Introducció al *business intelligence* i *big data*

Generant valor a partir de les dades

Josep Curto

PID_00242519

Universitat Oberta de Catalunya

Índex

Introducció 5

Objectius 7

1. Presa de decisions orientada a la dada 9
1.1. Què és el *business intelligence*? 10
1.2. Diferències entre BI, BA i *big data* 11
1.3. Beneficis 13
1.4. Quan és necessari? 14
1.4.1. Com detectar que no hi ha una estratègia de gestió de dades? 15
1.4.2. *Business Intelligence Maturity Model* 17

2. Gestió de la dada 19
2.1. Què significa gestionar la dada? 19
2.2. Emmagatzematge de la dada en BI 20
2.2.1. *Data Warehousing* 21
2.2.2. Elements en *data warehousing*: fets, dimensions i mètriques 22
2.3. Captura, transformació i gestió de la dada en BI 26
2.3.1. Integració de dades 27
2.3.2. ETL 29

3. Explotació de la dada 32
3.1. Informes 32
3.1.1. Què és un informe 33
3.1.2. Tipus d'informes 33
3.1.3. Elements d'un informe 34
3.1.4. Tipus de mètriques 35
3.1.5. Tipus de gràfics 36
3.1.6. Cicle de vida d'un informe 44
3.2. OLAP 45
3.2.1. OLAP com a eina d'anàlisi 46
3.2.2. Tipus d'OLAP 47
3.3. Quadres de Comandament 49
3.3.1. Procés de creació d'un quadre de comandament 51
3.3.2. *Dashboard* davant *Balanced Scorecard* 52

4. Què és *business analytics*? 56
4.1. Definició de *business analytics* 56
4.2. Tipus de *business analytics* 56

4.3. Beneficis de *business analytics* 59

5. El nou context de negoci 60
5.1. Què ha canviat des del punt de vista de negoci 60
5.2. La naturalesa de la dada 61
5.2.1. Les magnituds físiques de la dada 62
5.2.2. On es troben les dades rellevants per al negoci? 64
5.2.3. Metadades: més enllà del valor de la dada 64
5.3. Les limitacions del *data Warehouse* 65

6. Què és *big data*? 67
6.1. Definició de *big data* 67
6.2. Tipus de *big data* 68
6.2.1. Classificació de NIST 68
6.3. Quan és necessari *big data*? 70
6.3.1. Presa de decisions 70
6.3.2. Operacions i intel·ligència operacional 71
6.3.3. Validació d'hipòtesis i resolució de problemes 71
6.3.4. Productes i serveis de dades 72
6.3.5. Comerç de dades 72

7. Tecnologies de *big data* 74
7.1. Emmagatzematge 76
7.2. Processament 78
7.3. Anàlisi 80
7.4. Visualització 83

Resum 85

Glossari 86

Bibliografia 88

Introducció

En els últims anys les empreses s'han embarcat en un procés de profunda transformació digital dins el marc del que es coneix com la quarta revolució industrial, que està donant pas a una nova manera d'organitzar els mitjans de producció. Les empreses s'estan transformant en "fàbriques intel·ligents" capaces d'una major adaptabilitat a les necessitats i als processos de producció, així com a una assignació més eficaç dels recursos. D'aquesta manera, han obert la via a una nova revolució industrial que s'ha anomenat també transformació digital. No es tracta només de la digitalització dels processos a través de l'automatització, sinó de l'ús de la informació, i les tecnologies de la informació (TI) i les comunicacions amb l'objectiu d'augmentar el valor per al client i l'avantatge competitiu de l'empresa. Per aquest motiu TI ha passat d'estar a la perifèria de l'organització a estar al centre, erigint-se en un dels seus pilars. Aquesta progressiva transformació de base tecnològica s'ha combinat amb altres aspectes, com l'adveniment de les xarxes socials, la democratització d'Internet o el desplegament de l'Internet de les Coses.

Lectura complementària

Schwab, K. (2016). *The Fourth Industrial Revolution*. Davos: World Economic Forum

Internet de les Coses

Internet de les Coses fa referència a la interconnexió digital d'objectes quotidians amb Internet. Ens hi referirem pel seu acrònim en anglès IoT, *Internet of Things*.

El resultat d'aquesta tempesta perfecta en què es troben totes les organitzacions és una explosió de la dada en volum, velocitat i varietat. I de manera natural, ha crescut la complexitat per capturar, processar, emmagatzemar, analitzar i visualitzar les dades.

Com a conseqüència, han aparegut nombrosos mètodes, tècniques i tecnologies que busquen ajudar les organitzacions a prendre millors decisions a partir de les dades i a extreure'n valor. Aquests mètodes, tècniques i tecnologies per a la captura, el processament, l'emmagatzematge, la gestió i l'anàlisi s'han anat estructurant progressivament en diferents estratègies que coneixem com ***business intelligence***, ***business analytics*** i ***big data***.

No obstant això, a mesura que aquestes estratègies s'han fet conegudes, les organitzacions les han anat implementant amb menys fortuna de l'esperada, tal com apunten els estudis d'Aberdeen Group, Dresner Advisory Services o Harvard Business Review. Si les eines han anat madurant al llarg dels últims anys, com és que les organitzacions segueixen tenint tants problemes en la implantació d'aquest tipus de projectes? Hi ha, per tant, encara diverses preguntes per a qualsevol empresa:

Referències bibliogràfiques

Michael Lock (2012). *Managing the TCO of BI: The Path to ROI is Paved with Adoption. Aberdeen Group*.
Howard Dresner (2015). *Wisdom of Crowds Business Intelligence Market Study*. Dresner Advisory Services.
Donald A. Marchand; Joe Peppard (2013). *Why IT Fumbles Analytics*. Harvard Business Review.

- Què és *business intelligence*?
- Què és *business analytics*?
- Què és *big data*?
- Què significa per a la meva organització?
- Quan és rellevant?

- Hi està preparada la meva organització?
- Com desplegar amb èxit aquest tipus d'iniciatives?
- Quines barreres presenten aquest tipus de projectes?
- Quines tecnologies existeixen dins *business intelligence*, *business analytics* i *big data*?

Responent les anteriors preguntes, el present material busca capacitar estudiants i professionals en el context de l'anàlisi de la informació amb l'objectiu de desenvolupar estratègies de negoci que incloguin *business intelligence*, *business analytics* i *big data*, en el si de la seva pròpia organització. I en conseqüència, poder detectar en la pròpia organització casos d'ús i problemàtiques que necessiten aquest tipus d'enfocament.

Objectius

Aquest material didàctic està adreçat a:

- Desenvolupadors i consultors que volen conèixer *business intelligence* i *big data*.
- Desenvolupadors i consultors que volen ajudar en el desenvolupament d'estratègies de negoci que incloguin *business intelligence* i *big data*.
- Gestors que estan interessats en la transformació digital de la seva organització i en la inclusió de *business intelligence* i *big data* com un dels seus pilars fonamentals.
- Estudiants de qualsevol disciplina, especialment els de formació no tecnològica en l'àmbit de l'empresa.

I té els següents objectius:

1. Entendre els conceptes de *business intelligence* i *big data*, les situacions en què cal desplegar una solució d'aquest tipus i els avantatges que proporciona.
2. Promoure'n la necessitat: per què és necessari tenir una estratègia de negoci que inclogui *business intelligence* i *big data*?
3. Presentar i discutir les tecnologies que engloben *business intelligence* i *big data*.
4. Donar a conèixer casos d'ús i exemples.

Si bé l'obra és autocontinguda en la mesura del possible, s'introduiran els conceptes necessaris per al seguiment del material.

1. Presa de decisions orientada a la dada

La gestió d'una organització es fonamenta en prendre decisions adequades pel que fa a clients, productes, empleats, proveïdors i processos de negoci en tots els seus departaments, des de finances fins a *màrqueting*. Per tant, cal tenir mecanismes que donin suport a una presa de decisions eficient.

En els darrers anys, ha emergit una nova forma de competir que es fonamenta en prendre decisions basades en dades i evidències deixant enrere la intuïció. Aquesta forma de competir combina diferents estratègies per generar valor de negoci: ***business intelligence***, ***business analytics*** i ***big data***. No és estrany que els CIOs de les principals empreses del món destaquin per cinquè any consecutiu que la seva principal prioritat tecnològica són aquesta mena d'iniciatives*.

CIO

Quan parlem de CIO, fem referència al *Chief Information Office*, responsable de les tecnologies de la informació en una organització.

***Building the Digital Platform: Insights From the 2016 CIO Agenda Report.* Gartner.**

Així, l'explotació de la informació en el context de les organitzacions ha passat de ser una necessitat més a ser la prioritat de màxima rellevància. L'objectiu és poder prendre millors i més ràpides i informades decisions de negoci. Què significa prendre millors decisions? Considerem el següent exemple.

> En el context actual, conèixer el client és primordial. Es busca comprendre patrons de comportament del client. Interessa, per tant, conèixer aspectes com: quina és la probabilitat que un client pagui cada mes pel servei contractat?, qui va comprar quins productes?, quins productes són els millors (per regió, per canal, etc.)?, quins objectes es tendeixen a comprar junts?, quins altres productes podem recomanar als nostres clients?
>
> Amazon, coneguda empresa d'*e-commerce*, fa recomanacions en temps real on combina l'històric de vendes dels seus clients i les preferències mostrades quan busquem un producte determinat. Aquesta iniciativa fonamentada en la dada permet incrementar els ingressos per client.

Moltes organitzacions encara no han desplegat aquest tipus d'iniciatives i, entre aquelles que ho han fet, no totes han aconseguit assolir l'èxit esperat. Aquest tipus d'iniciatives són complexes ja que suposen una profunda transformació en l'organització. No només es tracta d'implementar un sistema d'informació sinó de canviar la manera com opera una organització en cadascun dels seus departaments. Per això, en aquest material començarem parlant de la intel·ligència de negoci.

El mercat de *business intelligence* existeix des de fa bastants anys i ha evolucionat cap a solucions amb més prestacions, i podem considerar que ha arribat a una significativa maduresa. Destaquem, per exemple, que:

- S'ha produït una consolidació al mercat, mitjançant la compra d'empreses petites per part dels principals agents del mercat, per complementar la seva proposta de valor (en destaquem SAP, IBM, Microsoft o Oracle).

- Han aparegut noves empreses amb focus en la innovació que cobreixen nous nínxols en el mercat de la intel·ligència de negoci, com la visualització, l'anàlisi predictiva, les *virtual appliances*, és a dir, solucions que combinen *maquinari* i *programari*, i/o el *real-time business intelligence* (en destaquem Tableau, QlikView o Yellowfin).

- Les principals solucions BI *codi obert*, és a dir, Pentaho, JasperSoft i Actuate, han estat adquirides per Hitachi Data Systems, Tibco i OpenText respectivament.

Open source

Quan parlem de *codi obert* (o codi lliure), fem referència a programes informàtics el codi dels quals és lliure i està disponible per a tot el món per ser revisat, modificat i/o millorat.

- Hem assistit a l'aparició d'una nova generació de solucions enfocades a la generació de valor de conjunts de dades complexes, cosa que freqüentment es coneix com a *big data*, que expandeix el valor de la intel·ligència de negoci.

- Han començat a consolidar-se les propostes d'intel·ligència de negoci vinculades *Cloud Computing* (en destaquem GoodData, Sisence o Iberinform).

Algunes de les eines en el context de la intel·ligència de negoci acumulen diversos anys de desenvolupament i evolució, i estan recolzades per organitzacions que tenen un clar model de negoci i que generen sinergies entre elles en forma d'ecosistemes. Podem trobar tant eines de bases de dades com de mineria de dades. Tal és la maduresa d'aquestes solucions que és possible desenvolupar i implementar projectes d'intel·ligència de negoci per a tot tipus d'organitzacions, tant pimes com grans organitzacions.

En aquest mòdul, es presenten els conceptes i diferències entre ***business intelligence***, ***business analytics*** i ***big data***, i els seus principals usos a l'empresa.

1.1. Què és el *business intelligence*?

El context de la societat de la informació ha propiciat la necessitat de tenir millors, més ràpids i més eficients mètodes per extreure i transformar les dades d'una organització en informació i distribuir-la al llarg de la cadena de valor.

No obstant això, aquesta necessitat, que actualment es considera crítica en la gran majoria d'empreses, no és nova. A l'octubre de 1958, Hans Peter Luhn, investigador d'IBM, en l'article *A Business Intelligence System*, va encunyar un terme que respon a aquesta problemàtica com l'habilitat de copsar les relacions de fets presentats de manera que guiïn les accions cap a una meta desitjada.

Cadena de valor empresarial

Quan parlem de la cadena de valor empresarial, descrita i popularitzada per Michael E. Porter en la seva obra *Competitive Advantage: Creating and Sustaining superior performance*, fem referència a un model teòric que permet descriure les activitats que generen valor en una organització.

No és fins al 1989 que Howard Dresden, en aquell moment analista de Gartner, proposa una definició formal del concepte.

Conceptes i mètodes per millorar les decisions de negoci mitjançant l'ús de sistemes de suport basats en fets.

Des de llavors, el concepte de què estem parlant ha evolucionat unint sota el seu paraigua diferents tecnologies, metodologies i termes. És, per tant, necessari establir una definició formal d'ús en el present material.

S'entén per *business intelligence* el conjunt de metodologies, aplicacions, pràctiques i capacitats enfocades a la creació i administració d'informació que permet als usuaris d'una organització prendre millors decisions.

En essència, mitjançant la intel·ligència de negoci, podem trencar amb la següent màxima:

> "Allò que no es defineix no es pot mesurar. Allò que no es mesura no es pot millorar. Allò que no es millora, sempre es degrada."
>
> William Thomson

Dins el context de la intel·ligència de negoci s'inclouen múltiples tecnologies. Algunes d'elles són: *data warehouse*, *Reporting*, Anàlisi OLAP (*Online Analytical Processing*), Anàlisi visual, Anàlisi predictiva, Quadre de comandament, Quadre de comandament integral, Mineria de dades, Gestió del rendiment, Previsions, Regles de negoci, *dashboards*, Integració de dades –que inclou ETL (*Extract*, *Transform and Load*)–, etc.

La figura 1 representa els casos d'ús, la composició tradicional d'una plataforma de dades i el paper de la intel·ligència de negoci.

Al llarg d'aquest material entrarem en detall en algunes d'elles per tenir clars els components mínims que ha de tenir aquest tipus de sistemes.

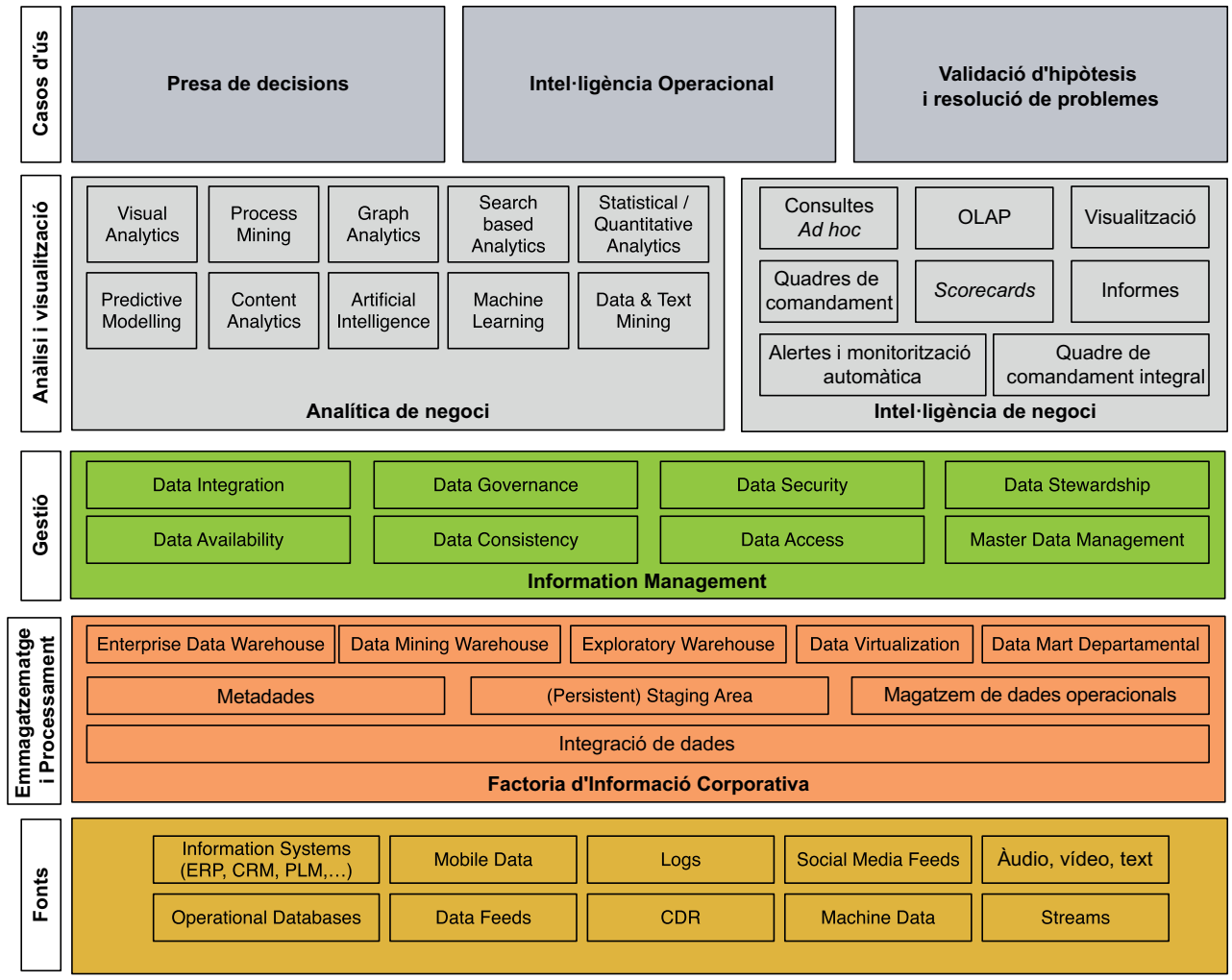
1.2. Diferències entre BI, BA i *big data*

Hem comentat que les organitzacions tenen a la seva disposició diferents estratègies per a l'explotació de la dada. Freqüentment es combinen entre elles per respondre a una necessitat de negoci, però cadascuna d'elles té diferents casos d'ús. Per poder entendre les diferències, ens cal definir-les també.

Primer definim què és *business analytics* (BA) o analítica de negoci.

S'entén per *business analytics* el conjunt d'estratègies, tecnologies i sistemes que permeten analitzar el rendiment passat d'una organització per poder predir comportaments futurs, així com per detectar patrons ocults en la informació.

Figura 1. Plataforma de dades



Font: Josep Curto

Actualment també parlem de ***data science*** com el següent pas a *business analytics*.

Data science

Quan parlem de *data science*, fem referència a un camp multidisciplinari que busca generar coneixement de dades complexes combinant algoritmes, tècniques i coneixements de matemàtiques / estadística, programació i negoci.

Ara definim què és *big data*.

S'entén per *big data* el conjunt d'estratègies, tecnologies i sistemes per a l'emmagatzematge, processament, anàlisi i visualització de conjunts de dades complexes, que freqüentment està definida per volum, velocitat i varietat de la dada.

