

Es busquen esdeveniments com devolucions de mercaderies, reclamacions de clients, retards en l'obtenció de productes d'informació completats, altes quantitats de treball requerides per obtenir productes d'informació produïts, etc.

En aquest enfocament, es fan servir entrevistes de negoci que permeten determinar el nivell de confiança en l'exactitud de les dades procedents dels sistemes d'informació, a més del seu nivell de satisfacció respecte a aconseguir tot el que necessiten.

També pot incloure la cerca de decisions preses per la corporació que van ser decisions equivocades. A continuació, les dades s'examinen per determinar si tenen inexactituds que contribueixen als problemes, a més de l'abast de la contribució. Generalment, aquest examen apunta al problema específic i no és un exercici de perfils de dades exhaustiu, encara que podria estendre's si hi ha l'evidència d'un problema de qualitat generalitzat de dades.

2.10.3. Comparació de mètodes

Cap dels enfocaments és superior a l'altre: tots dos aporten valor al procés. El mètode de *dins cap a fora* és generalment més fàcil d'aconseguir i necessita menys temps per a la seva execució. Un sol analista pot analitzar una gran quantitat de dades en poc temps. L'enfocament de *fora cap a dins* requereix dedicar molt de temps a entrevistar persones d'altres departaments.

Vegem-ne dos exemples:

- Imaginem una empresa on, en la generació de comandes, hi ha un error en la inclusió de l'identificador del proveïdor. Freqüentment, en funció del volum de comanda, en certs sectors els proveïdors accedeixen a descomptes (per fidelització). En aquest escenari, el proveïdor pot estar perdent quantitats significatives cada any a causa de l'error, i ser completament desconeixedor del que està succeint. L'enfocament de *dins cap a fora* és l'adequat en aquest cas.

- L'oposat també és cert. Imaginem que el sistema d'informació assigna identificadors erronis a algunes peces d'un fabricant. Això deriva en errors d'enviament i de devolucions. L'anàlisi de *fora cap a dins* podria identificar la disparitat en els codis.

2.11. La qualitat de la dada en el context del govern de la dada

En el context de govern de la dada, la qualitat de la dada és una funció més per dur a terme. Com ja sabem, cada funció té diferents activitats (planificació, control, de desenvolupament i operatives), cadascuna de les quals és duta a terme pel rol corresponent.

Per la qualitat de dades, aquestes activitats són:

- Desenvolupar i promoure la conscienciació sobre la qualitat de les dades (activitat operativa).
- Perfilar, analitzar i avaluar la qualitat de les dades (activitat de desenvolupament).
- Definir requisits de qualitat de dades i regles de negoci (activitat de desenvolupament).
- Provar i validar els requisits de qualitat de les dades (activitat de desenvolupament).
- Definir indicadors de qualitat de dades i nivells de servei (activitat de planificació).
- Mesurar i monitorar la qualitat de les dades (activitat de control).
- Gestionar problemes de qualitat de dades (activitat de control).
- Corregir defectes de qualitat de dades (activitat operativa).
- Dissenyar i implementar procediments de qualitat de dades operatives (activitat de desenvolupament).
- Monitorar els procediments operacionals i el rendiment de la gestió de qualitat de dades (activitat de control).
- Auditar la qualitat de dades (activitat de control).

El programa de qualitat de dades cobreix aquestes funcions, que formen part del marc més general.

3. Desenvolupem un programa de qualitat de dades

3.1. Desenvolupament d'un programa de qualitat de dades

Com ja hem comentat, un programa de qualitat de dades proporciona una guia de bones pràctiques per millorar i aprofitar millor les dades. Implementar un programa de qualitat de les dades en una organització exigeix superar una sèrie de reptes que han de ser coneguts i superats. Les raons per les quals les organitzacions no segueixen una iniciativa formal i planificada de gestió de la qualitat de les dades inclouen les següents:

- Cap de les unitats de negoci o departament sent que és responsable del problema.
- Es necessita cooperació interdepartamental.
- Es requereix que l'organització reconegui que té problemes significatius.
- Es necessita disciplina.
- Es requereix una inversió de recursos financers i humans.
- Es percep que és un procés costós.
- El retorn de la inversió és, sovint, difícil de quantificar.

Dins el desenvolupament del programa, podem identificar dues grans tasques crucials:

1) Avaluació. És a dir, des d'on partim i les mesures que cal aplicar-hi.

2) Seguiment (de les solucions aplicades). És a dir, un cop detectats els errors i les oportunitats, i decidides les tasques i accions que cal dur a terme, valorar quin impacte han tingut aquestes en l'organització, i insistir, si cal, en el procés de millora.

Aquestes dues grans tasques es descomponen en diferents passos que es repeteixen de forma cíclica, com il·lustra la figura 4:

- **Valoració**: en aquest pas, s'identifiquen els elements que s'han d'avaluar per la qualitat de les dades. Normalment, aquests seran elements de dades

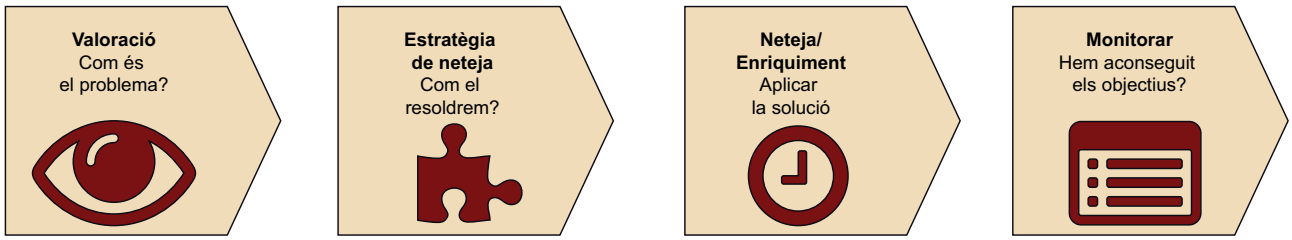
considerats crítics per a les operacions empresarials i els informes de gestió associats, i caldrà avaluar quines dimensions de qualitat de dades farem servir i quina és la seva ponderació associada.

- **Estratègia de neteja**: per a cada dimensió de qualitat de dades, es defineixen els valors o rangs que representen dades de qualitat bona i dolenta. Això permet classificar les dades i determinar quines necessiten ser tractades i quina mesura s'hi ha d'aplicar.