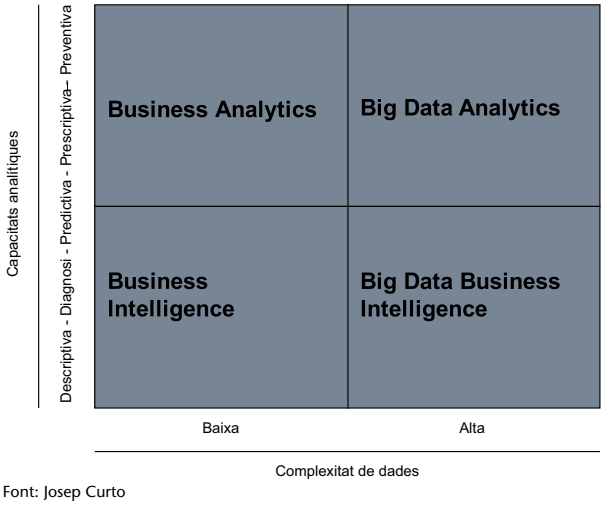
Compararem aquestes estratègies respecte de diferents factors: eines, focus, ús, tipus de la dada, complexitat de la dada i abast. A més, indicarem el seu nivell de maduresa en el mercat. La taula 1 descriu les diferències entre aquestes estratègies.

Taula 1. Diferències entre BI, BA i *big data*

Estratègia	***Business intelligence***	***Business analytics***	***Big data***
Maduresa	Alta	Alta	Emergent
Eines	Consultes, Alertes, Reporting, OLAP, etc.	Classificació, Clustering, Regressió, etc.	Machine learning, Deep Learning, Visualització, etc.
Focus	Què i com va passar, quants, amb quina freqüència, quin és el problema, què cal fer	Per què està passant, què passaria si tot continua igual, què passarà a continuació, què és el millor que pot passar	Capturar, emmagatzemar, processar, analitzar
Ús	Reactiu	Predictiu, proactiu, prescriptiu	Tots els anteriors
Tipus de dada	Estructurat	Estructurat i Semi-estructurat	Tot tipus, principalment no estructurat
Complexitat de la dada	Baixa	Baixa / Mitjana	Alta
Abast	Direcció	Processos	Vertical / Processos

Tot i que queda clar que aquestes estratègies són diferents, el normal és que es combinin en els projectes d'explotació de la dada. La figura 2 permet identificar casos d'ús respecte de la complexitat de la dada i les capacitats analítiques que s'han de desenvolupar en l'organització.

Figura 2. Combinació d'estratègies



Font: Josep Curto

La lectura d'aquest gràfic es realitza a partir dels seus eixos. Per exemple, quan necessitem desenvolupar capacitats analítiques descriptives i la complexitat de dada sigui alta, combinarem *big data* amb *business intelligence* formant així un únic sistema.

Una cadena de supermercats com, per exemple, Carrefour fa servir la intel·ligència de negoci per comprendre el rendiment de cadascun dels seus centres comercials, usa l'analítica de negoci per identificar els principals grups de clients (i les característiques que els defineixen) i usa *big data* per implementar i desplegar un sistema de recomanació de productes per incentivar les compres.

1.3. Beneficis

La implantació d'aquests sistemes d'informació proporciona diversos beneficis entre els quals podem destacar:

- Proporcionar una visió única, conformada, històrica, persistent i de qualitat de tota la informació rellevant per a l'organització.

- Crear, gestionar i mantenir mètriques, indicadors claus de rendiment (*Key Performance Indicador* –KPI–) i indicadors claus de metes (*Key Goal Indicator* –KGI–) fonamentals per a l'empresa.

- Habilitar l'accés a informació actualitzada tant a nivell agregat com detallat.

- Reduir les diferències entre els enfocaments del departament TI i la resta de departaments a través de la implementació del projecte, un cop el departament TI ha comprès les necessitats de negoci.

- Millor comprensió i documentació dels sistemes d'informació en el context d'una organització, un cop han estat identificades les fonts rellevants d'informació per al negoci.

- Millorar la gestió i la competència de l'organització com a resultat de ser capaços de:

– Diferenciar allò rellevant d'allò que és superflu a través de la identificació de les mètriques adequades de negoci.
– Accedir més ràpid a la informació a partir de l'automatització de l'extracció i consolidació de dades.
– Tenir major agilitat en la presa de decisions.

- Crear un cercle virtuós de la informació: les dades es transformen en informació que genera un coneixement, que permet prendre millors decisions, que es tradueix en millors resultats i que genera noves dades.

1.4. Quan és necessari?

Tal com Thomas Davenport en el seu llibre *Competing on Analytics* explica, està emergint una nova forma d'estratègia competitiva basada en l'ús de l'estadística descriptiva, models productius i tècniques complexes d'optimització, dades d'alta qualitat i una presa de decisions basada en fets. En aquest context, la intel·ligència de negoci és el pas previ per a aquesta estratègia atès que ajuda a establir les bases per al seu futur desplegament.

Hi ha situacions en què la implantació d'un sistema de *business intelligence* resulta adequada. Destaquem, entre totes les que existeixen:

- La presa de decisions es realitza de forma intuïtiva en l'organització.
- Identificació de problemes de qualitat d'informació.
- Ús massiu d'Excel com a repositori d'informació corporatiu o d'usuari. El que es coneix com Excel caos.
- Necessitat de creuar informació entre departaments de manera àgil.
- Evitar sitges d'informació.
- Les campanyes de màrqueting no són efectives a causa de la informació base usada.
- Hi ha massa informació en l'organització com per ser analitzada de la manera habitual. S'ha arribat a la massa crítica de dades.
- Cal automatitzar els processos d'extracció i distribució d'informació.

En definitiva, els sistemes de *business intelligence* busquen respondre les preguntes:

- Què va passar?
- Què passa ara?
- Per què va passar?
- Què passarà?

Desplegar un projecte d'intel·ligència de negoci en el si d'una organització no és un procés senzill. Les bones pràctiques indiquen que, per arribar a bon port, cal tenir una estratègia d'intel·ligència de negoci que coordini les tecnologies, l'ús, els processos de maduresa i la metodologia a utilitzar.

1.4.1. Com detectar que no hi ha una estratègia de gestió de dades?

És possible detectar, a través dels següents punts i percepcions, que no hi ha una estratègia definida en el si d'una organització:

- Els usuaris identifiquen el departament d'informàtica com l'origen dels seus problemes d'intel·ligència de negoci.
- La direcció considera que la intel·ligència de negoci és un altre centre de cost.
- El departament d'IT no comprèn les necessitats de negoci i el sistema d'intel·ligència de negoci no ajuda els usuaris de negoci.
- El sistema de BI es considera com un sistema de suport sota *help desk* en lloc d'atendre directament el negoci.
- No es coneix la diferència entre intel·ligència de negoci i la gestió del rendiment.
- No és possible mesurar l'ús del sistema d'intel·ligència de negoci.
- No és possible mesurar el retorn de la inversió (*Return On Invest* –ROI–) del projecte de *business intelligence*.

- Es considera que l'estratègia per al *data warehouse* és la mateixa que per al sistema d'intel·ligència de negoci.
- No hi ha un pla per desenvolupar, contractar, retenir i fer créixer l'equip de BI.
- Els directors de negoci desconeixen si l'empresa té una estratègia per al BI que puguin usar per comprendre el rendiment de les seves unitats de negoci.
- No existeix un responsable funcional, és a dir, un director d'intel·ligència de negoci (o bé, l'assignat no és l'adequat).
- No existeix un centre de competència que permeti definir l'estratègia d'intel·ligència de negoci.
- Hi ha nombroses solucions en l'organització, distribuïdes en diferents departaments, que repeteixen funcionalitat.
- No hi ha un pla de formació real i consistent d'ús de les eines.
- Algú creu que és un èxit que la informació consolidada estigui a disposició dels usuaris finals al cap de dues setmanes.
- Els usuaris creuen que la informació del sistema d'intel·ligència de negoci no és correcta.
- No hi ha una cultura analítica en què la dada i els fets són rellevants per prendre decisions, sigui quin sigui el nivell de l'organització.

El desenvolupament d'una estratègia de negoci és un procés a llarg termini que inclou nombroses activitats, entre les quals convé destacar:

- Parar atenció a les necessitats que requereixen BI en l'organització, perquè s'acostuma a satisfer els usuaris o departaments que criden més fort, cosa que no vol dir que donin més valor a la companyia. Per exemple, els departaments de finances són un cas típic de baixa atenció en solucions BI.
- Identificar quins processos de negoci necessiten diferents aplicacions analítiques que treballin de manera continuada per assegurar que no hi ha sitges de funcionalitat.
- Desenvolupar un *framework* de mètriques a nivell empresarial com el pilar d'una gestió del rendiment a nivell corporatiu.
- Incloure els resultats d'aplicacions analítiques (mineria de dades o altres) en els processos de negoci amb l'objectiu d'afegir valor a tot tipus de decisions.
- Establir els estàndards de BI en l'organització per racionalitzar tant les tecnologies existents com les futures adquisicions.
- Revisar i avaluar el document actual de solucions en un context de risc / recompenses.
- Considerar inversions tàctiques el retorn d'inversió de les quals estigui dins d'un període de temps d'un any. A més, tenir en compte les diferents anàlisis de mercat, de solucions i, fins i tot, el *hype cycle* de Gartner per conèixer l'estat de l'art.
- Aprendre dels èxits i fracassos d'altres empreses revisant casos d'estudi i consultat les empreses del sector per determinar què ha funcionat i què no.

- Crear un centre de competència (o d'excel·lència) de BI (BICC). Té l'objectiu d'unir coneixement en tecnologies, metodologies, estratègia, amb suport a nivell executiu i amb analistes de negoci implicats, i que tingui responsabilitat compartida en èxits i fracassos.
- Alinear el departament IT i el negoci en cas de no poder organitzar un BICC, fonamental per treballar com a equip integrat. El departament d'IT ha d'entendre les necessitats i proposar la millor solució, ajustant-se a la necessitat particular, i escalable a altres de futures.
- Evangelitzar l'organització.

1.4.2. *Business Intelligence Maturity Model*

Si bé l'objectiu d'aquest material no és donar pautes per definir una estratègia de *business intelligence* sinó una introducció de conceptes, un bon punt de partida és identificar quin és el grau de maduresa de l'organització respecte de la intel·ligència de negoci.

El BIMM (*Business Intelligence Maturity Model*) és un model de maduresa que permet classificar la nostra organització des del punt de vista del grau de maduresa d'implantació de sistemes *business intelligence*.

Vegem-ne les fases:

Fase 1: No existeix BI. Les dades es troben en els sistemes d'informació operacionals, com la comptabilitat, la facturació o la nòmina, escampats en altres suports o fins i tot només continguts en el *saber fer* de l'organització. Les decisions es basen en la intuïció, l'experiència, però no en dades consistents. L'ús de dades corporatives en la presa de decisions no ha estat detectat i tampoc l'ús d'una eina adequada al fet.

En aquesta fase, el sistema d'administració comercial (el control de les vendes per venedors, regions, productes, clients, preus, descomptes...) pren les seves decisions basades en el coneixement (experiència i intuïció) de cadascun dels seus comercials.

Fase 2: No existeix BI, però les dades són accessibles. No hi ha un processat formal de les dades per a la presa de decisions, encara que alguns usuaris tenen accés a informació de qualitat i són capaços de justificar decisions amb aquesta informació. Sovint, aquest procés es realitza mitjançant Excel o algun sistema simple per generar informes. S'intueix que hi ha d'haver solucions per millorar aquest procés però es desconeix l'existència del *business intelligence*.

En aquesta fase, el sistema d'administració comercial ha identificat la necessitat d'usar les dades per prendre millors decisions. Alguns comercials prenen les seves decisions basades en dades del sistema de control de vendes i generen informes utilitzant Excel.

Fase 3: Aparició de processos formals de presa de decisions basada en dades. S'estableix un equip que controla les dades i que permet fer informes contra aquells que permeten prendre decisions fonamentades. Les dades són extretes directament dels sistemes transaccionals sense processos de qualitat o preparació per a l'anàlisi automàtica i no hi ha un magatzem únic per a les dades rellevants.

> En aquesta fase, el sistema d'administració comercial ha creat un equip que prepara els informes, tot i que el procés segueix sent manual i suposa moltes hores de treball preparant les dades.

Fase 4: *data warehouse*. L'impacte negatiu contra els sistemes transaccionals porta a la conclusió que un repositori de dades és necessari per a l'organització. Es percep el *data warehouse* com una solució desitjada. El *reporting* segueix sent personal.

> En aquesta fase, el sistema d'administració comercial ha creat un repositori únic de qualitat amb les dades rellevants per a l'anàlisi.

Fase 5: *data warehouse* creix i el *reporting* es formalitza. El *data warehouse* funciona i es desitja que tots se'n beneficiïn, de manera que el *reporting* corporatiu es formalitza. Es parla d'OLAP (anàlisi multidimensional), però només alguns identifiquen realment els seus beneficis.

> En aquesta fase, el sistema d'administració comercial estén el repositori únic a altres àrees i s'automatitzen els informes.

Fase 6: Desplegament d'OLAP. Després de cert temps, ni el *reporting* ni la manera d'accedir al *data warehouse* és satisfactòria per respondre preguntes sofisticades. OLAP es desplega per a aquests perfils. Les decisions comencen a impactar de manera significativa en els processos de negoci en tota l'organització.

> En aquesta fase, el sistema d'administració comercial comença a fer servir l'anàlisi multidimensional (OLAP) per a alguns perfils avançats.

Fase 7: *business intelligence* es formalitza. Apareix la necessitat d'implantar altres processos d'intel·ligència de negoci, com *data mining*, *Balanced ScoreCard*... I els processos de qualitat de dades impacten en processos com *Customer Relationship Management* (CRM), *Supply Chain Management* (SCM)... S'ha establert una cultura corporativa que entén clarament les diferències entre sistemes OLTP i DSS.

> En aquesta fase, el sistema d'administració comercial fa servir un sistema d'intel·ligència de negoci i s'inicia la fase d'implementar altres anàlisis més avançades.

Existeixen altres models de maduresa com el model Delta analític de Thomas Davenport o el de TDWI*.

***https://tdwi.org/pages/maturity-model/analytics-maturity-model-assessment-tool.aspx**

2. Gestió de la dada

En l'apartat anterior, hem introduït el concepte d'intel·ligència de negoci. Aquest concepte fa referència, al mateix temps, tant a un sistema d'informació com a una estratègia de negoci per millorar la presa de decisions. El desplegament d'aquest tipus d'estratègies / sistemes passa per la gestió eficient de la dada.

En aquest apartat revisem què significa gestionar la dada en el context de la intel·ligència de negoci.

2.1. Què significa gestionar la dada?

Dins d'una organització, les persones que han de prendre decisions tenen expectatives que han de ser cobertes. Ens referim al fet que, per prendre una decisió, la dada ha d'estar **disponible** i **accessible**, ser de **qualitat**, **en un moment adequat**, **securitzada** i **transformada en informació**.

Això vol dir:

- **Disponible**: la dada ha estat capturada i emmagatzemada en un dipòsit.
- **Accessible**: hi ha un mecanisme per al consum de dada que habilita el seu accés per part de tercers, tant sistemes com persones.
- **Qualitat**: la dada que s'ha validat té el nivell de qualitat suficient per a la presa de decisions.
- **En un moment adequat**: s'han tingut en compte les necessitats temporals de negoci per a la disponibilitat i accessibilitat de la dada.
- **Securitzada**: la dada està protegida i només poden accedir-hi aquells que tenen permisos.
- **Informació**: la dada s'ha transformat en informació a través de l'anàlisi.

Al capdavall, gestionar la dada significar ser capaç d'emmagatzemar i processar la dada de manera eficient complint les expectatives anteriors perquè respongui a les necessitats actuals d'una organització.

En els següents subapartats discutirem els elements de la intel·ligència de negoci que permeten emmagatzemar i processar la dada de forma eficient.

2.2. Emmagatzematge de la dada en BI

Com ja s'ha comentat, un sistema d'intel·ligència de negoci està format per diferents elements, però de totes les peces, la principal és el *data warehouse* o magatzem de dades. Necessitem definir aquest concepte.

> Un *data warehouse* és un repositori de dades que proporciona una visió global, comuna i integrada de les dades de l'organització, independentment de com seran utilitzades posteriorment pels consumidors o usuaris; amb les propietats següents: estable, coherent, fiable i amb informació històrica.

Donat que abasta un àmbit global de l'organització i amb un ampli abast històric, el volum de dades pot ser molt gran (centenars de terabytes o fins i tot petabytes). Les bases de dades relacionals són el suport tècnic més usat per emmagatzemar les estructures d'aquestes dades i els seus grans volums.

> El *data warehouse* de Wal-Mart desa informació de les transaccions de més de 100 milions de clients i les dades logístiques de més de 25.000 proveïdors. Ja el 1992, aquest magatzem de dades va arribar a tenir més d'1 terabyte d'informació rellevant per comprendre el comportament de clients i optimitzar la relació amb els proveïdors.

Resumint, el *data warehouse* presenta les següents característiques:

- **Orientat a un tema**: organitza una col·lecció d'informació al voltant d'un tema central.
- **Integrat**: inclou dades d'orígens diversos i presenta consistència de dades.
- **Variable en el temps**: es realitzen fotografies de les dades basades en dates o fets.
- **No volàtil**: la informació és persistent i tan sols de lectura per als usuaris finals.

Habitualment, el *data warehouse* està constituït per una base de dades relacional (que desa els seus registres per files), però no és l'única opció factible, també és possible considerar les bases de dades orientades a columnes (que desen les dades per columnes), basades en la lògica associativa (que identifiquen conceptes de negoci utilitzant lògica) o *appliances* especialitzades (optimitzades per al rendiment en l'anàlisi).

2.2.1. *Data Warehousing*

Hem de tenir en compte que en el context d'un *data warehouse* existeixen altres elements que es combinen per poder respondre a les necessitats de negoci:

- ***Data Warehousing***: és el procés d'extreure i filtrar dades de les operacions comunes de l'organització, procedents dels diferents sistemes d'informació operacionals i/o sistemes externs, per transformar-los, integrar-los i emmagatzemar-los en un magatzem de dades, per tal d'accedir-hi i donar suport en el procés de presa de decisions de l'organització.

Mineria de dades

Quan parlem de mineria de dades, o *data mining*, fem referència al camp interdisciplinari amb l'objectiu general de predir resultats i/o descobrir relacions en les dades. Pot ser descriptiu, és a dir, descobrir patrons que descriuen les dades, o predictiu, per pronosticar el comportament del model basat en les dades disponibles.

- ***Data Mart***: és un subconjunt de les dades del *data warehouse* amb l'objectiu de respondre a una determinada anàlisi, funció o necessitat, i amb una població d'usuaris específica. Està pensat per cobrir les necessitats d'un grup de treball o d'un determinat departament dins de l'organització. Per exemple, un possible ús seria per a la mineria de dades o per a la informació de màrqueting.

- ***Operational Data Store* (ODS)**: és un tipus de magatzem de dades que proporciona només els últims valors de les dades i no el seu historial; generalment és a més admissible un petit desfasament o retard en les dades operacionals. També podem tenir-ne d'específics per a la mineria de dades i l'exploració de la dada.