- **Neteja i enriquiment**: en aquest pas, s'apliquen els criteris de neteja i de millora de les dades i processos per prevenir errors futurs.

- **Monitorar**: finalment, es revisen els resultats i es determina si la qualitat de les dades és acceptable o no. En introduir canvis en els models de negoci, és possible que apareguin nous problemes de qualitat de dades.

Figura 4. Passos en el marc de la qualitat de les dades



Font: David Cabanillas, adaptat de Deloitte

Per controlar el flux anterior i assegurar que es compleixin uns terminis d'execució, és aconsellable dur a terme les accions següents:

1) Determinar un calendari d'actuació: la programació ha d'estar expressada en termes clars, incloent un glossari —si es considera necessari— i ha de ser compartida entre la part de negoci i TI.

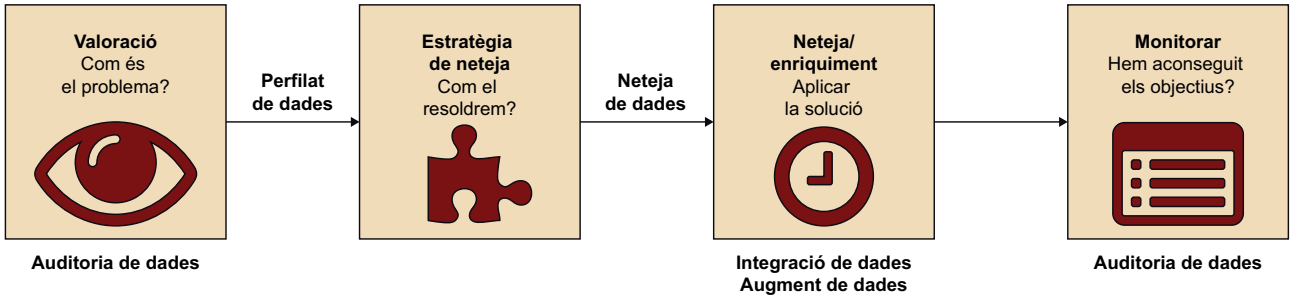
2) Establir la freqüència d'avaluació: la periodicitat de les accions de descobriment i d'avaluació de la qualitat de les dades es determinarà en funció de la criticitat de la informació continguda en les dades i la rellevància de les àrees a les quals afectin.

3) Designar els responsables: l'absència de propietaris de les dades és el principal problema de fons en un elevat percentatge de qüestions relacionades amb la qualitat de dades. Assignar responsabilitats i arribar a un consens és la millor prevenció.

4) Definir els requisits de *reporting:* la retroalimentació és imprescindible per fomentar el flux de coneixement i garantir-ne l'actualització. Per maximitzar l'eficiència d'aquesta comunicació, han d'establir-se els termes en què es durà a terme, abans de començar el programa.

En els subapartats següents, es descriuen les tasques relacionades amb cadascun dels passos del programa. La figura 5 relaciona cadascuna d'aquestes tasques amb el pas concret del programa de qualitat de dades en la qual es troba.

Figura 5. Tasques i passos en el marc de la qualitat de dades



Font: David Cabanillas, adaptat de Deloitte

3.1.1. Perfilat de dades

Una de les tasques inicials és el perfilat de dades. Dins del programa de qualitat de dades, el perfilat correspon a les fases de **valoració** i **estratègia de neteja**. Tal com comenta Ralph Kimball:

> S'entén com **perfilat de dades**, o *data profiling*, l'anàlisi del contingut, l'estructura i anomalies de les dades.

El resultat d'aquesta tasca són perfils de dades, cosa que proporciona un mitjà metòdic, repetible, consistent i basat en mètriques per avaluar les seves dades. En què consisteix el perfilat de dades? Inclou diferents tipus d'anàlisi per classificar la dada:

- **Anàlisi de completesa**. Dona resposta a la pregunta següent: amb quina freqüència un atribut té valors, està buit o és nul?

- **Anàlisi d'unicitat**. Dona resposta a les preguntes següents: quants valors únics trobem en un atribut? N'hi ha duplicats? Aquest fenomen és esperat i normal?

- **Anàlisi de distribució**. Respon a la pregunta: quina és la distribució de freqüències dels valors per a un determinat atribut?

- **Anàlisi de rangs**. Respon a la pregunta: quins són els valors mínims, màxims, la mediana i la mitjana per als valors d'un determinat atribut?

- **Anàlisi de patrons**. Respon a les preguntes següents: quins formats trobem per a un atribut? Com es distribueixen els valors en aquests formats?

Imaginem que tenim la base de dades de clients, i un dels atributs és el DNI (Document Nacional d'Identitat a Espanya). Hi ha diferents regles que defineixen i validen el fet que un valor en aquesta columna és, efectivament, un DNI:

- El valor ha de tenir nou caràcters.
- Els primers vuit han de ser nombres.
- L'últim ha de ser una lletra.
- La lletra es calcula prenent totes les lletres, excepte la Ñ, la I i la O, perquè poden induir a errors, en un ordre concret (que no és l'ordre alfabètic lògic, sinó aquest: TRWAGMYFPDXBNJZSQVHLCKET), i seleccionant aquella que coincideix en la posició igual a la resta de dividir el número del DNI entre 23.

Fer servir aquestes regles correspon en aquest cas a aplicar l'anàlisi de patrons. En alguns casos, aquestes regles són deduïbles; en d'altres, la combinació d'aquestes anàlisis ajuda a descobrir tant els requisits de qualitat de dades com la forma d'avaluar la qualitat de la dada.

Cal comentar que els requisits i les regles descoberts no només s'han d'aplicar sobre el conjunt de dades que s'està analitzant, sinó que freqüentment impliquen transformar les fonts d'origen, fins i tot la modificació de sistemes d'informació i de processos de negoci.

3.1.2. Neteja de dades

Un cop coneguts els problemes que tenen les dades, cal aplicar accions per corregir aquests problemes.

S'entén com **neteja de dades**, o *data cleansing*, el procés d'arreglar o esborrar dades que són incorrectes, incompletes, duplicades o amb un format incorrecte.

També és possible fer referència a aquest concepte com *data scrubbing* o *data correction*. Dins el programa de qualitat de dades, la neteja de dades correspon a les fases d'**estratègia de neteja** i **neteja i enriquiment**.

L'objectiu d'aquesta tasca és millorar la confiança de l'organització en les seves dades, però podem trobar-nos en diferents escenaris. Per exemple, problemes relacionats amb:

- **Dades procedents d'una única font**. Els errors:

– En un nivell d'esquema, fan referència al disseny (falta d'integritat de les restriccions, disseny ineficient) i s'observen per problemes d'unicitat, integritat referencial, etc.

– En un nivell d'instància, fan referència a problemes d'introducció de dades i s'observen duplicats, valors contradictoris, errors ortogràfics, etc.

- **Dades procedents de diferents fonts d'origen**. Els errors:

– En un nivell d'esquema, fan referència a dissenys de dades i esquemes heterogenis, fet que es tradueix en conflictes d'estructura, de noms, etc.

– En un nivell d'instància, fan referència a inconsistència en les dades, que es tradueix en inconsistències en afegir, combinar i creuar dades.

Quan es troben problemes de qualitat de dades (per exemple, en importar dades al magatzem de dades), hi ha quatre accions viables que es poden prendre:

- Rebutjar l'error.
- Acceptar l'error.
- Corregir l'error.
- Aplicar un valor predeterminat.

Quan la precisió és més important que la completesa, pot ser apropiat rebutjar l'error. Quan se sap que les dades contenen errors, però estan dins del nivell de tolerància, llavors pot ser apropiat acceptar l'error. Quan el valor es pot determinar, llavors l'error es pot corregir. Finalment, quan no es pot determinar el valor correcte i la integritat és molt important, llavors un valor predeterminat pot ser substituït per les dades errònies. En qualsevol cas, és important que els administradors de les dades entenguin les implicacions de la solució escollida i aquesta estigui consensuada amb la part de negoci.

I quines opcions tenim per a la neteja de dades? Disposem de diferents tècniques:

- **Duplicació**: permet detectar registres duplicats.
- **Cerca i reemplaçament**: permet trobar i reemplaçar registres a partir d'un criteri.
- **Divisió**: permet dividir un registre en diversos valors segons un criteri.
- **Diccionaris/referències**: permeten validar registres respecte a un conjunt de valors que s'accepten com a referent de qualitat.

Aquestes tècniques solen combinar-se entre elles per resoldre els problemes de qualitat, i poden ser manuals o automàtiques (en què el sistema identifica i proposa solucions).

3.1.3. Auditoria de dades

L'objectiu de l'auditoria és comprendre el grau en què els problemes de qualitat de les dades existeixen en la nostra organització, és a dir, l'extensió i la gravetat dels defectes de les dades. Aquest procés implica revisar les mètriques clau, a part dels valors, per crear conclusions. Els informes d'auditoria poden crear-se per mesurar el progrés en l'assoliment dels objectius de qualitat de les

dades i complir amb els acords en la primera fase del programa de qualitat de dades.*

*** Exemple de plantilla per a informe: https://goo.gl/4TvXtj**

D'aquesta manera, l'informe d'auditoria determina:

- Declaracions concretes sobre el *status quo* de la qualitat de les dades.
- Recomanacions d'orientació, és a dir, solucions per optimitzar la qualitat de les dades a curt termini i mantenir-ne el nivell assolit a llarg termini.
- Identificar possibles oportunitats d'optimització dels processos.
- Visió general del possible potencial d'estalvi o beneficis aconseguits i els costos associats a aplicar les solucions.

Imaginem que tenim dades geolocalitzades a la nostra organització. El procés d'auditoria seguiria aquest procés:

- Revisió de bases de dades internes.
- Revisió de bases de dades externes.
- Revisar dades fora de rang mitjançant una referència GIS (*geographic information system*).
- Revisar dades fora de rang amb eines estadístiques.
- Informar de la qualitat de les dades geolocalitzades.

3.1.4. Integració de dades

Com és possible deduir, en els processos de qualitat de dades hem de transformar les dades amb l'objectiu d'augmentar la seva qualitat. En aquest sentit, no cal reinventar la roda: dins de la qualitat de dades, una de les tasques més rellevants és la integració de dades.

Lectura complementària

J. Curto (2017). *Introducción al Business Intelligence*. Editorial UOC.

S'entén per **integració de dades**, o *data integration*, el conjunt d'aplicacions, productes, tècniques i tecnologies que permeten una visió única i consistent de les nostres dades de negoci.

És a dir, la integració permet accedir a les dades presents en diferents fonts, recuperar-les, combinar-les, transformar-les (seguint les regles de qualitat de dades) i aplicar els canvis en les fonts de destinació.

Imaginem que tenim un llistat amb dos arxius de productes. Una empresa pot haver venut els mateixos productes en diferents sucursals, però els productes es poden vendre sota noms diferents: la marca i els patrons de descripció en cada arxiu es van basar en el personal d'entrada de dades. El primer repte en la integració de dades és reconèixer que el mateix client existeix en cadascuna de les dues fonts, i el segon repte és combinar les dades en una sola vista del producte (consolidació). Amb les dades del client, sovint trobem un camp comú, per exemple, el DNI, que es poden utilitzar per identificar-lo i integrar-lo.

3.1.5. Augment de dades

L'augment de dades és l'últim pas. En aquest pas, es busca augmentar el valor del conjunt de dades inicial mitjançant la incorporació de dades externes.

Per exemple, aquest procés pot ser utilitzat per afegir les coordenades geogràfiques a les adreces de la base de dades de clients.

Tenim diferents opcions respecte a les fonts de dades:

- **De pagament**: són fonts de dades preparades per tercers per al seu consum. Poden ser accessibles per mitjà de fitxers independents o una API (*application programming interface*). La qualitat de la dada la gestiona el proveïdor, i el pagament pot tenir diferents modalitats (pagament únic, subscripció, etc.). Entre aquestes fonts, podem considerar Crunchbase* o Bloomberg.**

*** Més informació a: http://www.crunchbase.com**

**** Més informació a: http://www.bloomberg.com**

- ***Open data***: són fonts de dades de tercers a les quals es pot accedir lliurement i fàcilment. Com en el cas anterior, el proveïdor assegura el nivell de qualitat de la dada. Alguns ajuntaments ofereixen les seves dades en aquesta modalitat. Per exemple, l'Ajuntament de Barcelona.***

***** Més informació a: http://opendata.bcn.cat**

- **Públiques**: són fonts de dades disponibles de forma pública, però l'accés de les quals no està preparat per al consum. Freqüentment, aquestes dades poden recuperar-se de manera automàtica mitjançant tècniques de *web scrapping*. Un exemple d'això és recuperar informació de Wikipedia.