- ***Staging Àrea***: és el sistema que roman entre les fonts de dades operacionals i el *data warehouse* amb l'objectiu de:

– Facilitar l'extracció de dades des de fonts d'origen amb una heterogeneïtat i complexitat gran.
– Millorar la qualitat de dades.
– Ser usat com a cau de dades operacionals amb el qual, posteriorment, es realitza el procés de *data warehousing*.
– Accedir amb detall a informació no continguda en el *data warehouse*.

Memòria cau

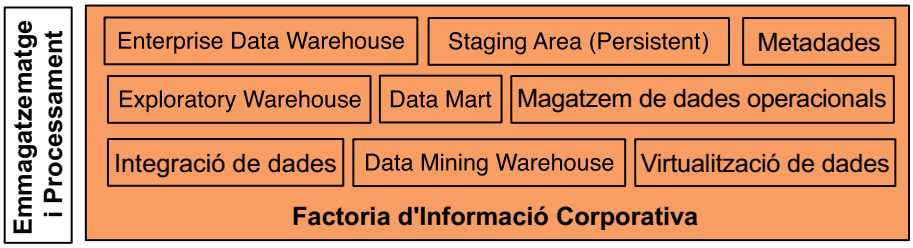
Quan parlem de memòria cau, fem referència al procés d'usar la memòria com a emmagatzematge temporal de dades.

- **Processos ETL**: és una tecnologia d'integració de dades basada en la consolidació de dades; tradicionalment s'usa per alimentar magatzems de dades de qualsevol tipus: *data warehouse*, *data mart*, *staging àrea* i ODS. Usualment es combina amb altres tècniques de consolidació de dades. Aquesta tecnologia permet extreure, transformar i carregar dades.

- **Metadades**: són dades estructurades i codificades que descriuen característiques del procés de *data warehousing* i dels diferents elements que s'han tingut en compte en l'arquitectura del *data warehouse*.

La figura 3 resumeix els diferents components que trobem en el context del *data warehouse*.

Figura 3. Components del *data warehouse*



Font: Josep Curto

2.2.2. Elements en *data warehousing*: fets, dimensions i mètriques

Forma normal

Quan parlem de forma normal, fem referència al procés que consisteix a designar i aplicar una sèrie de regles per al disseny de base de dades amb l'objectiu d'eliminar dades repetides i tenir integritat en les dades.

L'estructura relacional d'una base de dades operacional segueix les formes normals en el seu disseny. En un *data warehouse* no s'ha de seguir aquest patró de disseny. La idea principal és que la informació sigui emmagatzemada de forma desnormalitzada per optimitzar les consultes. Per fer això, hem d'identificar en la nostra organització els processos de negoci, les perspectives d'anàlisi per al procés de negoci i les mesures quantificables associades a tots ells.

És a dir, s'estructura la dada en processos de negoci, vistes d'anàlisi i mesures per comprendre la seva evolució. D'aquesta manera parlarem de:

- **Taula de fet**: és la representació en el *data warehouse* dels processos de negoci de l'organització. Per exemple, una venda pot identificar-se com un procés de negoci de manera que és factible, si correspon a la nostra organització, considerar la taula de fet vendes.

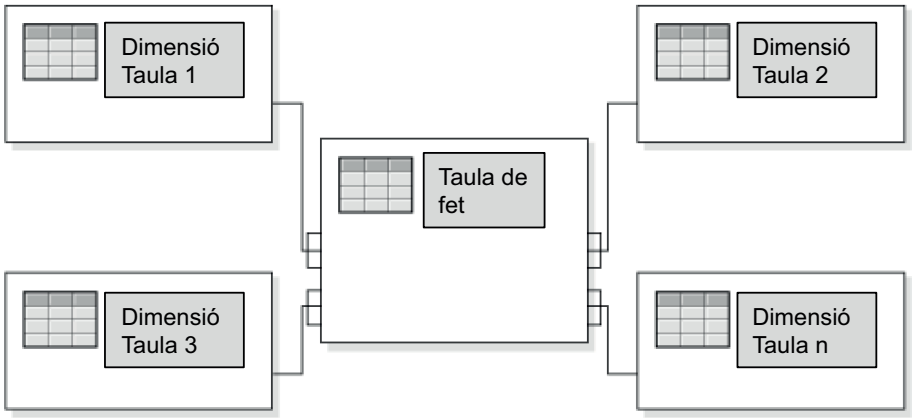
- **Dimensió**: és la representació en el *data warehouse* d'una vista per a un cert procés de negoci. Si tornem a l'exemple d'una venda, per a ella, tenim el client que ha comprat i la data en què s'ha realitzat la compra. Aquests conceptes poden ser considerats com vistes per a aquest procés de negoci. Pot ser interessant recuperar totes les compres realitzades per un client. Això ens fa entendre per què la identifiquem com una dimensió.

- **Mètrica**: són els indicadors d'un procés de negoci. Aquells conceptes quantificables que permeten mesurar el nostre procés de negoci. Per exemple, en una venda, en tenim l'import.

Existeixen principalment dos tipus d'esquemes per estructurar les dades en un magatzem de dades:

- **Esquema en estrella**: consisteix a estructurar la informació en processos, vistes i mètriques amb un esquema que recorda un estel (d'aquí el nom). A nivell de disseny, consisteix en una taula de fets (el que en els llibres trobarem com a *fact table*) en el centre per al fet objecte d'anàlisi, i una o diverses taules de dimensió per cada punt de vista d'anàlisi que participa de la descripció d'aquest fet. En la taula de fet trobem els atributs destinats a mesurar (quantificar): les seves mètriques. La taula de fets només presenta unions amb dimensions. La figura 4 il·lustra aquest esquema.

Figura 4. Esquema en estel



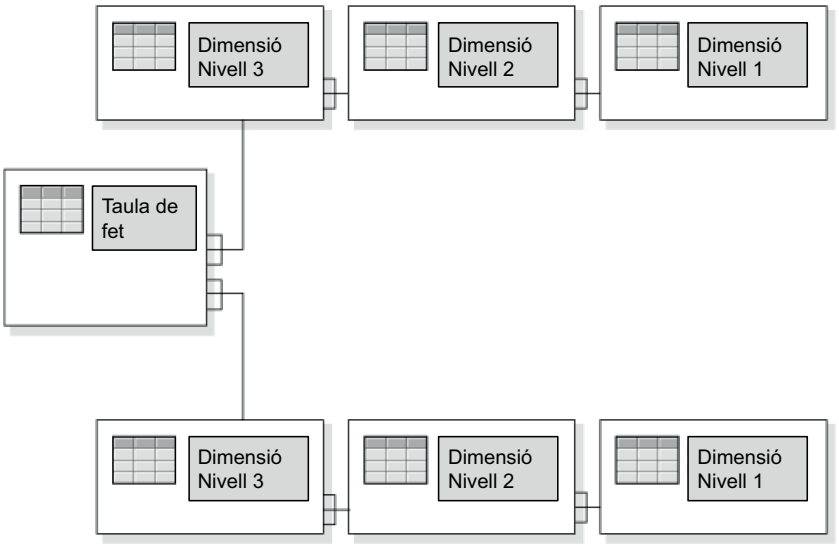
Font: Josep Curto

- **Esquema en floc de neu**: és un esquema de representació derivat de l'esquema en estrella, en què les taules de dimensió es normalitzen en diverses taules. Per aquesta raó, la taula de fets deixa de ser l'única taula de l'esquema que es relaciona amb altres taules i apareixen noves unions. És possible distingir dos tipus d'esquemes en floc de neu:

– **Complet**: en el qual totes les taules de dimensió en l'esquema en estrella apareixen ara normalitzades.
– **Parcial**: només es porta a terme la normalització d'algunes d'elles.

La figura 5 il·lustra aquest esquema.

Figura 5. Esquema en en floc de neu



Font: Josep Curto

És convenient aprofundir en els conceptes de taula de fet, dimensió i mètrica.

Taula de fet

Ja sabem que una taula de fet representa un procés de negoci. A nivell de disseny, és una taula que permet desar dos tipus d'atributs diferenciats:

- Mesures del procés / activitat / flux de treball / esdeveniment que es pretén modelitzar.

- Claus foranes cap a registres en una taula de dimensió (o en altres paraules, com ja sabem, cap a una vista de negoci).

Hi ha diferents **tipus de taules de fet**:

- ***Transaction Fact Table* (taula de fets de transaccions)**: representa esdeveniments que succeeixen en un determinat espai-temps. Es caracteritza perquè permet analitzar les dades amb el màxim detall. Per exemple, podem pensar en una venda que té com a resultat mètriques com ara el seu import.

- ***Factless Fact Tables / Coverage Table* (taules de fets sense mesures)**: són taules que no tenen mesures, cosa que té sentit donat que representen el fet que l'esdeveniment succeeixi. Freqüentment, a aquestes taules s'afegeixen comptadors per facilitar les consultes SQL. Per exemple, podem pensar en l'assistència a un acte benèfic en el qual, per cada persona que hi assisteix, tenim un registre, però podríem no tenir cap altra mètrica associada.

- ***Periodic Snapshot Fact Table* (taules de fets periòdiques)**: són taules de fet usades per recollir informació de manera periòdica, en intervals de temps regulars. Depenent de la situació mesura o de la necessitat de negoci, aquest tipus de taules de fet són una agregació de les anteriors o estan dissenyades específicament. Per exemple, podem pensar en el balanç mensual. Les dades es recullen acumulades de manera mensual.

- ***Accumulating Snapshot Fact Table* (taules de fet agregades)**: representen el cicle de vida complet d'una activitat o procés, que té un principi i un final. Es caracteritzen per presentar diverses dimensions relacionades amb els esdeveniments presents en un procés. Per exemple, podem pensar en un procés de matriculació d'un estudiant i que recopila, durant el seu període de vida, dades que solen substituir-ne d'anteriors (superació i recopilació d'assignatures, per exemple).

Dimensió

Sabem que una dimensió recull principalment els punts d'anàlisi d'un fet. Per exemple, una venda es pot analitzar respecte del dia de venda, producte, client, venedor o canal de venda, entre d'altres.

Hi ha diferents tipus de dimensions:

- ***Slowly Changing Dimensions* (SCD)**: són dimensions que tenen en compte la gestió dels canvis històrics en les dades. En funció de les necessitats de negoci, la dada s'esborra, es modificar o es desa per a la seva comparació.

- **Degenerades**: són dimensions que només tenen un atribut i molt sovint es deixen a la taula de fet. Per exemple, el sexe d'un pacient.

- ***Junk***: Són dimensions que contenen informació volàtil que s'usa puntualment i que no es desa de manera permanent en el *data warehouse*.

- **Conformades**: són dimensions que es fan servir per compartir informació entre taules de fet, cosa que permet fer consultes comunes i creuar informació. L'exemple més fàcil és la dimensió temporal.

- ***Bridge* (pont)**: permet definir relacions entre taules de fet, necessàries per definir, per exemple, la relació entre un pilot i els seus diversos patrocinadors.

- ***Role-playing* (rols)**: que tenen assignat un significat. Per exemple, podem tenir la dimensió data, però també data de lliurament.

- **Alta cardinalitat o *monster***: que contenen una gran quantitat de dades difícilment consultables de manera íntegra. Una bona pràctica és trencar la dimensió en dues taules: una que contingui els valors estàtics i una altra que contingui els valors volàtils. Un exemple clar pot ser la informació del client. Hem de ser conscients de quina és la informació primordial sobre ell que es fa servir puntualment en els informes o altres anàlisis.

Mètriques

També podem distingir diferents tipus de mesures, basades en el tipus d'informació que recopilen, així com la seva funcionalitat associada:

- **Mètriques**: valors que recullen el procés d'una activitat o els seus resultats. Aquestes mesures procedeixen del resultat de l'activitat de negoci.

– **Mètriques de realització d'activitat (*leading*)**: mesuren la realització d'una activitat. Per exemple, la participació d'una persona en un esdeveniment.

– **Mètriques de resultat d'una activitat (*lagging*)**: recullen els resultats d'una activitat. Per exemple, la quantitat de punts d'un jugador en un partit.

- **Indicadors clau**: entenem per aquest concepte valors corresponents que cal assolir i que suposen el grau d'assumpció dels objectius. Aquestes mesures proporcionen informació sobre el rendiment d'una activitat o sobre la consecució d'una meta.

– ***Key Performance Indicator* (KPI)**: Indicadors clau de rendiment. Més enllà de l'eficàcia, es defineixen uns valors que ens expliquen en quin rang òptim de rendiment ens hauríem de situar a l'hora d'aconseguir els objectius. Són mètriques del procés.

– ***Key Goal Indicator* (KGI)**: Indicadors de metes. Defineixen mesures per informar la direcció general si un procés TIC ha assolit els seus requisits de negoci i s'expressen, en general, en termes de criteris d'informació.

Hem d'afegir que existeixen també indicadors d'acompliment. Els indicadors clau d'acompliment (en definitiva, són KPI) defineixen mesures que determinen com de bé s'està exercint el procés de TI per arribar a la meta. Són els principals indicadors que assenyalen si serà factible aconseguir un objectiu o no, i són bons indicadors de les capacitats, les pràctiques i les habilitats. Els indicadors de metes de baix nivell es converteixen en indicadors d'acompliment per als nivells alts.

2.3. Captura, transformació i gestió de la dada en BI

En l'anterior subapartat hem discutit com s'emmagatzemen les dades en un sistema d'intel·ligència de negoci. És important recalcar que en la intel·ligència de negoci s'estructura per endavant el format d'anàlisi.

Després de crear un model que representa els processos de negoci rellevants per a l'organització i les diferents perspectives d'anàlisi, i d'haver-lo implementat en el *data warehouse*, el pas següent és la càrrega de les dades. No només es tracta de carregar dades en el repositori sinó també de preocupar-se per altres aspectes com quina és la millor manera de capturar la dada, quin tipus de transformacions són necessàries per transformar les dades en informació i quines accions són necessàries per gestionar de manera eficient la dada.

Per capturar, transformar i gestionar la dada de manera eficient, s'usa la integració de dades, que proporciona una visió única de totes les dades de negoci es trobin on es trobin.

Aquest subapartat es centrarà en la integració de dades en general i en els processos ETL (Extracció, Transformació i Càrrega) en particular, que és una de les tecnologies d'integració de dades que es fa servir en els projectes d'implantació de *business intelligence*. L'objectiu d'aquest apartat és conèixer les diferents opcions d'integració de dades en l'àmbit de la intel·ligència de negoci i, en particular, conèixer el disseny de processos ETL.

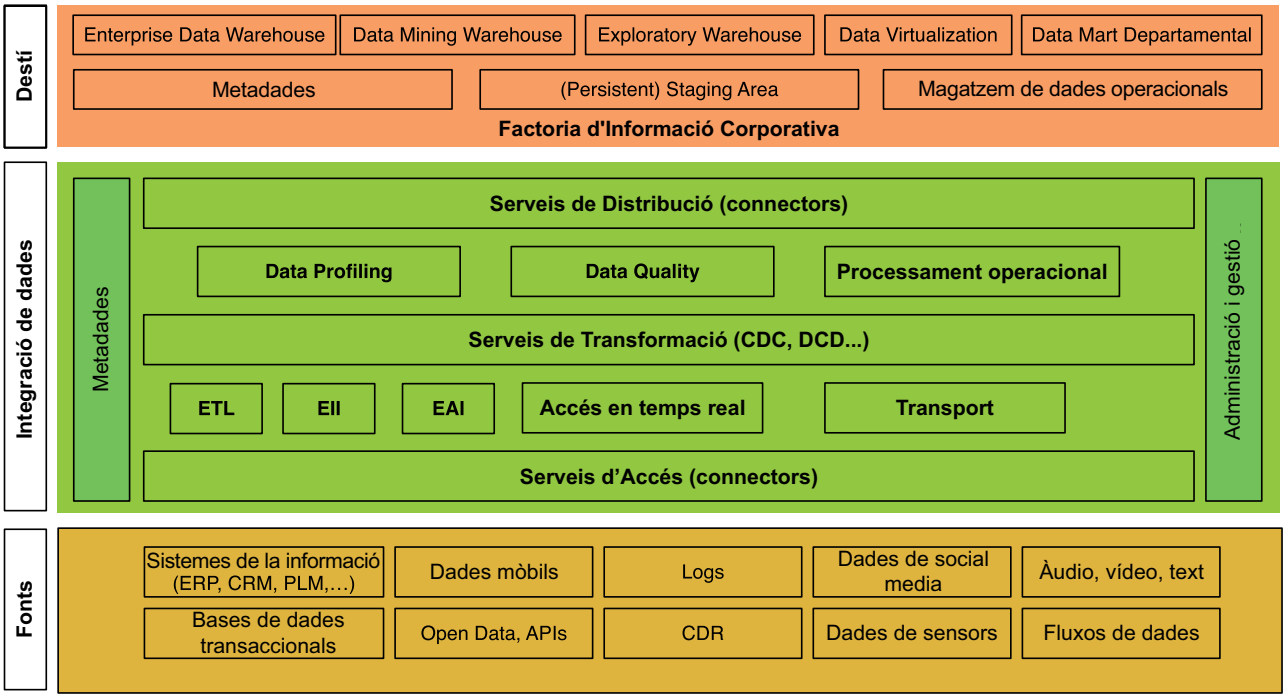
2.3.1. Integració de dades

La integració de dades inclou diversos components destinats a la gestió eficient de la dada com poden ser:

- Serveis d'accés / lliurament de dades (via adaptadors / connectors)
- Gestió de serveis
- *Data profiling* o perfilat de dades, que permet conèixer i analitzar les dades per conèixer la seva estructura, contingut, relacions i regles que puguin derivar-se de les dades
- *Data Quality* o qualitat de dades, que permet conèixer, analitzar i gestionar el nivell de qualitat d'un conjunt de dades
- Processos operacionals
- Serveis de transformació: CDC, SCD, validació, agregació
- Serveis d'accés en temps real
- *Extract, Transform and Load* (ETL)
- *Enterprise Information Integration* (MII)
- *Enterprise Application Integration* (EAI)
- Capa de transport de dades
- Gestió de metadades

La figura 6 il·lustra els components d'integració dins el context de la intel·ligència de negoci.

Figura 6. Components Integració de dades



Font: Josep Curto

Les eines d'integració de dades estan evolucionant per incloure més prestacions de curació i qualitat de dades basades en algoritmes analítics automàtics o semi-automàtics. Per exemple, en entorns internacionals, on el salari d'un empleat pot denominar-se de manera diferent (*wages vs. salary*) i fins i tot incloure prestacions diferents, l'eina d'integració de dades pot reconèixer que són el mateix concepte de negoci.

Curació de dades

Quan parlem de curació de dades, fem referència al procés que mira d'assegurar que les dades siguin fiables i recuperables per a finalitats futures d'investigació o reutilització.

El punt de partida adequat és definir formalment el concepte d'integració de dades.

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

Pel que fa a la definició:

- Les aplicacions són solucions fetes a mida que permeten la integració de dades en base a l'ús de productes d'integració.
- Els productes comercials desenvolupats per tercers capaciten la integració mitjançant l'ús de tecnologies d'integració.
- Les tecnologies d'integració són solucions per realitzar la integració de dades.