- ***Crowdsourcing***: són fonts de dades que es generen a partir de la col·laboració dels consumidors. Per exemple, demanar ajuda a la nostra comunitat per a la traducció del nostre programari.

Open data

Les dades obertes són dades que poden ser utilitzades, reutilitzades i redistribuïdes lliurement per qualsevol persona, i que es troben subjectes, com a màxim, al requeriment d'atribució i de compartir-se de la mateixa manera en què apareixen.

En la combinació de les dades pròpies i les de tercers, les tècniques d'integració de dades tenen un paper fonamental.

3.2. Millors pràctiques

Millors pràctiques

Les millors pràctiques són un conjunt d'accions que han rendit un servei en un determinat context i que s'espera que, en contextos similars, rendeixin resultats similars.

Com hem après durant aquest apartat, un programa de qualitat de dades passa per diferents fases i té diferents tasques. Dur a bon port aquest tipus de projectes no és senzill, ja que requereix orquestrar processos, persones i tecnologia.

En aquest apartat, ens centrarem en revisar les millors pràctiques que poden ajudar al desplegament d'aquest tipus de projectes.

La majoria de les organitzacions deixen la qualitat de les dades als administradors de bases de dades (amb la suposició que si aquests són amos de les dades, són també responsables de la seva qualitat), però aquest enfocament no és sempre el més assenyat. Les organitzacions han de adherir-se a un programa de qualitat de dades per assegurar que les seves dades són de qualitat. Al programa, es marquen els objectius i fins a on es vol i s'ha d'arribar a la qualitat de les dades, s'hi descriuen els passos i, finalment, s'hi avaluen els resultats. Com ja hem comentat, per aplicar i seguir el programa, tant negoci com TI s'han de coordinar.

Recordeu!

La prevenció és millor que la cura. Hem de mirar d'evitar introduir dades de baixa qualitat en els nostres sistemes.

Què hem de tenir en compte com a millors pràctiques? Hi ha una sèrie de principis clau que cal considerar en l'establiment d'un programa eficaç:

- **Començar de manera acotada**: l'objectiu final és dissenyar una visió completa de l'estat de qualitat de dades en l'organització. No obstant això, pot ser una tasca inabordable directament. Un bon full de ruta d'implementació ha de començar amb una àrea acotada (per exemple, àrea de clients) i construir-hi a sobre el programa al llarg del temps.

- **La qualitat de les dades és un procés continu**: un programa no és un esdeveniment únic, sinó un procés continu que evolucionarà amb el temps. Sovint, es tracta d'un canvi cultural en una organització i, per tant, la seva implementació porta temps. El programa ha de ser reavaluat periòdicament, i modificat quan sigui necessari.

- **Enfocament en els objectius de qualitat de les dades**: en aquest tipus d'iniciatives, no es tracta de solucionar ràpidament problemes de qualitat de dades sense les arrels del problema. Cal establir les mesures de qualitat de dades importants per a l'organització, entendre els nivells de qualitat de dades requerides pel negoci i establir els objectius de qualitat de dades apropiades.

- **Combinar qualitat de dades en el desenvolupament/implementació de programari**: sovint, la qualitat de dades es considera un procés posterior al desenvolupament/implementació de programari. Aquest enfocament reactiu suposa un major esforç per l'organització. Cal establir un enfocament proactiu que consideri ja des del principi la qualitat en tots els processos de negoci.

- **Començar amb les dades d'alt retorn**: una forma de convèncer l'organització de la necessitat de tenir un programa global de qualitat de dades consisteix a centrar-se inicialment en les àrees específiques de dades que proporcionaran el retorn d'inversió més alt si s'apliquen aquestes mesures.

- **Maximitzar la participació multidepartamental en el programa**: aquí la col·laboració és crucial en el disseny i la implementació de mesures tant en un àmbit de negoci (usuaris de dades) com de TI (administradors de dades).

- **Hi ha més d'una solució del programa**: les dades solen fluir a través d'una organització des de moltes fonts/aplicacions, mitjançant diferents serveis d'integració de dades i en molts repositoris. La qualitat de les dades es pot gestionar en qualsevol punt del flux de dades, depenent de les polítiques de gestió de dades, els requisits empresarials, la propietat de les dades, l'arquitectura del sistema, etc. Comprendre les mesures de qualitat de dades, el cicle de qualitat de dades i mantenir-ne els principis són aspectes clau per assegurar una base sòlida del programa.

Flux de dades

El flux de dades és el moviment que tenen les dades en un sistema determinat.

- Elements que cal tenir en compte: aplicar la qualitat de la dada transforma l'organització, de manera que cal documentar els estats abans i després. Això significa que és convenient tenir en compte els elements següents:

– Llista de conjunts de dades i elements que cal tractar.
– Llista de tipus de dades i categories.
– Catàleg, esquema o mapa d'on resideixen les dades.
– Discussió de solucions de neteja per categoria de dades.
– Diagrames de flux de dades existents.
– Diagrames de flux de treball existents.
– Pla per decidir quan i on s'accedeix a les dades per netejar-les.
– Anàlisi de com canviarà el flux de dades després de la implementació del projecte.
– Discussió de com el flux de treball canviarà després de la implementació del projecte.
– Llista d'actors afectats pel projecte.
– Pla per educar les parts interessades pel que fa als beneficis del projecte.
– Pla per a la formació d'operadors i usuaris.
– Llista de mesures de qualitat de dades i mètriques per monitorar.
– Pla de quan i on monitorar.
– Pla per a la neteja inicial i, després, regular.

Flux de treball

Com s'estructuren les tasques, com es duen a terme, quin és el seu ordre correlatiu, com se sincronitzen, com flueix la informació que suporta les tasques i com se li fa un seguiment al compliment de les tasques.

3.3. Impacte

Perquè la part de negoci i TI vagin de la mà en l'aplicació del programa de qualitat de dades, cal valorar el seu impacte abans de la seva execució. Són diversos els punts que cal tenir en compte en aquesta valoració:

- **Eficiència operativa**: temps i costos de neteja de dades o correccions de processament.
- **Mètriques**: mesures de rendiment inexactes per als empleats.

- **Risc/compliment**: les dades incorrectes condueixen a un risc *X*, violacions de normes.
- **Ingressos**: cost d'oportunitat perduda, identificació d'oportunitats.
- **Productivitat**: disminució de la capacitat de processament directe mitjançant l'automatització serveis.
- **Eficiència de l'adquisició**: major facilitat d'ús per al personal (vendes, centre de trucades, etc.), major facilitat d'interacció per als usuaris.
- **Reducció**: reducció del temps des de la comanda fins al lliurament.
- **Rendiment**: deteriorament en la presa de decisions.

Una bona metodologia que cal seguir és l'ús d'una plantilla/*template* sobre els diferents impactes que produirà el programa de qualitat de dades. Un exemple és l'estructura que es mostra a la figura 6:

- **ID problema:** identificador assignat per al problema.
- **Problema de la dada:** descripció del problema.
- **Impacte:** descripció de l'impacte comercial atribuïble a l'emissió de dades; pot haver-hi més d'un impacte per a cada problema de dades.
- **Mesura:** mesurament en el grau d'impacte.
- **Importància:** una estimació de la quantificació dels impactes en un àmbit monetari/temporal.

Figura 6. *Template*

ID del problema	Problema de la dada	Impacte	Mesura	Importància

Font: David Cabanillas

4. Tècniques i tecnologia per a la qualitat de la dada

El programa de qualitat de dades se sustenta en tècniques i tecnologia que permeten automatitzar les principals tasques que hem discutit en anteriors apartats.

4.1. Tècniques

Quan parlem de tècniques en el context de qualitat de dades, en fem referència a dos tipus:

- Aquelles que permeten l'anàlisi visual de les dades, amb l'objectiu de detectar-hi anomalies.
- Aquelles que permeten automatitzar el procés d'identificar problemes de qualitat i el seu tractament.

4.1.1. Tècniques visuals

En el moment d'identificar problemes de qualitat de la dada, una primera tècnica és la visualització. Aquesta tècnica pot donar suport al descobriment de patrons en les dades. La manipulació visual de dades (fent servir agregacions, agrupaments, classificacions, escales de colors, diferents tipus de gràfics, etc.) permet, de vegades, identificar els problemes següents:

- Valors que falten.
- Valors fora de rang.
- Resultats de negoci que no encaixen.

En aquest sentit, eines analítiques visuals com Tableau,* QlikSense,** GGobi*** o Improvise**** encapsulen aquestes tècniques i permeten als analistes construir vistes multidimensionals de les dades que ajuden a avaluar els problemes de qualitat de les dades.

* http://www.tableau.com
** http://www.qlik.com
*** http://www.ggobi.org
**** https://goo.gl/pMG6LS

A més, un cop creades les mètriques de control, el *scorecard* permet fer un seguiment del programa de qualitat, com il·lustra la figura 7.

Tot i que les tècniques visuals són interessants, no sempre resulten suficients per a l'anàlisi de la qualitat de la dada, i per això cal recórrer a tècniques d'automatització.

Figura 7. Exemple d'*scorecard*

DATA ELEMENT LEVEL

TABLE XYZ

Expected fields to be populated 100%

#	Table Column Name	8/4/2012	8/11/2012	8/18/2012	8/25/2012	9/1/2012	9/8/2012	Trend
1	Key Field 1	100%	100%	100%	100%	100%	100%	
2	Key Field 2	100%	100%	100%	100%	100%	100%	
3	Key Field 3	100%	100%	100%	100%	100%	100%	
4	Key Field 4	100%	100%	100%	100%	100%	100%	
5	Field 05	91%	72%	67%	70%	70%	70%	
6	Field 06	72%	78%	80%	81%	81%	83%	
7	Field 07	94%	96%	96%	100%	100%	98%	
8	Field 08	88%	74%	72%	72%	72%	70%	
9	Field 09	81%	74%	65%	70%	67%	64%	
10	Field 10	84%	70%	63%	72%	70%	66%	
11	Field 11	88%	74%	70%	74%	72%	70%	
12	Field 12	84%	74%	72%	74%	72%	72%	
13	Field 13	84%	74%	70%	74%	72%	68%	
14	Field 14	94%	98%	98%	95%	95%	98%	
15	Field 15	66%	74%	72%	70%	77%	79%	
16	Field 16	78%	80%	85%	88%	84%	83%	
17	Field 17	47%	52%	46%	44%	44%	47%	
18	Field 18	19%	15%	17%	16%	16%	17%	
19	Field 19	100%	100%	100%	100%	100%	100%	
	Average Score	83%	79%	77%	79%	79%	78%	
		B	C	C	C	C	C	

Font: Midior Consulting

4.1.2. Tècniques per a l'automatització

Com succeeix en altres àmbits, ja no és possible tractar de forma manual la qualitat de la dada. De fet, la gran majoria de les tasques dins d'aquesta disciplina es fonamenten en l'anàlisi descriptiva, la mineria de dades i la mineria de textos.

1) L'**anàlisi descriptiva** permet diferenciar els valors incomplets, buits, atípics i rangs de valors, a més de conèixer la distribució de freqüències, valors mínims/màxims, etc. Aquesta anàlisi es farà per a cada columna del conjunt de dades analitzades.

2) La **mineria de textos** permet analitzar atributs en format de text per identificar-hi valors atípics.

3) La **mineria de dades** permet:

- Trobar agrupacions de dades i patrons. L'anàlisi de patrons s'utilitza per determinar si els valors de dades en un camp o camps coincideixen amb el format o l'estructura esperats. Les tècniques de *clustering* s'utilitzen per detectar una varietat d'errors relatius a una mètrica de distància escollida.

clustering

El *clustering* o clusterització és un procediment d'agrupació d'una sèrie d'elements d'acord amb un criteri.

– La distància euclidiana és útil per a la detecció d'emissions numèriques i d'unitats de mesura coherents.

– Les distàncies basades en caràcters (distància de Levenshtein), *token-based* (*atomic strings*) i en la fonètica (*soundex*) són útils per detectar inconsistències en el text, com ara errors ortogràfics, diferents ordenaments de termes i paraules fonèticament similars.

- Identificar anomalies. Mitjançant l'anàlisi dels valors atípics (ja sigui per mitjà de l'anàlisi estadística, o tècniques més avançades com l'anàlisi de sèries temporals).

- Completar o combinar dades. Les dades es poden vincular implícitament definint criteris d'unió en valors similars, usant un valor únic generat o codis de coincidència basats en algoritmes de lògica difusa (*fuzzy logic*). També és possible fer servir altres tècniques d'extrapolació. Per exemple, fent servir funcions heurístiques per extrapolar el salari actual d'un client, a partir del seu salari de fa cinc anys.