La integració de dades juga un paper cada vegada més fonamental en les organitzacions ja que actualment la dada pot residir en plataformes internes (sistemes operacionals i decisionals, *cloud* privat) i externes (*Internet of Things*, *cloud* públic, dispositius mòbils), i cal poder capturar i distribuir la dada als sistemes d'anàlisi necessaris.

Cloud computing

Quan parlem de *cloud computing* (o computació en núvol), fem referència a un terme general per a la prestació de serveis allotjats a través d'Internet. Té diferents modalitats: privada (proporcionada per una organització), pública (proporcionada per tercers) o híbrida (combinació de les dues anteriors).

Hi ha diferents tècniques d'integració de dades:

- **Propagació de dades**: consisteix a copiar dades d'un lloc d'origen a un entorn destinació local o remot. Les dades es poden extreure de l'origen mitjançant programes que generin un fitxer que ha de ser transportat a la destinació, on s'utilitzarà com a fitxer d'entrada per carregar a la base de dades de destinació. Una aproximació més eficient és descarregar només les dades que han canviat en origen respecte de l'última propagació realitzada, generant un fitxer de càrrega incremental que també serà transportat a la destinació.
- **Consolidació de dades**: consisteix a capturar els canvis realitzats en diversos entorns origen i propagar-los a un únic entorn destí, on s'emmagatze-

ma una còpia de totes aquestes dades. En són exemples un *data warehouse* o ODS, alimentat per diversos entorns de producció.

- **Federació de dades**: proporciona a les aplicacions una visió lògica virtual comuna d'una o més bases de dades. Aquesta tècnica permet accedir a diferents entorns d'origen de dades, que poden estar en els mateixos o en diferents gestors de dades i màquines, i crear una visió d'aquest conjunt de bases de dades com si fos, a la pràctica, una base de dades única i integrada.

- **CDC (*Change Data Capture*)**: s'utilitzen per capturar els canvis produïts per les aplicacions operacionals en les bases de dades d'origen, de tal manera que poden ser emmagatzemades i/o propagades als entorns destí perquè mantinguin la consistència amb els entorns origen. Hi ha quatre tècniques principals: CDC per aplicació (l'aplicació genera l'actualització), CDC per *timestamp* (la base de dades genera el canvi basat en dates), CDC per *triggers* (la base de dades genera el canvi en funció d'accions d'actualització de les dades que conté) i CDC per captura de LOG (consisteix a monitoritzar els canvis a nivell del *log* de registres de canvis de l'aplicació).

- **Tècniques híbrides**: la tècnica triada a la pràctica per a la integració de dades dependrà dels requisits de negoci per a la integració, però també, en gran mesura, dels requisits tecnològics i de les probables restriccions pressupostàries. De fet, se solen emprar diverses tècniques d'integració de manera que es constitueix el que s'anomena una tècnica híbrida.

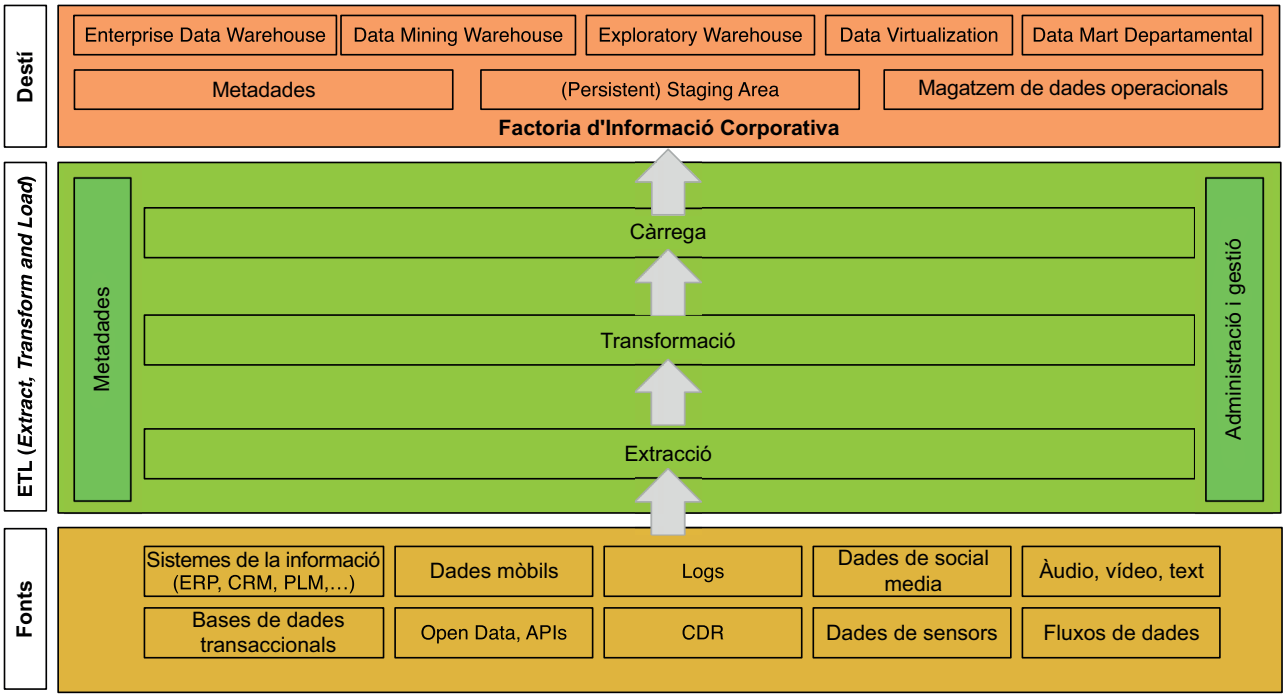
2.3.2. ETL

En el context de la intel·ligència de negoci, dins de totes les tècniques d'integració de dades, les eines ETL han estat l'opció usual per alimentar el *data warehouse*. La funcionalitat bàsica d'aquestes eines està composta per:

- Gestió i administració de serveis
- Extracció de dades
- Transformació de dades
- Càrrega de dades
- Gestió de dades

La figura 7 il·lustra els components d'ETL dins el context de la intel·ligència de negoci.

Figura 7. Components ETL



Font: Josep Curto

ETL és una tecnologia que permet extreure dades de l'entorn origen, transformar-les segons les nostres necessitats de negoci per a la integració de dades i carregar-les en els entorns destinació. Els entorns origen i destinació són generalment bases de dades i/o fitxers, però en ocasions poden ser també cues de missatges d'un determinat *middleware*, juntament amb els fitxers o altres fonts estructurades, semiestructurades o no estructurades. Està basada en tècniques de consolidació.

Les eines d'ETL, a la pràctica, mouen o transporten dades entre entorns origen i destí, però també documenten com aquestes dades són transformades (si ho són) entre l'origen i el destí, emmagatzemant aquesta informació en un catàleg propi de metadades; intercanvien aquestes metadades amb altres aplicacions que puguin necessitar-les i administren totes les execucions i processos de l'ETL: planificació del transport de dades, *log* d'errors, *log* de canvis i estadístiques associades als processos de moviment de dades. Aquest tipus d'eines solen tenir una interfície gràfica d'usuari, i permeten dissenyar i administrar i controlar cadascun dels processos de l'entorn ETL.

Com funciona un procés ETL?

- S'identifiquen els conjunts de dades a extreure.
- S'accedeix a les fonts de dades d'origen i es recuperen.
- Es realitzen les transformacions necessàries en les dades (per exemple, canviar el format de data).
- Es carrega la dada transformada en la dada de destinació (per exemple, el *data warehouse*).

N'hi ha de diferents tipus:

- ETL de generació de codi (mitjançant un llenguatge de programació)
- ETL basat en una eina especialitzada
- ETL integrat a la base de dades

3. Explotació de la dada

Hem vist que per generar valor en el context de la intel·ligència de negoci és necessari emmagatzemar la dada de forma adequada, prenent en consideració només la informació que és rellevant. El següent pas és explotar la dada emmagatzemat. Cosa que, en definitiva, ens permeti prendre decisions i dur a terme accions informades.

En aquest apartat revisem els següents enfocaments per a l'explotació de la dada: informes, OLAP i quadres de comandament.

3.1. Informes

El punt d'entrada tradicional per a una eina d'intel·ligència de negoci en el context d'una organització és la necessitat d'informes operacionals.

Al llarg de la vida d'una empresa, la quantitat de dades que es generen per la seva activitat de negoci creix de forma exponencial, i aquesta informació es desa tant en les bases de dades de les aplicacions de negoci com en fitxers en diversos formats.

Cal generar i distribuir informes per conèixer l'estat del negoci i poder prendre decisions a tots els nivells: operatiu, tàctic i estratègic.

El primer enfocament és modificar les aplicacions de negoci perquè puguin generar els informes. Sovint, l'impacte en les aplicacions és considerable, ja que afecta el rendiment tant dels informes com de les operacions que suporta l'aplicació.

En el moment en què es busca una solució que permeti generar informes sense impactar en el rendiment de les aplicacions de negoci, és quan es considera el *data warehouse*.

Cal comentar que:

- Les eines d'informes existeixen des de fa molt temps i, per això mateix, són solucions madures que permeten cobrir les necessitats dels usuaris finals respecte dels informes.

- Cada fabricant suporta la creació de tot tipus d'informes; en funció de l'enfocament, la dependència dels usuaris finals respecte al departament IT pot ser diferent.

- Les fonts d'origen dels informes són diverses, des del propi *data warehouse*, OLAP, metadades o ODS.

3.1.1. Què és un informe

Les eines d'informes (o també anomenades de *reporting*) permeten respondre principalment la pregunta **Què va passar?** Atès que aquesta és la primera pregunta que es formulen els usuaris de negoci, la gran majoria de les solucions de *business intelligence* del mercat inclouen un motor de generació d'informes. Definim primer què és un informe.

> Un informe és un document a través del qual es presenten els resultats d'un o diversos processos de negoci. Sol contenir text acompanyat d'elements com taules o gràfics per agilitzar la compressió de la informació presentada.

Els informes estan destinats a usuaris de negoci que tenen la necessitat de conèixer la informació consolidada i agregada per a la presa de decisions.

> Imaginem una empresa que té diverses botigues distribuïdes en una ciutat. El director d'aquesta cadena necessita conèixer el rendiment de cadascuna de les botigues per poder gestionar-les de manera eficient. És per això que necessita un informe on es presentin i analitzin els resultats tant a nivell agregat (per conèixer-ne la rellevància d'una botiga respecte del total) com a nivell de botiga (per conèixer els detalls). És a dir, estem parlant de vendes, costos, clients, productes venuts, productes al magatzem, personal...

Ara podem definir formalment les eines de *reporting*.

> S'entén per plataforma de *reporting* aquelles solucions que permeten dissenyar i gestionar (distribuir, planificar i administrar) informes en el context d'una organització o en una de les seves àrees.

3.1.2. Tipus d'informes

Hi ha diferents tipus d'informes en funció de la interacció oferta a l'usuari final i la independència respecte del departament TI:

- **Estàtics**: tenen un format preestablert inamovible.

- **Paramètrics**: presenten paràmetres d'entrada i permeten múltiples consultes.

- ***Ad hoc***: són creats per l'usuari final a partir de la capa de metadades que permet usar el llenguatge de negoci propi. Aquesta capa és molt rellevant perquè oculta el llenguatge tècnic als usuaris de negoci i a més, permet augmentar el valor de l'informe afegint-hi nova informació (com regles de negoci o noves mètriques).

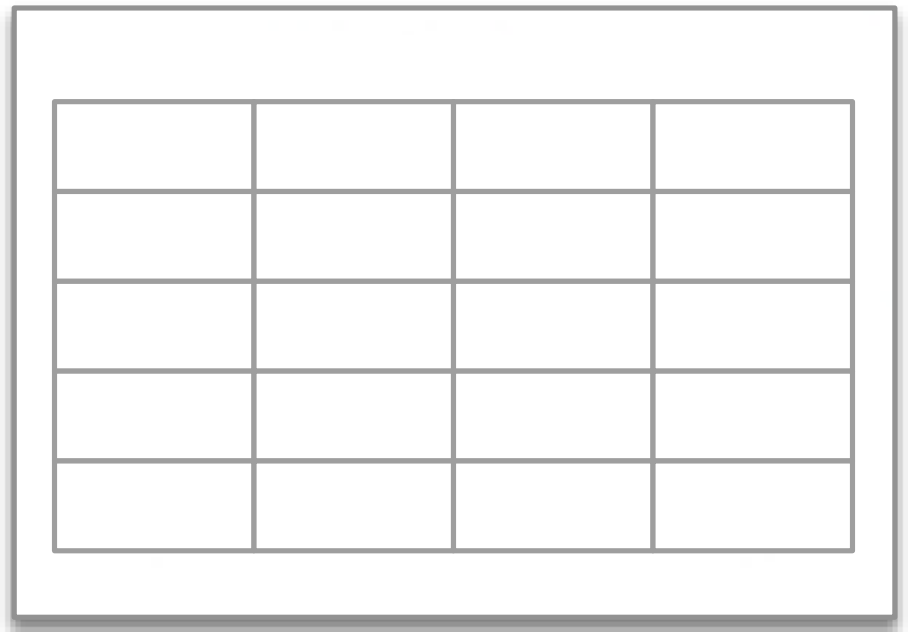
3.1.3. Elements d'un informe

Principalment un informe pot estar format per diversos elements que representen taules de fet, dimensions i mètriques:

- **Text**: descriu l'estat del procés de negoci, proporciona les descripcions necessàries per entendre la resta d'elements de l'informe, així com etiquetes (títol) i/o metadades (data d'execució, fórmules de cálcul...).

- **Taules**: aquest element té forma de matriu (files i columnes) i permet presentar una gran quantitat d'informació com il·lustra la figura 8.

Figura 8. Taules



Font: Josep Curto

- **Gràfics**: aquest element persegueix l'objectiu de mostrar informació amb un alt impacte visual que serveixi per obtenir informació agregada amb molta més rapidesa que a través de taules.

- **Mapes**: aquest element permet mostrar informació geolocalitzada com es mostra a la figura 9.

Figura 9. Mapa

Font: Josep Curto

- **Mètriques**: aquest element permet conèixer quantitativament l'estat d'un procés de negoci. Són paràmetres que es mesuren com, per exemple, les vendes d'un condicionador de cabell per regió.

- **Alertes visuals i automàtiques**: permeten definir avisos automàtics dels canvis d'estat d'un procés de negoci. Aquestes alertes estan formades per elements gràfics com dates, icones o colors ressaltats, i han d'estar automatitzades en funció de regles de negoci encapsulades en l'informe.

Entrarem en detall en algun d'aquests elements.

3.1.4. Tipus de mètriques

Els informes inclouen mètriques de negoci. Per això és necessari definir els diferents tipus de mesures existents basades en el tipus d'informació que recopilen així com la funcionalitat associada a elles:

- **Mètriques**: valors que recullen el procés d'una activitat o els seus resultats. Aquestes mesures procedeixen del resultat de l'activitat de negoci.

– **Mètriques de realització d'activitat (*leading*)**: mesuren la realització d'una activitat. Per exemple, la participació d'una persona en un esdeveniment.

– **Mètriques de resultat d'una activitat (*lagging*)**: recullen els resultats d'una activitat. Per exemple, la quantitat de punts d'un jugador en un partit.

- **Indicadors clau**: entenem per aquest concepte valors corresponents que cal assolir i que suposen el grau d'assumpció dels objectius. Aquestes mesures proporcionen informació sobre el rendiment d'una activitat o sobre la consecució d'una meta.

– ***Key Performance Indicator* (KPI)**: Indicadors clau de rendiment. Més enllà de l'eficàcia, es defineixen uns valors que ens expliquen en quin rang òptim de rendiment ens hauríem de situar a l'hora d'aconseguir els objectius. Són mètriques del procés. Per exemple, la ràtio de creixement d'altes en un servei.

– ***Key Goal Indicator* (KGI)**: Indicadors de metes. Defineixen mesures per informar la direcció general si un procés TIC ha assolit els seus requisits de negoci. En general, s'expressen en termes de criteris d'informació. Si considerem el KPI anterior, seria marcar un valor objectiu de creixement del servei que es pretén aconseguir, per exemple, un 2%.

3.1.5. Tipus de gràfics

En el procés de confecció d'un informe, un dels punts més complicats és la selecció del tipus de gràfic. Hem de començar primer per la definició formal del concepte.

S'entén per gràfic la representació visual d'una sèrie de dades.

El gràfic pot ser una eina eficaç ja que:

- Permet presentar la informació de forma clara, senzilla i precisa.
- Facilita la comparació de dades i habilita destacar tendències i diferències.

L'ús del gràfic dependrà del tipus de dada i el podem classificar en:

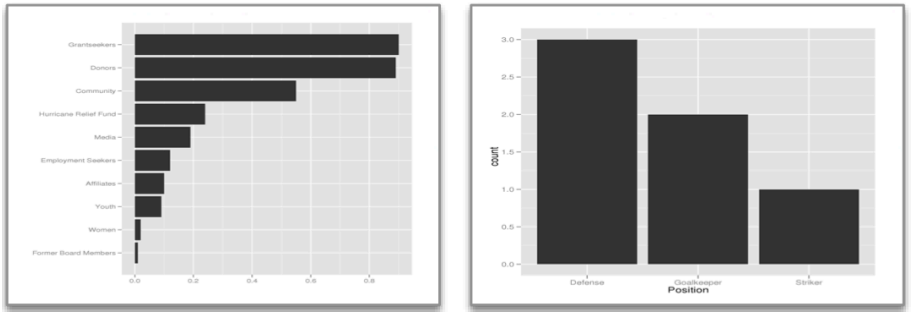
- **Qualitatiu**: es refereix a qualitats o modalitats que no es poden expressar numèricament. Poden ser ordinals (segueixen un ordre) o categòric (sense ordre).

- **Quantitatiu**: es refereix a quantitats o valors numèrics. Poden ser discrets (prenen valors sencers) o continus (prenen qualsevol valor en un interval).

Revisem ara alguns dels tipus de gràfics més rellevants:

- **Gràfics de barres**: és una representació gràfica en un eix cartesià de les freqüències d'una variable qualitativa o discreta. L'orientació pot ser vertical o horitzontal. Es poden classificar en senzill (representa una única sèrie de dades), agrupat (conté diverses sèries de dades) o apilat (es divideix en segments de diferents colors o textures, i cadascun d'ells en representa una sèrie), com es mostra a la figura 10.