Funcions heurístiques

Les funcions heurístiques tenen coneixement d'informació de proximitat.

- Analitzar i identificar causes d'errors. La tècnica de *root cause analysis* permet comparar escenaris amb errors i sense per identificar els factors que els han motivat.

Com es fan servir aquestes tècniques?

L'anàlisi de columnes proporciona una inspecció de dades en la qual un analista passarà per tots els registres, o per un subconjunt de registres dins d'una taula, i iniciarà la identificació de diverses estadístiques sobre les dades. Els tipus d'anàlisi inclouen el nombre total de registres, el seu tipus de dades, quants registres contenen valors nuls o mancants, cardinalitat o unicitat, valors mínim i màxim, mitjana, desviació estàndard, duplicats i molt més. Tenir una idea de com es veuen les dades ajudarà a determinar quanta feina caldrà per solucionar o corregir les imprecisions o inconsistències en les dades.

Lectura complementària

P. G. Elmagarmid ; V. S. Verykios (2007). «Duplicate record detection: A survey». *IEEE Transactions on Knowledge and Data Engineering* (vol. 19, núm. 1, pàg. 1-16).

Això no acaba aquí, ja que l'analista pot aplicar l'anàlisi de dominis, i focalitzar-se en valors i rangs de dades esperats o acceptats. És a dir, decidir si un valor de dades específic és acceptable o cau dins d'un rang acceptable de valors. Un exemple d'aquests criteris podria ser un camp de sexe on els únics valors acceptables són *male* (M) o *female* (F). Un altre exemple podrien ser les cinquanta

abreviatures de dos estats de caràcters en el format apropiat (sense espais, majúscules, sense períodes, etc.). Un informe d'anàlisi de domini produiria un gràfic que indicaria percentatges de registres que queien dins o fora del valor acceptable.

Un altre pas consistiria a identificar valors atípics, és a dir, valors fora de rang. Els valors extrems poden ser valors atípics univariats estàndard, o específics d'un tipus. Per exemple, els valors atípics de la sèrie temporal prenen generalment dues formes: un *outlier* additiu és un moviment inesperat i transitori en un valor mesurat al llarg del temps, mentre que un *outlier* inesperat és un moviment inesperat que persisteix en el temps.

Lectura complementària

V. Chandola; A. Banerjee; V. Kumar (2009). «Anomaly detection: A survey». *ACM Comput. Surv.* (núm. 41, vol. 3, art. 15).

Outlier

És una observació numèricament distant de la resta de les dades, és a dir, un valor atípic.

I està clar que cal continuar amb les potencials dependències amb altres conjunts de dades o sistemes.

Un cop feta l'anàlisi i conegut l'estat, es farien servir les tècniques de transformació per aplicar les mesures pertinents, que van, com ja hem vist, des de rebutjar l'error fins a transformar la dada.

Lectura complementària

Y. Huhtala; J. Kärkkäinen; P. Porkka; H. Toivonen (1999). «Tane: An efficient algorithm for discovering functional and approximate dependencies». *The Computer Journal* (núm. 2, vol. 42).

4.2. Tecnologia

Des de la perspectiva tecnològica, en la qualitat de dades tenim diferents tipus d'eines. El procés de qualitat de dades s'ha dut a terme de forma manual, mitjançant *scripts* (SQL, llibreries de R com dplyr,* llibreries per Python com Pandas,** etc.) i mitjançant programes informàtics propietaris i codi obert.

*** Més informació a: https://goo.gl/v3RbFJ**

**** Més informació a: http://pandas.pydata.org**

Aquí ens centrarem en els programes. Depenent del fabricant, les eines inclouen més o menys característiques. Hi ha, principalment, tres tipus:

1) Especialitzades en la qualitat de dades.

2) Les que inclouen funcionalitat de qualitat de dades, sense que en sigui el seu focus. Per exemple, les eines d'integració de dades.

3) Les que complementen les eines de qualitat de dades. Per exemple, les eines de gestió del projecte de qualitat de dades, o per fer proves.

4.2.1. Eines de qualitat de dades

Aquest tipus d'eines estan especialitzades en la qualitat de dades. De vegades, formen part d'una cartera més gran de productes (com una plataforma d'inte-

gració de dades o una solució per al govern de la dada), i altres són simplement un producte independent.

Això significa que inclouen les característiques següents:

1) anàlisi exploratòria de dades

2) perfilat de dades

3) capacitat de crear regles de negoci

4) transformacions de dades i *workflow* associades a qualitat de dades

5) gestió d'excepcions

6) gestió i accés a taules de referència

7) identificador de duplicats i identitats

8) dominis de dades de negoci i capacitats d'autodescobriment de dominis

9) gestió de metadades

Destaquem solucions com les d'Informatica,* Tamr,** DataCleaner,*** Trillium Enterprise Data Quality**** o Datiris Profiler,***** tot i que cal comentar que els principals fabricants del mercat disposen de solucions pròpies.

*** http://www.informatica.com**
**** http://www.tamr.com**
***** https://goo.gl/N9o3Gu**
****** https://goo.gl/KahJFS**
******* http://www.datiris.com**

La figura 8 il·lustra la plataforma de qualitat de dades. La plataforma es connecta amb aquells sistemes d'informació (interns i externs) que tenen dades rellevants. Aquesta comunicació, com ja sabem, és bidireccional per poder propagar els canvis.

Dins d'aquesta categoria, podem trobar un nou tipus d'eines híbrides que combinen la manipulació ràpida de dades, la qualitat de dades i la integració de dades, en una nova categoria denominada *data wrangling*, el principal client de la qual és l'analista de dades i el científic de la dada. Destaquem eines com Trifacta* o Open Refine.** A més, les eines mateixes de *data science* com Dataiku*** o Tibco Spotfire***** inclouen també aquestes capacitats. Cal destacar, igualment, que Pentaho Data Integration***** inclou múltiples *plugins* que estenen la seva funcionalitat en l'àmbit de la qualitat de la dada.

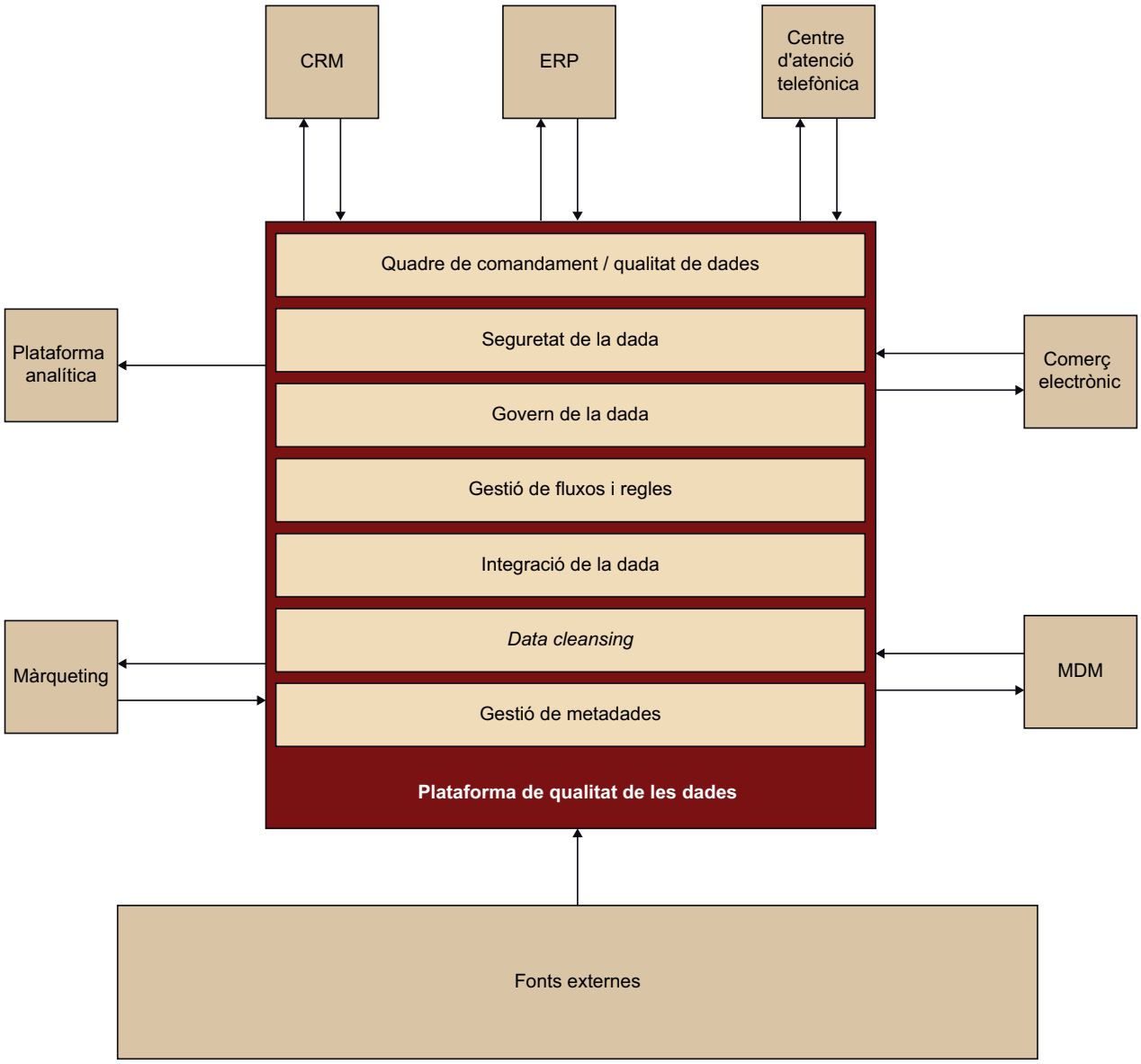
*** http://www.trifacta.com**
**** http://openrefine.org**
***** http://www.dataiku.com**
******* http://spotfire.tibco.com**
****** http://www.pentaho.com**

En el context de *big data*, han aparegut empreses especialitzades com Zaloni.*

*** Més informació a: http://www.zaloni.com**

Les solucions actuals per a la qualitat de dades solen incloure capacitats de monitoratge i creació d'*scorecards*.

Figura 8. Plataforma de qualitat de dades



Font: David Cabanillas

4.2.2. Eines d'integració de dades

Dins de les eines d'integració de dades, destaquen les eines ETL, que s'utilitzen per a:

- Extreure dades de fonts de dades homogènies o heterogènies.
- Transformar les dades per emmagatzemar-les en un format o una estructura apropiats per al propòsit de consulta i anàlisi.
- Carregar les dades en la destinació final (base de dades, més específicament, magatzem de dades operatives, *data mart* i magatzem de dades).

Aquestes eines han començat a incloure passos de transformació vinculats a la qualitat de dades, com ara validar el format d'un camp (correu electrònic, compte bancari, etc.) o fins i tot aplicar *data profiling*. Eines com Pentaho Data Integration* o Talend** formen part d'aquesta categoria.

* http://www.pentaho.com
** http://www.talend.com

4.2.3. Eines per a la gestió del programa (de qualitat de dades)

L'aplicació del programa en un àmbit d'organització es pot executar com si es tractés d'un projecte. És a dir, és possible aplicar les eines de gestió de projectes tradicionals per a la implantació del pla de qualitat de dades. Eines com Basecamp,* Trello** o Asana*** poden ser de gran utilitat per coordinar el programa de qualitat de dades.

* http://basecamp.com/
** http://trello.com
*** http://asana.com

4.2.4. Proves de qualitat

Trobem diferents tipus de proves que poden aplicar-se al *data warehouse* (magatzem de dades) i bases de dades quan es volen garantir els processos de l'organització en termes de qualitat total. Algunes de les més interessants són les següents:

1) Proves unitàries: consisteixen a validar cadascun dels components d'una solució, encara que aquest tipus de test ha de dur-se a terme durant l'etapa de desenvolupament, mai després. Els elements més crítics i que s'han de sotmetre a aquest tipus de prova són, almenys, la lògica ETL, regles de negoci i càlculs implementats a la capa d'OLAP (*online analytical processing*) i la lògica d'indicadors clau o KPI (*key performance indicator*). Aquest tipus de proves es fan en diverses ocasions al llarg del curs d'un projecte, i poden automatitzar-se.

2) Proves del sistema d'integració: depenen de l'èxit obtingut en les proves unitàries, i han d'aconseguir dues metes principals:

a) Garantir que es pot construir i desplegar amb èxit: per això és necessari dur a terme proves d'acumulació del sistema.

b) Assegurar que no sorgeixen problemes durant l'execució del treball: amb aquest objectiu, un cop implementats i configurats, tots els treballs han de ser executats i les dades, processades.