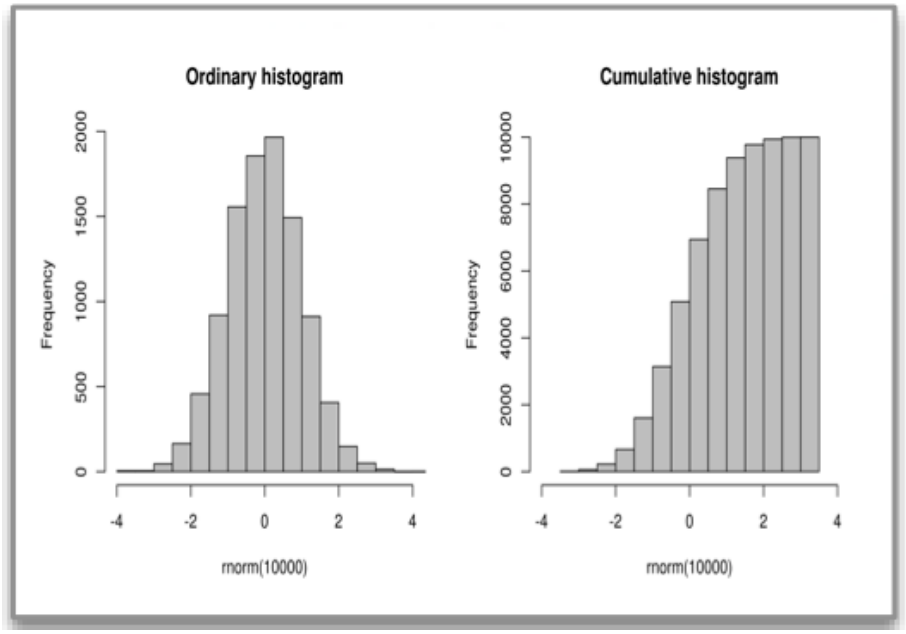
Figura 10. Gràfic de barres



Font: Josep Curto

- **Histograma**: s'usa per representar les freqüències d'una variable quantitativa contínua. En un dels eixos es posicionen les classes de la variable contínua (els intervals o les marques de classe, que són els punts mitjans de cada interval) i en l'altre eix, les freqüències. Existeixen també els histogrames bi-direccionals que contenen dues sèries de dades les barres de freqüències de les quals creixen en sentits oposats, com es mostra a la figura 11.

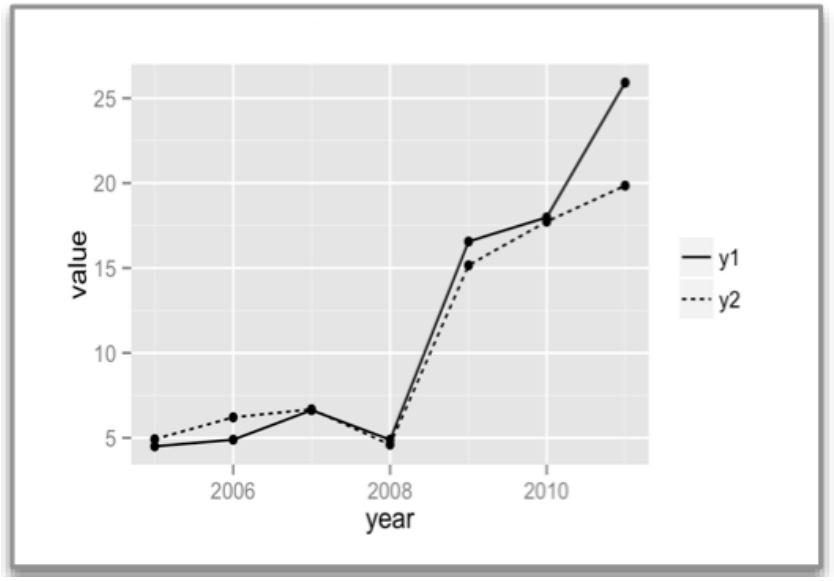
Figura 11. Histograma



Font: Josep Curto

- **Gràfic de línies**: és una representació gràfica en un eix cartesià de la relació que hi ha entre dues variables. Se sol utilitzar per presentar tendències temporals com es mostra a la figura 12.

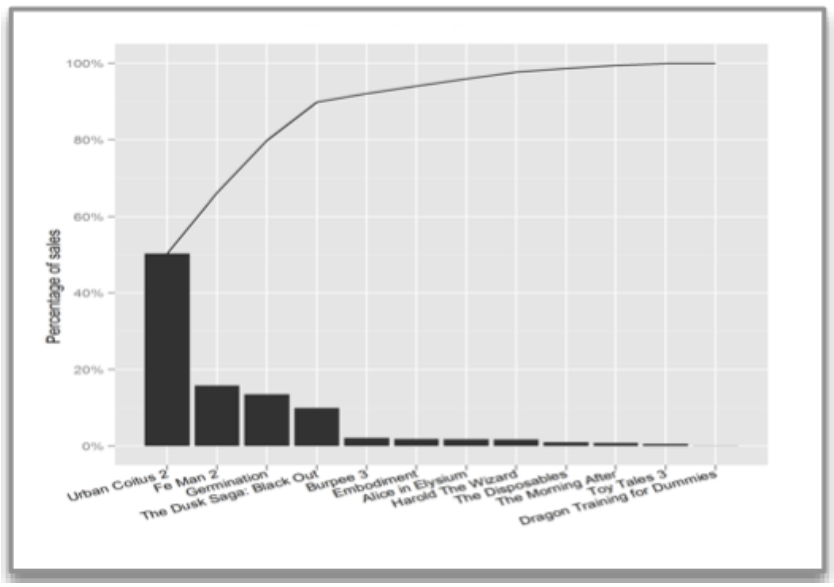
Figura 12. Gràfic de línies



Font: Josep Curto

- **Gràfic de Pareto**: és un tipus de gràfic de barres verticals ordenat per freqüències de manera descendent, que identifica i dóna un ordre de prioritat a les dades com es mostra a la figura 13.

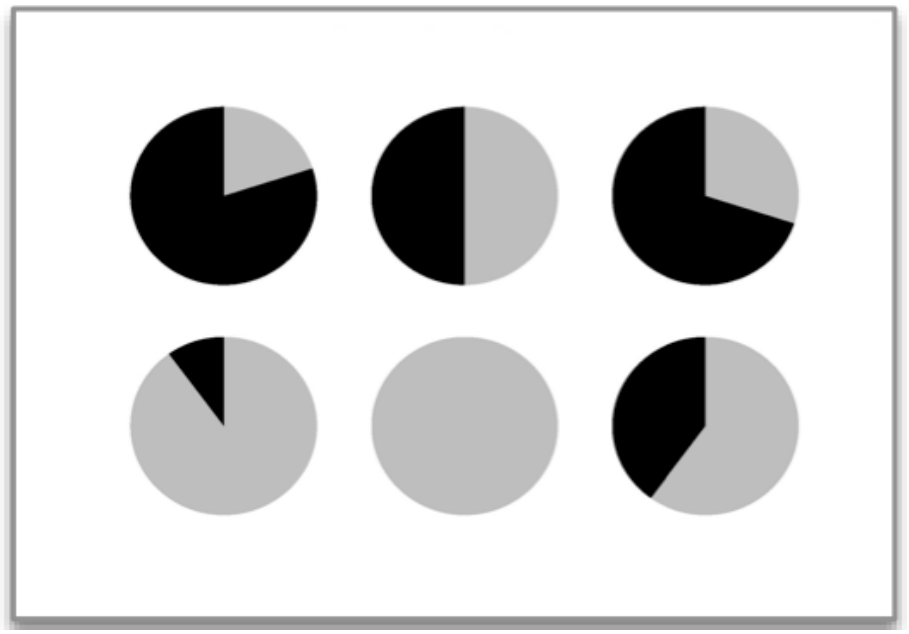
Figura 13. Gràfic de pareto



Font: Josep Curto

- **Gràfic de sectors**: és una representació circular de les freqüències relatives d'una variable qualitativa o discreta que permet, d'una manera senzilla i ràpida, la seva comparació com es mostra a la figura 14.

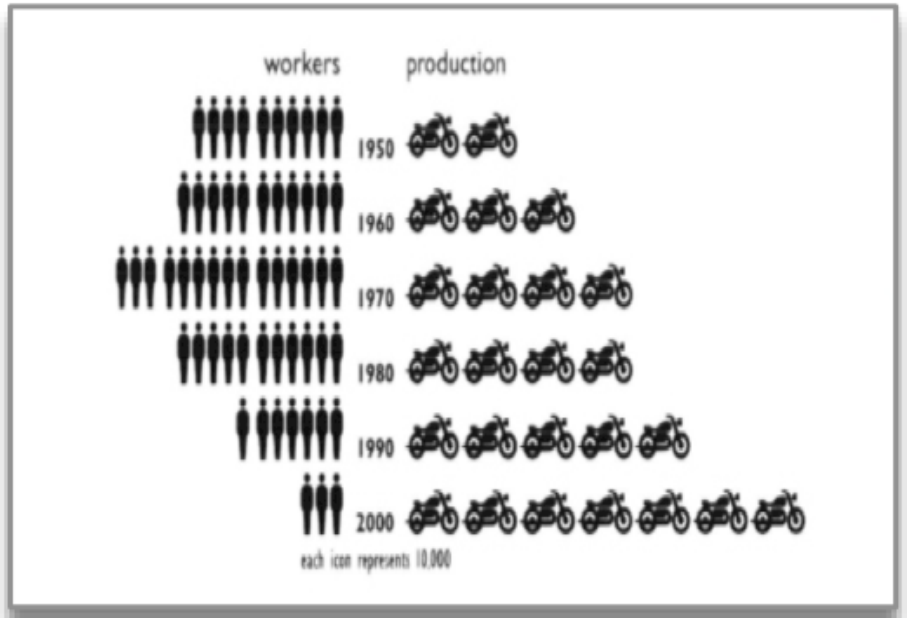
Figura 14. Gràfic de sectors



Font: Josep Curto

- **Pictograma**: és un gràfic que representa, mitjançant figures o símbols, les freqüències d'una variable qualitativa o discreta com es mostra a la figura 15.

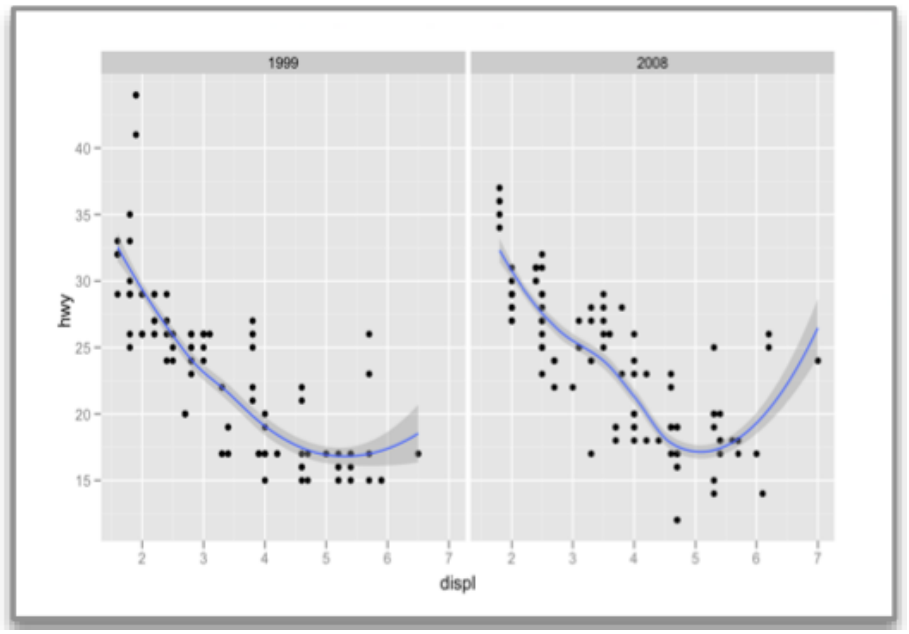
Figura 15. Pictograma



Font: Josep Curto

- **Gràfic de dispersió**: mostra en un eix cartesià la relació que hi ha entre dues variables i informa del grau de correlació entre elles. El tipus de correlació es pot deduir segons la forma del núvol de punts, a saber: nul·la, lineal o no lineal, com es mostra a la figura 16.

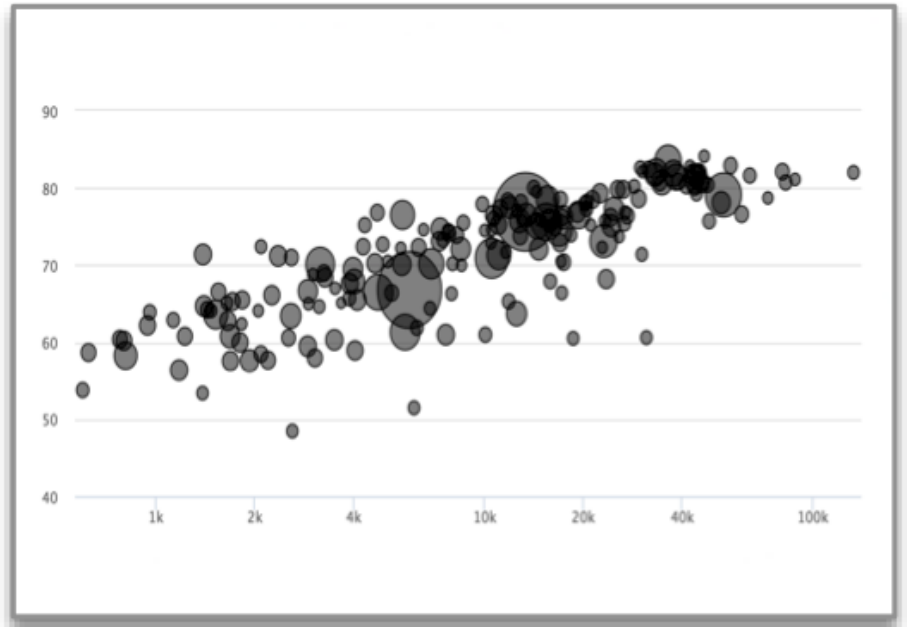
Figura 16. Gràfic de dispersió



Font: Josep Curto

- **Gràfic de bombolles**: és una variant del gràfic de dispersió al qual s'afegeix una tercera dimensió vinculada a la grandària dels punts (que es converteixen en bombolles), i fins i tot se'n pot afegir una quarta vinculada al color de cada bombolla. Per tant, permet estudiar la relació de tres variables com s'il·lustra a la figura 17.

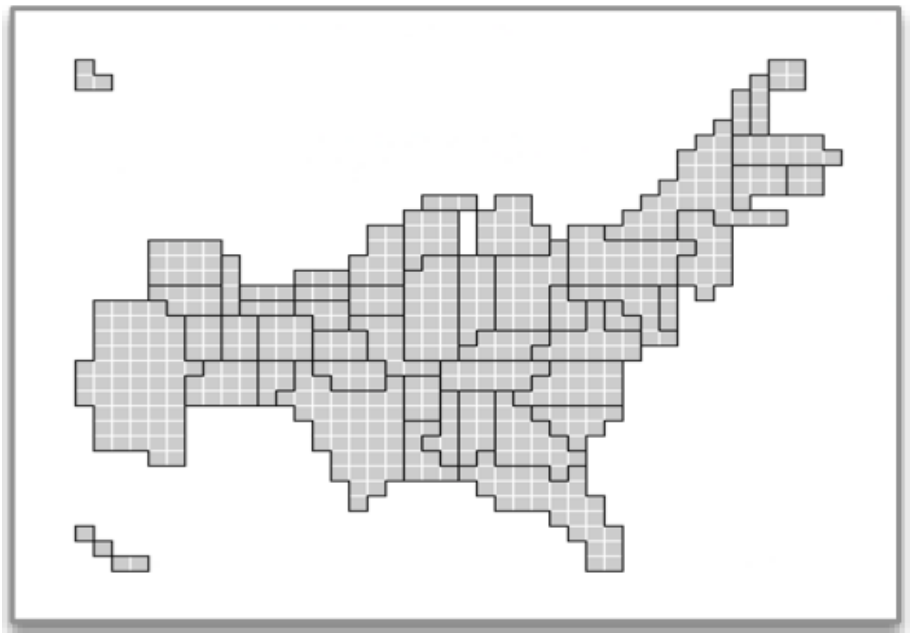
Figura 17. Gràfic de bombolla



Font: Josep Curto

- **Cartograma**: és un mapa en el qual es presenten dades per regions bé posant el número, o bé acolorint les diferents zones, en funció de la dada que representen com es mostra a la figura 18.

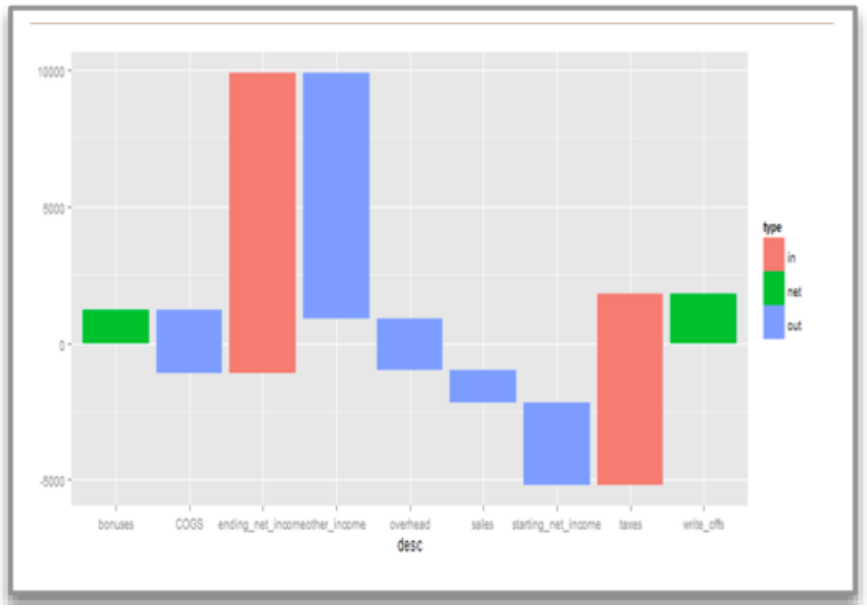
Figura 18. Cartograma



Font: Josep Curto

- **Gràfics en cascada**: és un tipus de gràfic normalment usat per comprendre com un valor inicial es veu afectat per una sèrie de canvis intermedis positius i negatius com es mostra a la figura 19.

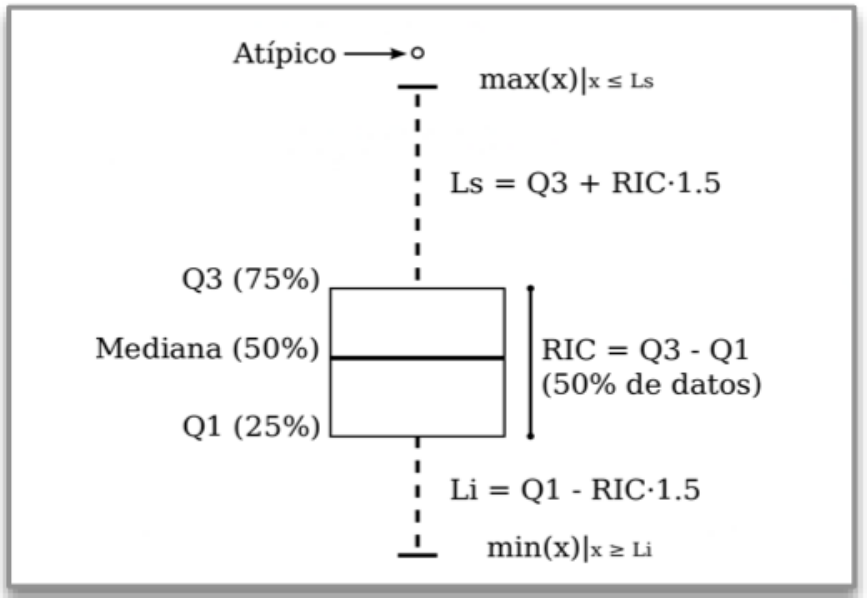
Figura 19. Gràfic en cascada



Font: Josep Curto

- **Diagrama de caixa**: és un tipus de gràfic que utilitza els quartils per representar un conjunt de dades. Permet observar d'una ullada la distribució de les dades i les seves principals característiques: centralitat, dispersió, simetria i mida de les cues, com es mostra a la figura 20.