L'adopció d'aquest tipus de proves en el cicle de desenvolupament del *data warehouse* i de bases de dades és un pas que serveix per confirmar que el sistema actua de la manera esperada, una vegada que les parts constituents de la solució es conjunten.

3) Proves de validació de dades: mitjançant aquest procés, se sotmeten a prova les dades dins d'un *data warehouse*. Una manera habitual de dur a terme aquesta prova consisteix en l'ús d'una eina de consulta *ad hoc* (per exemple,

Excel) que permeti recuperar dades en un format similar als informes operatius existents. Quan es detecta l'existència d'un vincle entre el *data warehouse* i l'informe operacional, es demostra que les dades són vàlides (llevat que, per descomptat, l'informe original sigui defectuós). Aquesta prova ha de ser duta a terme per un representant del negoci, ja que aquest perfil és el que millor coneix les dades i pot validar amb més garanties d'èxit.

4) Proves d'acceptació d'usuari: el seu objectiu és assegurar que les dades que es proporcionen a l'usuari final compleixen amb les seves expectatives, i que el mateix passa amb les eines que es posen a la seva disposició.

5) Proves de rendiment: s'ocupen de validar adequadament el rendiment de la solució en condicions de treball reals. Per això, en el *testing* cal considerar factors com l'arquitectura de dades, la configuració del *maquinari*, l'escalabilitat del sistema o la complexitat de les consultes.

6) Proves de regressió: aquest tipus de test és el procés de tornar a provar la funcionalitat per garantir que el desenvolupament del *data warehouse* i de les bases de dades no ha causat desperfectes en altres funcions i aplicacions. Cadascuna de les diferents categories de proves definides anteriorment ha de quedar subjecta a proves de regressió.

Resum

En aquest mòdul didàctic, hem presentat el concepte de qualitat de la dada, que té l'objectiu final d'augmentar la confiança en l'ús de la dada per a una presa de decisions eficient, òptima i ràpida en l'organització.

Primer, hem discutit els motius de la baixa qualitat de dades en les organitzacions, per què no s'inverteix en això i la seva necessitat, cosa que ens ha portat a definir el concepte.

A continuació, hem revisat en què consisteix un programa de qualitat de dades i els seus components, des de la maduresa (en quin estat està l'organització respecte a la qualitat de la dada) fins a les diferents metodologies existents (des de la dada al negoci, o a l'inversa). Tot això amb el focus posat en les persones, els processos i les dades.

Per poder desenvolupar el programa, hem discutit les diferents fases que el componen, a més de les millors pràctiques que hem de tenir en compte i la manera de mesurar l'impacte en l'organització.

Finalment, s'han revisat les tècniques i la tecnologia que formen part del que es coneix actualment com qualitat de dades.

Glossari

anàlisi i estandardització *m* i *f* Descomposició dels camps en parts i el format dels valors per incorporar dissenys basats en estàndards industrials, estàndards locals, regles de negoci definides per l'usuari i bases de coneixement de valors i patrons.

big data *m* Conjunt d'estratègies, tecnologies i sistemes per a l'emmagatzematge, el processament, l'anàlisi i la visualització de conjunts de dades complexes.

business intelligence *m* Conjunt de metodologies, aplicacions, pràctiques i capacitats enfocades a la creació i l'administració d'informació, que permet prendre millors decisions als usuaris d'una organització.

cicle de vida d'un actiu *m* Diferents etapes per les quals passa un actiu, des del seu naixement fins al seu final.

data quality *f* Tècniques per a la identificació, el control, l'increment i el manteniment de la qualitat de dades en una organització.

data quality management **(DQM)** *m* La gestió de la qualitat de les dades és un tipus d'administració que incorpora l'establiment de funcions, el seu desplegament, les polítiques, les responsabilitats i els processos pel que fa a l'adquisició, el manteniment, la disposició i la distribució de dades. Perquè una iniciativa de gestió de la qualitat de les dades tingui èxit, es requereix una sòlida associació entre els grups tecnològics i el negoci.

data warehouse *m* Repositori de dades que proporciona una visió global, comuna i integrada de les dades de l'organització, independent de com seran utilitzades posteriorment pels consumidors o usuaris, amb les propietats següents: estable, coherent, fiable i amb informació històrica.

enriquiment *m* Millorar el valor de les dades internes emmagatzemades, afegint-hi atributs relacionats de fonts internes o externes.

ETL *m* Processos que permeten l'extracció, la transformació i la càrrega de dades des de fonts d'origen fins a la destinació per al seu correcte consum.

neteja *f* Modificació de valors de dades per complir amb les restriccions de domini, restriccions d'integritat o altres regles de negoci que defineixen quan la qualitat de les dades és suficient per a l'organització.

matching *m* Identificar, vincular o combinar entrades relacionades dins o entre conjunts de dades.

monitoratge *f* Desplegament de controls que asseguren que les dades segueixen complint amb les regles de negoci que defineixen la qualitat de les dades per a l'organització.

perfilat *m* Captura d'estadístiques (metadades) que proporcionen informació sobre la qualitat de les dades i ajuden a identificar problemes de qualitat.

Programa de qualitat de dades *m* Eina per a l'avaluació de la qualitat de les dades dins d'una organització, i que permet identificar atributs de qualitat de dades, analitzar els atributs de qualitat de dades en el seu context actual o futur i proporcionar una guia per millorar la qualitat de les dades.

Bibliografia

Brackett, M.; Earley, P. S. (2009). *The DAMA Guide to The Data Management Body of Knowledge (DAMA-DMBOK Guide)*. Nova York: DAMA.

Curto, J. (2017). *Introducción al Business Intelligence (nueva edición ampliada y revisada)*. Barcelona: Editorial UOC.

Hoberman, S. (2015). *Data Model Scorecard: Applying the Industry Standard on Data Model Quality*. Nova York: Technics Publications.

Jugulum, R. (2014). *Competing with High Quality Data: Concepts, Tools, and Techniques for Building a Successful Approach to Data Quality*. Nova York: John Wiley & Sons.

Loshin, R. (2010). *The Practitioner's Guide to Data Quality Improvement*. Nova York: Morgan Kaufmann.

Mosley, M. (2009). *DAMA-DMBOK functional framework*. Nova York: DAMA.

McGilvray, D. (2008). *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information (TM)*. Nova York: Elsevier Science.

Sebastian-Coleman, L. (2013). *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework* . Nova York: Morgan Kaufmann.