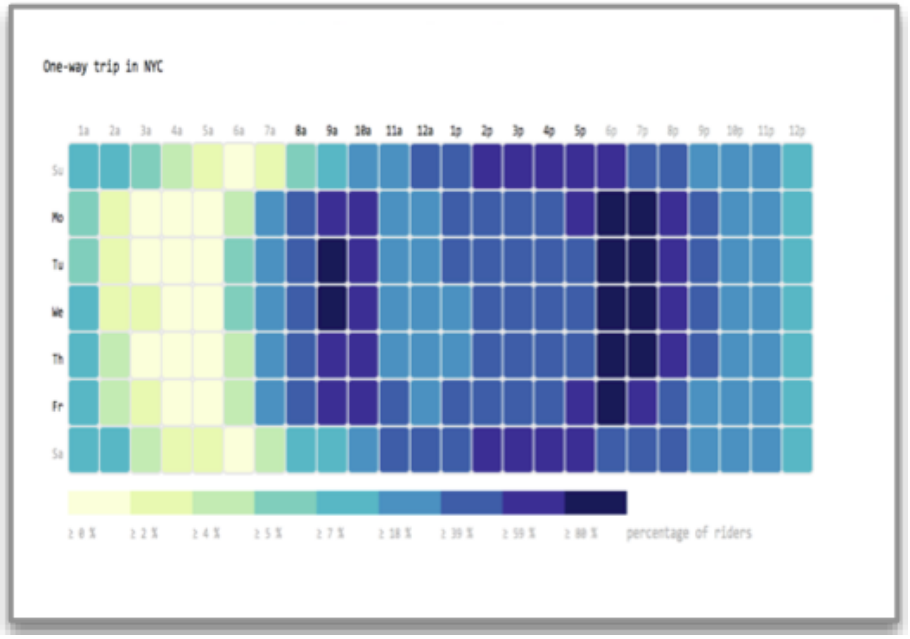
Figura 20. Diagrama de caixa



Font: Josep Curto

- **Mapa de calor**: és una representació gràfica de les dades on els valors individuals continguts en una matriu es representen com colors, com s'il·lustra a la figura 21.

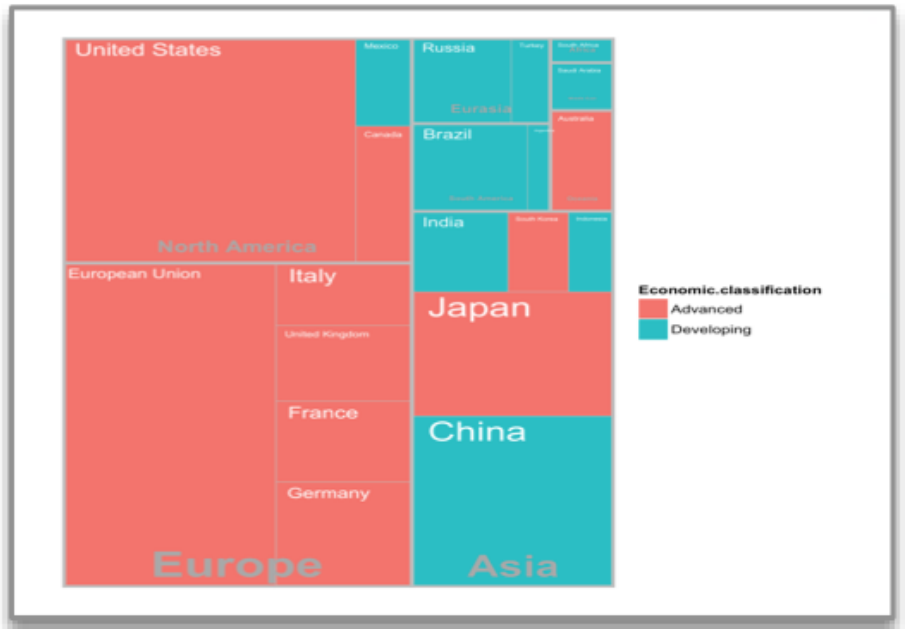
Figura 21. Mapa de calor



Font: Josep Curto

- ***Treemap***: és un mètode per a la visualització de dades jeràrquiques mitjançant l'ús de rectangles niats i de diferents mides com es mostra a la figura 22.

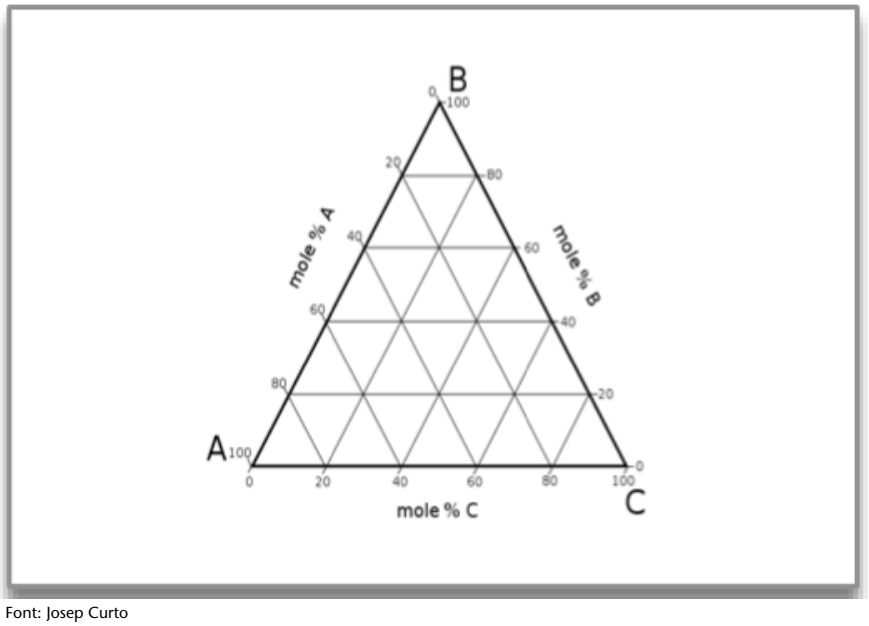
Figura 22. *Treemap*



Font: Josep Curto

- **Diagrames ternaris**: són usats per representar el percentatge relatiu de tres components on l'únic requeriment és que els tres components han de sumar un 100% com es mostra a la figura 23.

Figura 23. Diagrama ternari



Font: Josep Curto

Per escollir un gràfic cal seguir un procés sistemàtic a través d'una sèrie de preguntes:

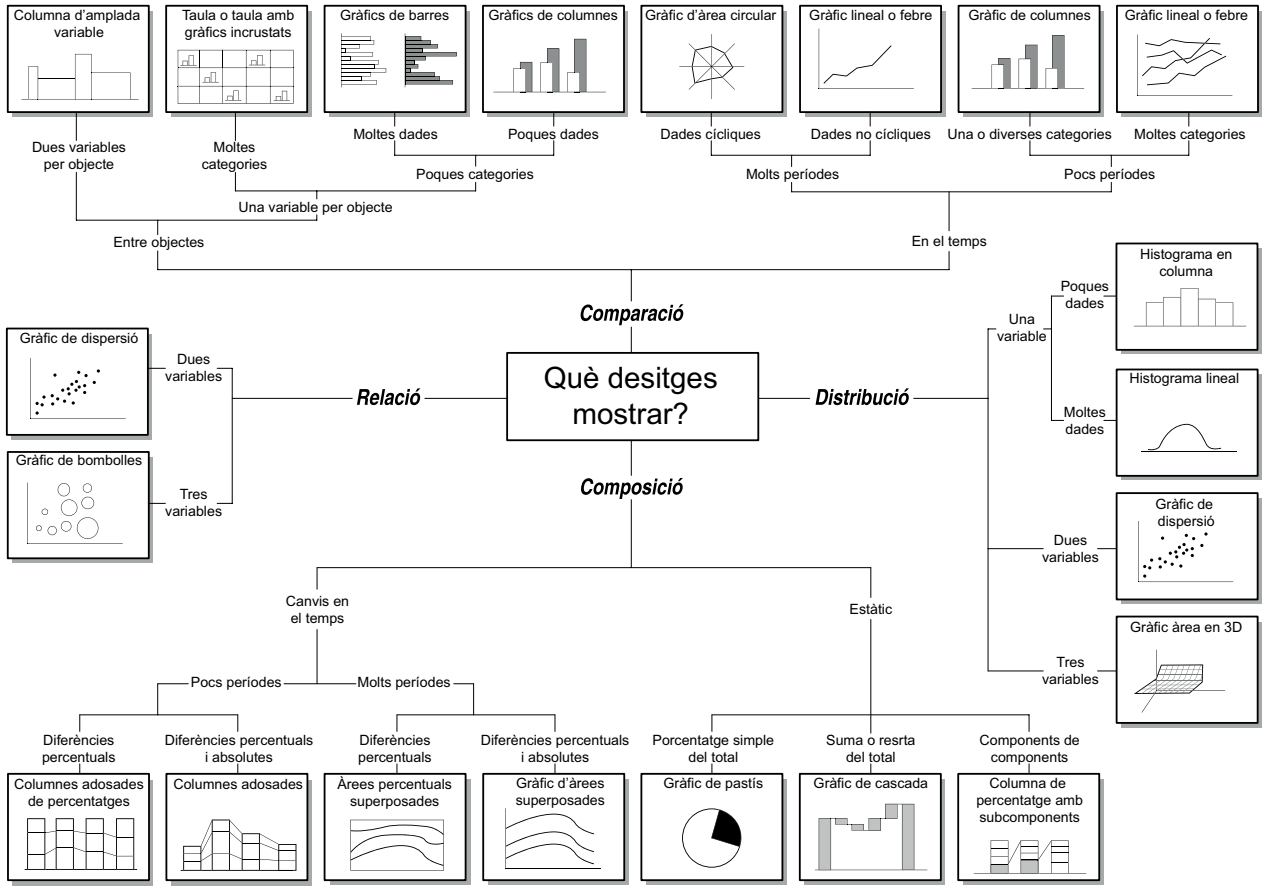
- **Què es vol mostrar?** Tenim diverses opcions: comparació, distribució, composició i relació.

- **Com és la dada?** Identificar el tipus de dada: quantitativa o qualitativa. Així com el tipus de variable: contínua o discreta.
- **Quantes variables tenim?** Podem tenir-ne una o més d'una, i tenir la necessitat de treballar amb elles.
- **És estàtic o canvia en el temps?** És a dir, cal identificar si hi ha una dimensió d'anàlisi temporal.
- **Depèn de la regió o de la cadena?** És a dir, cal identificar si hi ha una dimensió d'anàlisi geogràfica.

Es recomana revisar els criteris presentats en Extreme Presentation* que s'il·lustren a la figura 24.

*http://extremepresentation.com

Figura 24. Criteri de selecció



Font: Extreme Presentation

3.1.6. Cicle de vida d'un informe

Com ja s'ha definit, l'objectiu d'un informe és presentar els resultats d'una àrea o procés de negoci. En el moment de dissenyar un informe, no només cal tenir en compte la forma i el contingut que tindrà, sinó el seu cicle de vida perquè pugui continuar generant valor per a l'organització. És per això que hem d'introduir el que es coneix com el cicle de vida d'un informe, que es compon de les següents etapes:

- **Identificar**: consisteix a determinar els aspectes de negoci rellevants per a la seva comprensió i identificar les mètriques que representen aquests aspectes i que són rellevants per a la companyia i els seus gestors.

- **Mesurar:** consisteix a desenvolupar o revisar els sistemes d'informació que recopilen la informació necessària per a les mètriques. Inicialment, la companyia hauria de tenir ja implementats aquests sistemes, però no és estrany trobar-se amb la necessitat d'habilitar aquest tipus de sistemes.

- **Revisar**: consisteix a comprovar que la dada dels sistemes anteriors representa de manera efectiva, vàlida, completa i amb qualitat els processos de negoci, de manera que el sistema de *reporting* posterior tindrà aquestes característiques. En essència, estem parlant de governança de la dada.

Governança de dades

Quan parlem de governança de dades, fem referència a un conjunt d'estàndards, processos i polítiques que regeixen el desenvolupament i la utilització de les dades a nivell corporatiu.

- **Crear**: consisteix a crear l'informe i habilitar la seva distribució a les parts interessades.

- **Recopilar**: consisteix a recopilar de forma contínua el *feedback* per part dels usuaris així com les futures necessitats.

- **Millorar**: consisteix a implementar en el sistema de *reporting* les millores recopilades en el punt anterior. Aquestes millores poden ser en forma, contingut, distribució, qualitat de la dada, etc.

En aquest cicle és realment important detectar qui utilitzarà l'informe i per a què. A cada usuari, li agrada treballar la informació d'una manera i fa o no fa coses a partir d'aquesta premissa.

3.2. OLAP

Els informes proporcionen una visió estàtica del rendiment de l'organització. Per a alguns dels usuaris de negoci, normalment analistes, això no és suficient. Necessiten crear les seves pròpies anàlisis, filtrant la informació, creant noves mètriques, afegint la informació respecte de les diferents perspectives de negoci, establint relacions entre fets, etc.

Aquest tipus d'usuari necessita el que es coneix com OLAP (*Online Analytical Processing*), terme encunyat per Edgar F. Codd. Una manera senzilla d'entendre què significa aquest concepte és que es tracta d'una tecnologia que permet l'anàlisi multidimensional a través de taules matricials o pivotants (com les taules dinàmiques d'Excel). Si bé el terme OLAP s'introdueix per primera vegada el 1993, els seus conceptes base com, per exemple, l'anàlisi multidimensional, són molt més antics.

Tot i ser una tecnologia que ja té més de quatre dècades, les seves característiques i la seva evolució han produït que la gran majoria de solucions del mercat incloguin un motor OLAP.

Cal comentar que les eines OLAP dels diferents fabricants, si bé són similars, no són completament iguals atès que presenten diferents especificacions del model teòric.

3.2.1. OLAP com a eina d'anàlisi

OLAP forma part del que es coneix com a sistemes analítics que permeten respondre preguntes com **per què va passar?** Aquests sistemes poden trobar-se integrats en sistemes de *business intelligence* o ser simplement una aplicació independent.

Cal, abans de continuar, introduir una definició formal d'OLAP.

> S'entén per OLAP, o procés analític en línia, el mètode per organitzar i consultar dades sobre una estructura multidimensional. A diferència de les bases de dades relacionals, totes les potencials consultes estan calculades per endavant, fet que proporciona una major agilitat i flexibilitat a l'usuari del negoci.

Una eina OLAP està formada per un motor i un visor. El motor és, en realitat, el concepte que acabem de definir. El visor OLAP és una interfície que permet consultar, manipular, reordenar i filtrar dades existents en una estructura OLAP mitjançant una interfície gràfica d'usuari que disposa de funcions de consulta MDX entre d'altres.

MDX

Quan parlem d'MDX (*multidimensional Expressions*), fem referència a un llenguatge de consulta per a bases de dades OLAP.

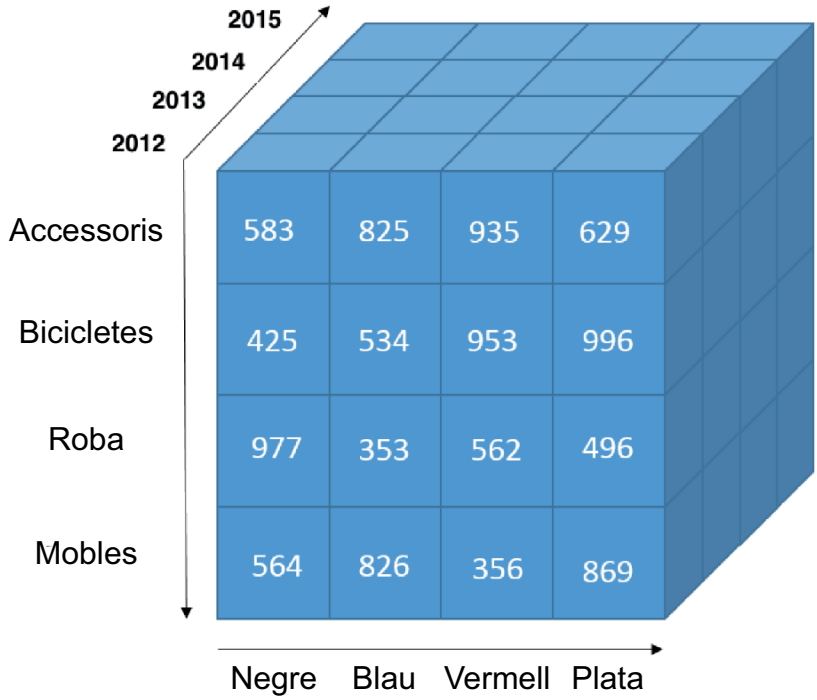
Les estructures OLAP permeten realitzar preguntes que serien summament complexes mitjançant el llenguatge de consulta a base de dades, conegut com SQL (*Structured Query Llenguatge*).

Considerem un exemple que ens permetrà entendre la potència d'aquest tipus d'eines.

> Imaginem que volem respondre la següent pregunta: quin és el marge de beneficis de la venda de bicicletes l'any 2014? Si tenim una anàlisi OLAP (de vegades anomenat cub), com el de l'exemple, format per l'any, els productes i la seva gamma de colors, la resposta és la intersecció entre els diferents elements. Cal observar que una estructura d'aquesta mena permet consultes molt més completes com, per exemple, comparar el marge de beneficis de 2013 i 2014 entre diferents productes, etc. A més, mitjançant el visor OLAP, proporcionen llibertat als usuaris finals per fer aquestes consultes de manera independent del departament d'IT.

La figura 25 il·lustra aquest exemple.

Figura 25. Cub



Font: Josep Curto

3.2.2. Tipus d'OLAP

Existeixen diferents tipus d'OLAP, que principalmente divergeixen en la manera com desen les dades:

- **MOLAP (*multidimensional OLAP*)**: és la forma clàssica d'OLAP i sovint ens hi referim amb aquest acrònim. MOLAP és una base de dades multidimensional (o més col·loquialment cub). En definitiva, es crea un fitxer que conté totes les possibles consultes precalculadas. A diferència de les bases de dades relacionals, aquestes formes d'emmagatzematge estan optimitzades per a la velocitat de càlcul. També s'optimitzen sovint per a la recuperació al llarg de patrons jeràrquics d'accés. Les dimensions de cada cub són típicament atributs com ara període, localització, producte o codi del compte. La manera com cada dimensió serà agregada es defineix per avançat.

- **ROLAP (*Relational* OLAP)**: és una forma d'OLAP on la dada és a la base de dades relacional, només en el moment de consulta, i es recuperen i es construeixen els resultats multidimensionals.

- **HOLAP (*hybrid* OLAP)**: No hi ha acord clar en la indústria pel que fa a què constitueix l'OLAP híbrid, excepte en el fet que és una base de dades en què les dades es divideixen entre emmagatzematge relacional i multidimensi-

onal. Per exemple, per a alguns venedors, HOLAP consisteix a utilitzar les taules relacionals per desar-hi les quantitats més grans de dades detallades; utilitza l'emmagatzematge multidimensional per a alguns aspectes de quantitats més petites de dades, menys detallades o agregades.

- ***Extreme* OLAP**: aquest nom s'està començant a utilitzar en la indústria per referir-se a un motor OLAP que treballa sobre alguna de les tecnologies *big data*.

- **DOLAP (*Desktop* OLAP)**: és un cas particular d'OLAP ja que està orientat a equips d'escriptori. Consisteix a obtenir la informació necessària des de la base de dades relacional i a desar-la en local. Les consultes i anàlisis són realitzades contra les dades emmagatzemades a l'escriptori.

- ***In-memory* OLAP**: aquest enfocament es fonamenta en l'ús de la memòria de l'ordinador per crear la consulta OLAP. En treballar només amb memòria, s'acceleren les operacions d'accés i consulta.

Cada tipus té certs avantatges, encara que hi ha desacord sobre les especificitats dels avantatges entre els diferents proveïdors:

- MOLAP és millor en sistemes més petits de dades, és més ràpid per calcular agregacions i retornar respostes, i necessita menys espai d'emmagatzematge. Darrerament, *in-memory* OLAP està posicionant-se com una opció per a MOLAP.

- ROLAP es considera més escalable. No obstant això, és difícil implementar eficientment el pre-procés de grans volums de dades, cosa que implica que es rebutgi amb freqüència. Altrament, el funcionament de les consultes podria no ser òptim.

- HOLAP està entre els dos en totes les àrees, però pot pre-processar ràpidament i escalar bé.

- *Extreme* OLAP serà usat quan l'empresa implementi projectes de *big data* i estigui interessada a tenir disponible anàlisi OLAP sobre grans volums de dades.

Tret d'*Extreme* OLAP, tots els tipus són, però, propensos a l'explosió de la base de dades. Aquest és un fenomen que genera la quantitat extensa d'espai d'emmagatzematge que és utilitzat per les bases de dades OLAP quan es resolen certes, però freqüents, condicions: alt nombre de dimensions, de resultats calculats per endavant i de dades multidimensionals escasses.

Les últimes tendències en OLAP inclouen la tecnologia *in-memory* així com la seva adaptació a tecnologies *big data*, com Apache Kylin* (creat per eBay) o Pinot** (creat per Linkedin).

*http://kylin.apache.org
**http://github.com/linkedin/pinot/

La dificultat en la implementació OLAP rau en la formació de les consultes, en l'elecció de les dades base i en el desenvolupament de l'esquema. Com a resultat, la majoria dels productes moderns es proveeixen de biblioteques enormes de consultes preconfigurades. Un altre problema que pot afectar el cub és treballar amb dades de baixa qualitat o incompletes.

3.3. Quadres de Comandament

Tant els informes com OLAP són eines que proporcionen informació als usuaris finals. La gran quantitat d'informació que normalment inclouen aquestes eines les pot fer inadequades per a usuaris que necessiten prendre decisions de manera ràpida a partir d'elles o que disposen de poc temps per fer la seva pròpia anàlisi.

El quadre de comandament prové del concepte francès *tableau de bord* i permet mostrar informació consolidada a alt nivell. Es focalitza en:

- Presentar una quantitat reduïda d'aspectes de negoci.
- L'ús majoritari d'elements gràfics.
- Inclusió d'elements interactius per potenciar l'anàlisi en profunditat i la comprensió de la informació consultada.

El quadre de comandament és una eina molt popular atès que permet entendre molt ràpidament la situació de negoci i és molt atractiva visualment. Per això, totes les solucions del mercat inclouen aquest tipus de solucions. L'oferta es diferencia principalment en la facilitat del procés de creació del quadre de comandament, en les opcions disponibles de visualització, i en la capacitat de treballar amb fluxos continus de dades i el reflex d'aquests canvis en temps real. Els quadres de comandament permeten l'anàlisi visual de la informació, el que es coneix com *Visual Analytics*.

El quadre de comandament sol usar-se també per a la direcció per objectius (DPO), que consisteix a identificar les àrees clau per a l'organització i definir els resultats esperats per a cadascuna d'elles i per a cadascun dels llocs directius. Per a cada àrea i directiu s'estableixen metes coordinades i negociades que es converteixen en indicadors de metes que permeten seguir-ne l'evolució la meta en un període de temps determinat. En definitiva, el quadre de comandament proporciona, en aquest escenari, suport a l'establiment de plans d'acció per assolir els objectius i controlar el seu procés d'assoliment.

Les últimes tendències que estan afectant els quadres de comandament inclouen *Data Visualization* i *Data Storytelling*. La primera fa referència a la inclusió d'una major quantitat d'elements gràfics per a la comprensió de la dada i l'ús de criteris per a l'ocupació d'aquests elements. La segona, al fet que l'eina

permet construir i explicar històries de negoci fonamentades en dades i fets per descriure què ha passat. No totes les eines del mercat inclouen aquesta tendència i només es troba en alguns productes innovadors.

Un quadre de comandament permet monitoritzar els processos de negoci atès que mostra informació crítica a través d'elements gràfics de fàcil comprensió. Aquest tipus d'eines, la periodicitat de refresc de les quals sol ser propera al temps real, és de gran utilitat per a tots aquells usuaris encarregats de prendre decisions diàriament.

Aquests sistemes poden trobar-se integrats en *suites* de *business intelligence* o ser simplement aplicacions independents.

Cal, abans de continuar, introduir una definició formal de quadre de comandament.

S'entén per quadre de comandament o *dashboard* el sistema que informa de l'evolució dels paràmetres fonamentals de negoci d'una organització o d'una de les seves àrees.

La informació que es presenta en un quadre de comandament es caracteritza per:

- Utilitzar diferents elements (gràfics, taules, alertes...).
- Combinar els elements de manera uniforme i precisa.
- Basar la informació presentada en indicadors clau de negoci.
- Presentar les tendències de negoci per propiciar la presa de decisions.

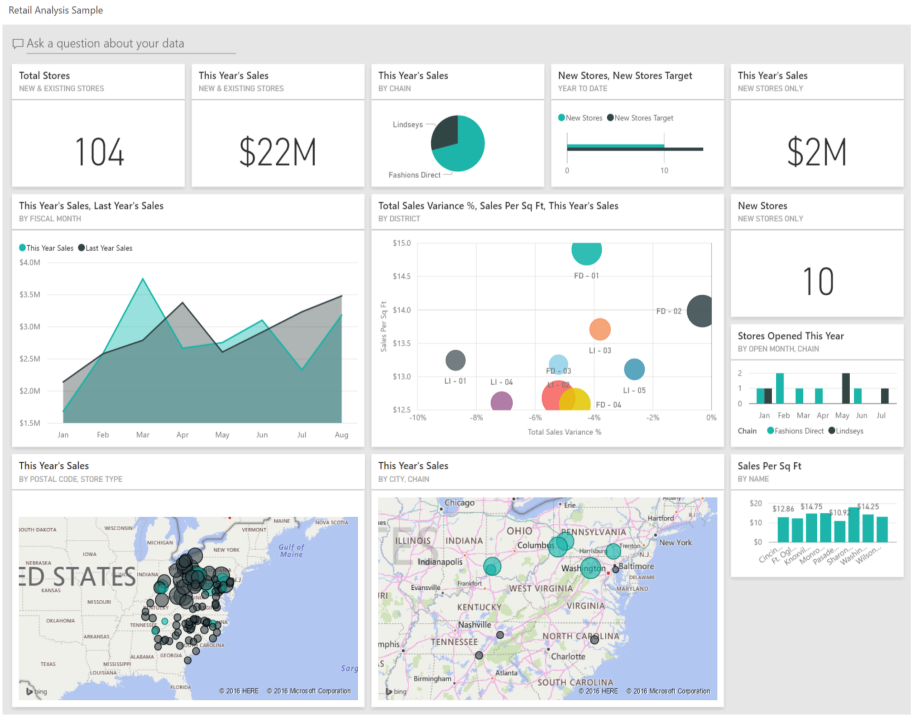
A què s'assembla un quadre de comandament? La figura 26 ens presenta un exemple que combina alguns dels diferents elements anteriors per al control del rendiment d'una companyia de *retail*. Aquest quadre de comandament permet conèixer l'evolució de les principals magnituds financeres, però al mateix temps aquelles específiques del negoci, com vendes per metre quadrat o expansió en botigues.

La tipologia d'usuaris que necessita aquestes eines és:

- Alta direcció, amb l'objectiu de comprendre el que succeeix en el negoci
- Gerents que han de monitoritzar processos de negoci
- Usuaris de negoci que necessiten poder fer una anàlisi exploratòria de la dada

Per tant, el quadre de comandament aporta valor a nivell estratègic, tàctic i operatiu.

Figura 26. Quadre de Comandament



Font: Power BI (MSFT)

Principalment, un quadre de comandament està format per diversos elements combinats. El quadre de comandament comparteix la majoria dels elements dels informes a excepció del fet que ha d'incloure **menús de navegació**, que faciliten a l'usuari final realitzar operacions amb els elements del quadre de comandament.

3.3.1. Procés de creació d'un quadre de comandament

El procés de crear un quadre de comandament és un procés iteratiu que combina diversos passos:

- Identificar la necessitat de negoci i els potencials usuaris del quadre de comandament.

- Triar les dades que es mostraran en el quadre de comandament. En aquest punt, cal tenir en compte les necessitats de l'usuari final, un cop s'han mantingut les reunions necessàries per identificar els requisits.

- Escollir el format de presentació. A partir de la informació que es mostrarà i les necessitats del client, és possible determinar quin tipus d'element d'un quadre de comandament és el més adequat. Es recomana realitzar un esbós.

- Integrar, combinar dades i presentar-les conjuntament. Un cop tenim els diferents elements, es realitza un esbós amb tots ells.

- Planificar l'interactivitat de l'usuari.

- Implementació del quadre de comandaments. En aquest punt entra l'eina seleccionada i inclou els següents passos:

– Aconseguir les dades i formatar-les per aconseguir els KPIs.
– Formatar els elements del quadre de comandament en funció de les capacitats de la solució escollida.

3.3.2. *Dashboard* davant *Balanced Scorecard*

Sovint, es confon el quadre de comandament o *dashboard* amb el quadre de comandament integral o *balanced scorecard*. La raó és la similitud dels noms en castellà. Necessitem definir aquest nou concepte.

S'entén per *balanced scorecard* el mètode de planificació estratègica basat en mètriques i processos ideat pels professors Kaplan i Norton, que relaciona factors mesurables de processos amb la consecució d'objectius estratègics.

La teoria del *Balanced Scorecard* va sorgir als anys 90 com a resposta davant la necessitat d'analitzar les organitzacions des d'un punt de vista diferent al financer, que estava quedant obsolet. L'objectiu era establir un nou model de mesures que permetés conèixer millor les organitzacions.

Per fer això, l'institut Nolan Norton va patrocinar un estudi d'un any amb l'objectiu de definir un *scorecard* corporatiu, en què van participar diverses companyies de múltiples sectors. D'aquest estudi va sorgir el concepte de *Balanced Scorecard* que organitzava indicadors clau de negoci en quatre grans grups o perspectives: financera, client, interna i innovació i aprenentatge.

Balanced reflecteix que els indicadors tracten de ser un equilibri entre els objectius a curt i llarg termini, entre les mesures financeres i les no financeres, entre els indicadors de retard o lideratge, i entre les perspectives internes i externes.

És així que el *Balanced Scorecard* permet traduir l'estratègia de l'empresa en un conjunt comprensible de mesures de rendiment que proporcionin el marc de mesura estratègica i de sistema de gestió.

Un quadre de comandament integral està format pels següents elements:

- **Perspectiva**: punt de vista respecte del qual es monitoritza el negoci. Segons aquesta metodologia, tota empresa té quatre perspectives: financera,

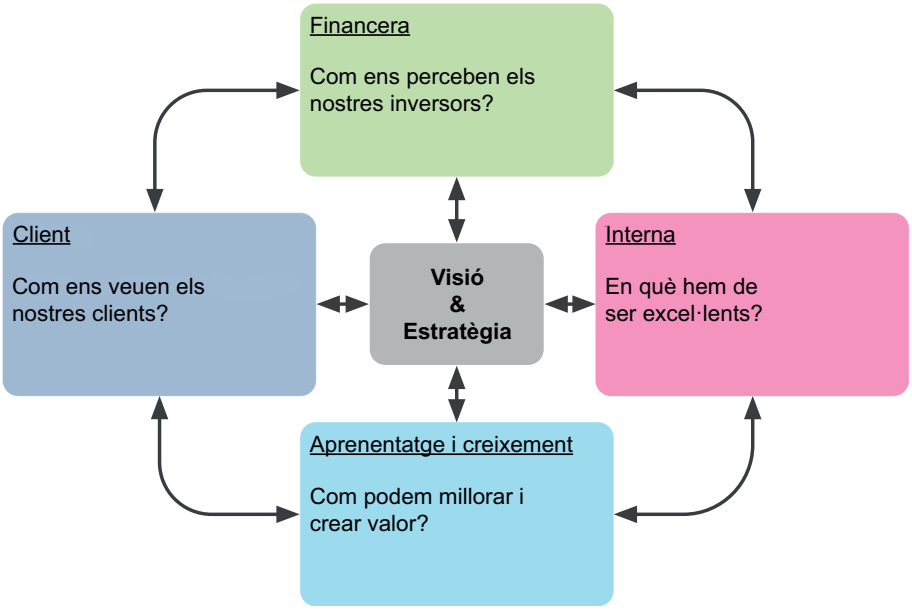
de client, de processos, i d'aprenentatge i creixement. Si bé pot estendre o reduir-se en nombre de perspectives. Detallem-ne ara les perspectives clàssiques:

– **Financera**: permet mesurar les conseqüències econòmiques de les accions preses en l'organització. Incorpora la visió dels accionistes i mesura la creació de valor de l'empresa.
– **Client**: reflecteix el posicionament de l'empresa en el mercat o en els segments de mercat on vol competir.
– **Interna**: pretén explicar les variables internes considerades com a crítiques, així com definir la cadena de valor generat pels processos interns de l'empresa.
– **Aprenentatge i creixement**: identifica la infraestructura que l'organització ha de construir per crear creixement i valor a llarg termini.

- **Objectius**: que s'han de complir en cadascuna de les perspectives.
- **Línies estratègiques**: engloben els objectius que segueixen una relació de causalitat.
- **Indicadors**: són principalment KPIs.
- **Relacions causa-efecte**: permeten comprendre com la consecució d'un objectiu impacta en un altre.
- **Plans d'acció**: accions que es realitzen per a la consecució d'un objectiu.
- **Pesos relatius**: importància d'un objectiu dins d'una perspectiva o d'una línia estratègica.
- **Matriu d'impacte**: permet dirimir com un pla d'acció afecta els objectius i en quina mesura que ho fa.

La figura 27 representa les diferents perspectives tradicionals.

Figura 27. Perspectives *Balanced ScoreCard*



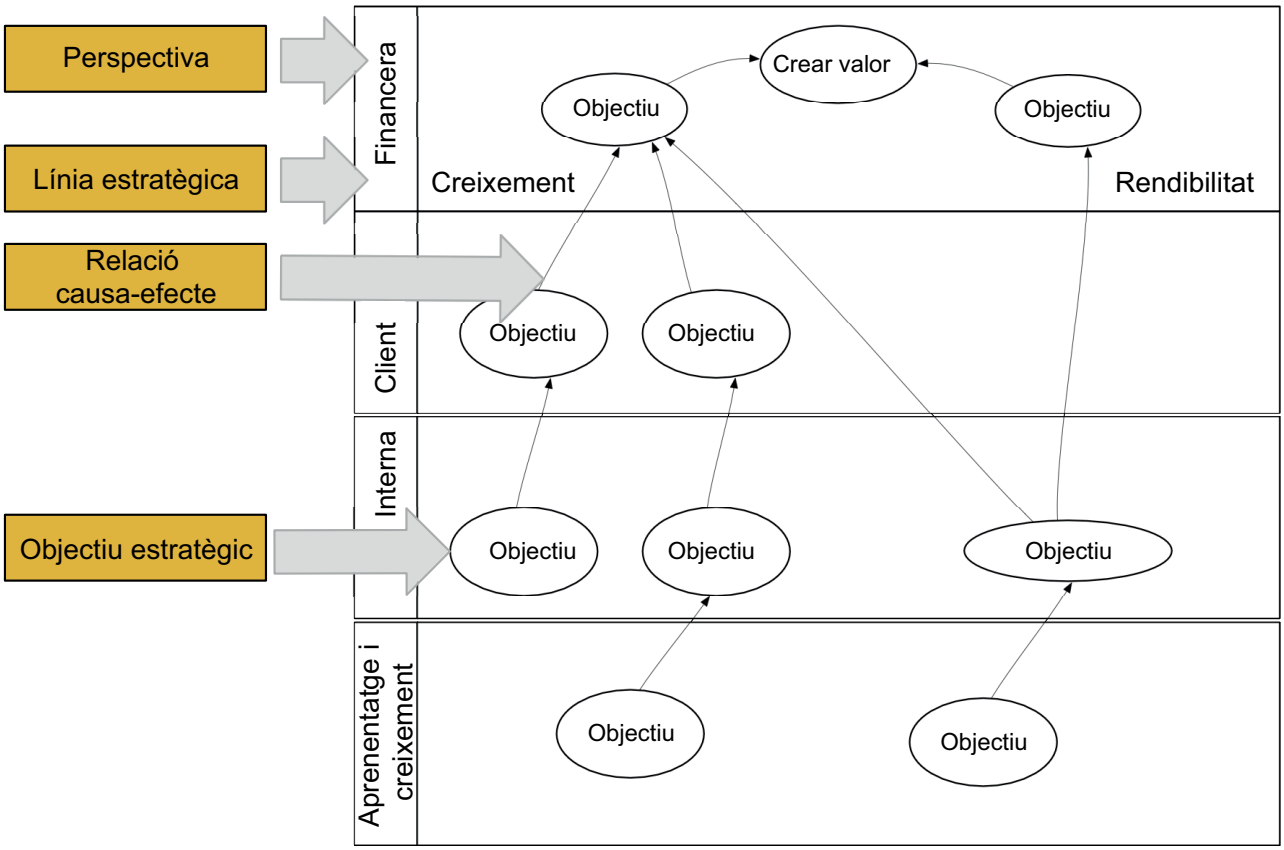
Font: Josep Curto

El procés de construcció d'un quadre de comandament integral és:

- Definir les perspectives de negoci. Sovint, les perspectives clàssiques són suficients per representar l'estratègia.
- Definir per a cada perspectiva els objectius estratègics.
- Definir per a cada objectiu plans d'acció per aconseguir aquests objectius.
- Definir indicadors per monitoritzar la consecució dels objectius.
- Definir les relacions de causalitat entre els objectius.
- Identificar les línies estratègiques a què pertanyen els objectius estratègics.

Aquest procés s'estructura a través d'un mapa estratègic que podem veure a la figura 28 que representa el procés anterior.

Figura 28. Construcció *Balanced ScoreCard*



Font: Josep Curto

Un punt important que cal destacar és que un *Balanced ScoreCard* ha de ser flexible i àgil, de manera que la recopilació d'informació s'ha de dur a terme de forma ràpida, senzilla i en el temps oportú perquè les accions que se'n derivin puguin realitzar-se de manera eficaç.

La implantació d'un quadre de comandament integral proporciona els següents beneficis:

- Defineix i clarifica l'estratègia.
- Subministra una imatge del futur mostrant el camí que condueix a ell.
- Comunica l'estratègia a tota l'organització.
- Permet alinear els objectius personals amb els departamentals.
- Facilita la vinculació entre el curt i el llarg termini.
- Permet formular amb claredat i senzillesa les variables més importants objecte de control.
- Constitueix un instrument de gestió.
- Facilita el consens en tota l'empresa ja que explicita el model de negoci de l'organització i el tradueix en indicadors.
- Permet comunicar els plans de l'empresa, unir els esforços en una sola direcció i evitar la dispersió. En aquest cas, el CME actua com un sistema de control per excepció.
- Permet detectar de forma automàtica desviacions en el pla estratègic o operatiu, i fins i tot explorar les dades operatives de la companyia fins a descobrir la causa original que ha donat lloc a aquestes desviacions.

Per tant, hi ha clares diferències entre un quadre de comandament i un quadre de comandament integral que es recullen a la taula 2.

Taula 2. Diferències entre Quadre de Comandament i Quadre de Comandament Integral

Característica	Quadre de Comandament	Quadre de Comandament Integral
Objectiu	Monitoritzar una àrea de negoci i prendre decisions operatives i/o tàctiques	Definir l'estratègia d'una organització i enllaçar l'estratègia amb l'operativa a través de plans d'acció
Elements	Taules, gràfics, llistes, alertes, menús, mapes...	Perspectives, objectius, indicadors, metes...

Està ara clar que són eines diferents que responen a necessitats diferents i que cal no confondre-les. La confusió sol venir del fet que un quadre de comandament pot suportar DPO, BSC i/o control estadístic.

4. Què és *business analytics*?

Com s'ha comentat abans, el més important per a una organització no és ser capaç d'emmagatzemar o processar dades, sinó de generar valor a partir d'elles. El valor pren la forma de l'anàlisi.

L'anàlisi es concentra en dues grans àrees: intel·ligència de negoci, amb enfocament a conèixer el rendiment passat, i que ja hem introduït, i *analítica de negoci*, amb enfocament a predir el rendiment futur i conèixer patrons ocults en la dada.

4.1. Definició de *business analytics*

Recuperem la definició que ja hem introduït.

> S'entén per ***business analytics*** el conjunt d'estratègies, tecnologies i sistemes per a la identificació i comprensió de patrons, i el desenvolupament de capacitats predictives respecte del rendiment de l'organització.

Business analytics permet respondre preguntes diferents de les de la intel·ligència de negoci com, per exemple:

- Per què va passar?
- Que passarà?
- Què passarà si canviem X?
- Quins patrons oculten les dades que no hem identificat?

Per poder respondre aquesta mena de preguntes, *business analytics* inclou diferents tipus d'anàlisi que revisarem a continuació.

4.2. Tipus de *business analytics*

En l'analítica de negoci tenim diferents tipus d'anàlisi. En destaquem els següents, tot i que no és una taxonomia exhaustiva ni exempta de solapaments:

- **Anàlisi estadística / quantitativa**: branca de les matemàtiques que investiga la recol·lecció, l'anàlisi, la interpretació i la presentació de dades d'u-

na mostra representativa; busca explicar les correlacions i dependències d'un fenomen físic o natural, d'ocurrència aleatòria o condicional. L'anàlisi quantitativa és un conjunt de tècniques d'anàlisi estadística que pot incloure, entre d'altres, l'anàlisi quantitativa del comportament.

El departament comercial pot utilitzar l'anàlisi estadística per analitzar si hi ha una correlació entre les vendes i l'època de l'any.

- **Mineria de dades**: és una tècnica que permet l'extracció d'informació i coneixement a partir de la dada.

  El departament comercial pot utilitzar la mineria de dades per agrupar els clients fonamentant-se en molts i diversos atributs al mateix temps com vendes, canal, perfil demogràfic, etc.

- **Mineria de textos**: és una tècnica que permet l'extracció d'informació i coneixement a partir de text.

  El departament comercial pot utilitzar la mineria de text per analitzar les opinions dels clients compartides en les xarxes socials.

- **Mineria de processos**: és una tècnica que permet l'anàlisi dels processos de negoci basada en *logs* d'esdeveniments.

  El departament comercial pot utilitzar la mineria de processos per analitzar el rendiment del procés comercial des que el client mostra interès fins que es realitza la compra.

- ***Machine Learning***: coneguda també com a aprenentatge automàtic, és una branca de la informàtica que ha evolucionat des de l'estudi i reconeixement de patrons cap a la intel·ligència artificial. Es fonamenta en diferents tipus d'algoritmes classificats en aprenentatge supervisat, no supervisat i basats en reforços.

  Una organització pot usar *machine learning* per crear un sistema de recomanació de productes.

- **Intel·ligència artificial**: és una àrea multidisciplinària que combina computació, matemàtiques, lògica... i que busca el disseny de sistemes capaços de resoldre per ells mateixos problemes quotidians, utilitzant com a paradigma la intel·ligència humana.

  Una organització pot usar la intel·ligència artificial per crear un agent de conversa (*chatbot*) que automatitza i simula interaccions humanes amb els clients per tal de, per exemple, proporcionar una resposta ràpida a peticions d'informació de producte o serveis.

- **Analítica de continguts**: és una tècnica que permet l'extracció d'informació i coneixement de contingut, com poden ser imatges o vídeos. El focus no només està en l'extracció de valor, sinó també en la composició automàtica de continguts personalitzats.

Sistemes cognitius

Dins dels sistemes que busquen simular la intel·ligència humana, destaquen els sistemes cognitius. La computació cognitiva fa referència a *maquinari* i *programari* que simula el funcionament del cervell humà per prendre decisions. L'aprenentatge es fonamenta en instruccions i experiència.

Una organització pot usar l'analítica de continguts per personalitzar les comunicacions amb els clients.

- **Analítica de grafs**: és una tècnica que permet l'extracció d'informació i coneixement de dades estructurades com un graf.

  Una organització pot usar l'analítica de grafs per detectar en les xarxes socials els seguidors més rellevants de l'organització.

- **Analítica visual**: és una tècnica que habilita l'exploració de dades i la detecció de patrons a través de tècniques de visualització.

  El departament comercial pot utilitzar l'analítica visual per analitzar el rendiment de les diferents zones geogràfiques.

- **Modelització predictiva**: és una tècnica per a la representació de models mitjançant tècniques estadístiques o matemàtiques (com equacions diferencials), que permet identificar representacions i fer prediccions.

  El departament comercial pot utilitzar la modelització predictiva per identificar els factors que incideixen en la compra dels productes i serveis, i estimar què passarà en períodes següents.

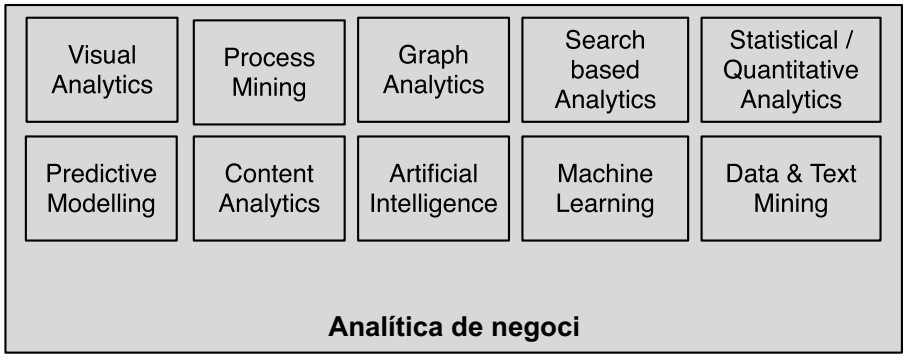
Equacions diferencials

Quan parlem d'equacions diferencials, estem fent referència a una equació matemàtica que relaciona una funció i les seves derivades.

L'estadística, la intel·ligència artificial, *machine learning* i la mineria de dades i text són disciplines relacionades i, en realitat, habiliten els casos d'ús presentats en aquesta taxonomia.

La figura 29 il·lustra els components en l'analítica de negoci.

Figura 29. Analítica de negoci



Font: Josep Curto

En general, quan parlem de *business analytics* fem referència a solucions encapsulades que només cal parametritzar. Aquestes solucions estan optimitzades per a un sector i un procés. Per exemple, podem parlar d'anàlisi de la taxa d'abandonament per al sector de les telecomunicacions o la detecció del frau en el sector financer.

Quan parlem d'aplicar aquesta tècnica –i per a això cal desenvolupar una solució *ad hoc*–, és quan estem considerant el que es coneix actualment com *data science*.

4.3. Beneficis de *business analytics*

La implantació de *business analytics* proporciona diversos beneficis entre els quals podem destacar:

- Permetre als usuaris de negoci plantejar hipòtesis i validar-les. Com, per exemple, quin és el factor més rellevant per al comportament dels nostres clients.

- Poder treballar amb escenaris diversos i comparar els seus resultats. Com, per exemple, usant arbres de decisió per veure quin test és el més probable davant d'un quadre de símptomes d'un pacient.

- Poder trobar i analitzar patrons en les dades. Com, per exemple, fer servir la segmentació de clients per definir diferents estratègies de preus per als clients.

- Poder automatitzar tasques manuals basades en regles i patrons. Com, per exemple, la identificació de preguntes repetides o similars en Quora.

- Poder fer recomanacions de productes i serveis de manera automàtica. Com, exemple, els sistemes de recomanació de Spotify o Netflix.

- Trobar els motius reals darrere d'un succés. Com, per exemple, en quin punt de la xarxa i per què ha fallat la retransmissió *online* d'un partit de futbol.

- Poder fer accions preventives. Com, per exemple, en la detecció de les àrees amb major potencialitat de crims a Rio durant els Jocs Olímpics del 2016.

5. El nou context de negoci

L'ús de dades per prendre millors decisions no és nou. De fet, des de fa temps les organitzacions s'han estat ancorant en estratègies com la intel·ligència de negoci i/o l'analítica de negoci. Però una sèrie de condicions en el mercat han propiciat que sigui necessària una nova estratègia per a l'anàlisi de dades: ***big data***.

En aquest apartat, ens centrarem en comprendre quines són aquestes noves condicions del mercat, què ha canviat de la naturalesa de la dada i, finalment, discutirem per què les tecnologies clàssiques de *business intelligence*, basades en el *data warehouse* i les bases de dades relacionals, moltes vegades no són suficients per gestionar el nou entorn de dades i de negoci.

5.1. Què ha canviat des del punt de vista de negoci

En les últimes dècades, les tecnologies de la informació poc a poc han anat agafant més rellevància en les organitzacions. S'han transformat en un component bàsic per a les operacions automatitzant, d'una banda, part o fins i tot el procés sencer i, de l'altra, proporcionant suport a les diferents necessitats departamentals (des de finances fins a màrqueting).

Aquesta ha estat una progressiva transformació digital i moltes empreses encara estan en aquest procés. En els últims anys, la transició s'ha accelerat per diversos factors com ara la democratització d'Internet, l'adveniment de les xarxes socials, l'emergència dels dispositius intel·ligents i/o el desplegament de l'Internet de les Coses. I com a resultat, les organitzacions es troben en un període de competitivitat i evolució disruptiva basada en TI* i fonamentada principalment en quatre factors: el *cloud*, social, mobilitat i analítica. Expliquem aquests factors tecnològics:

***A aquest període, Gartner l'anomena el nexe de forces; IDC l'explica com la tercera plataforma i Cognizant simplement fa servir una paraula, SMAC, acrònim de *Social, Mobile, Analytics* i *Cloud*.**

- ***Cloud***: fa referència a tecnologies que permeten consumir recursos TI (des d'emmagatzematge fins a un CRM) gestionats per tercers i que freqüentment comparteixen el seu ús.
- **Social**: fa referència a les tecnologies que permeten facilitar les interaccions socials.
- **Mobilitat**: fa referència a les tecnologies que habiliten accés a informació i interaccions amb independència de la localització.
- **Analítica**: fa referència a aquelles tecnologies que maximitzen la utilitat de la dada.

CRM

És l'acrònim de *Customer Relationship Management*, que fa referència a la gestió de la relació amb clients.

Aquests quatre factors provoquen que els models de negoci siguin diferents o, fins i tot, que se'n generin de nous. El *cloud* permet que tinguem flexibilitat en la implementació, en el desplegament, en l'escalabilitat i en la globalització; el social redefineix la forma en què interactuem amb clients, empleats i proveïdors; el mòbil amplia els canals d'interacció i desdibuixa el perímetre del que coneixem com a empresa; i, finalment, l'analítica significa ja no només que podem conèixer el que passa en l'organització sinó que podem utilizar la informació com a font d'avantatge competitiu. En definitiva, és una transformació de la relació entre persones, negoci i tecnologia.

Com a resultat, hem assistit a l'explosió de noves formes d'apropar-se al mercat i generar valor per al client i l'organització. Per exemple, sabem d'empreses que han aconseguit crear sistemes de recomanació per als seus clients, com Amazon, dissenyar productes basats en preferències, com Netflix, o identificar el risc creditici basat en fonts tan diferents d'informació com les xarxes socials o les compres a eBay i Amazon, com Kreditech*. Tot i que companyies com Facebook, Google o Netflix acaparen l'atenció pels seus avenços en l'ús de les tecnologies de dades, la realitat és que estem vivint una revolució d'ampli espectre i moltes altres empreses ja estan apostant per la implementació d'aquest tipus de projectes.

***http://www.kreditech.com**

De fet, és possible trobar exemples en nombrosos sectors; aquestes aplicacions tenen diverses formes i colors, i freqüentment estan profundament especialitzades. Per exemple, en el context dels *Massively Multiplayer Online Game* (MMOG), o videojocs multijugador massiu en línia, empreses com Jagex* ja monitoritzen les transaccions de micropagaments i el funcionament dels sistemes que suporten les operacions usant tecnologies de *big data*. En el sector de l'esport, equips com el FC Barcelona analitzen grans quantitats de dades en diferents formats (vídeos, estadístiques, dades geolocalitzats, etc.) per comprendre millor el rendiment propi com a equip i de forma individual, així com el dels equips contraris, i dissenyar conseqüentment estratègies més eficients per guanyar. En el sector de l'agricultura, *big data* permet millorar l'eficiència dels sistemes de reg, en ser la peça clau per integrar i analitzar dades d'estacions meteorològiques, informes de plagues i malalties, sensors en plantes, boques de reg i sòl de parcel·les, i sistemes d'informació com ERP, com en el cas del celler Luna Beberide**.

ERP

És l'acrònim d'*Enterprise Resource Planning*, que fa referència a la gestió dels recursos d'una organització.

***http://www.jagex.com**

****http://www.lunabeberide.es**

Però, com passa cada vegada que apareix una nova tecnologia innovadora i d'avantguarda que té el potencial de transformar profundament la societat, no resulta senzill portar a bon port la implementació. I una primera pregunta sorgeix: **En quina mesura ha canviat la dada?**

5.2. La naturalesa de la dada

Tal com hem comentat, estem vivint una explosió en la complexitat de la dada. Per entendre aquesta complexitat cal parlar sobre la seva naturalesa,

fer un incís sobre què entenem per les magnituds físiques de la dada i entrar en detall en dos punts cada vegada més rellevants: on es troben les dades importants per a una organització i el crucial paper de les metadades.

5.2.1. Les magnituds físiques de la dada

Hi ha tres magnituds físiques de la dada: volum, velocitat i varietat.

- Quan parlem de **volum**, fem referència a la grandària del conjunt de dades creat diàriament. En tan sols una dècada, les organitzacions han passat de treballar amb Terabytes a haver de lluitar amb Petabytes o magnituds superiors.

- Quan parlem de **velocitat**, fem referència tant al processament de dades com a la seva latència. El primer fa referència a la quantitat de dades en moviment (mesurat en termes de Gigabytes o Terabytes per segon). El segon fa referència a la suma de retards temporals que s'aplica a la ingestió de dades i a la seva anàlisi de manera separada o conjunta (mesurat en mil·lisegons). Això implica tractar amb dades des de processos *batch* fins a en temps real i/o *streaming*.

- Quan parlem de **varietat**, fem referència tant a la quantitat de fonts diferents que s'han de combinar (respecte de formats com vídeo, àudio, text...), com a l'heterogeneïtat de la dada (que pot ser estructurada, semiestructurada o no estructurada).

byte

Quan parlem de byte, fem referència a una unitat de mesura d'informació digital. Parlarem de diversos bytes:
Gigabyte (GB) 10^9 bytes
Terabyte (TB) 10^{12} bytes
Petabyte (PB) 10^{15} bytes
exabytes (EB) 10^{18} bytes
Zettabyte (ZB) 10^{21} bytes
Yottabyte (IB) 10^{24} bytes

Latència

Quan parlem de latència, fem referència a la suma de retards temporals en la captura, emmagatzematge, processament i anàlisi de la dada.

Més enllà de les magnituds físiques és possible trobar altres característiques com:

- **Veracitat**, que fa referència a la incertesa en la dada producte de la seva baixa qualitat, l'ambigüitat en la seva definició o simplificacions en la seva modelització.

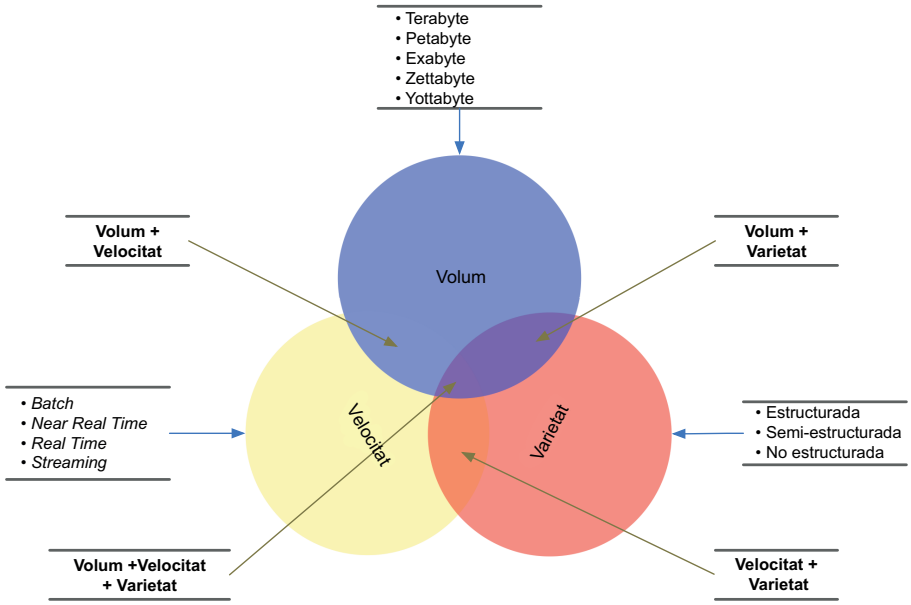
- **Variabilitat**, que fa referència al fet que els fluxos de dades poden tenir comportaments erràtics o inconsistents en certs períodes.

- **Vinculació**, que fa referència a la dificultat de relacionar diferents i diverses fonts de dades.

Cal comentar que no sempre ens trobem amb aquestes tres últimes característiques, tot depèn de la naturalesa del problema que es pretén resoldre. És per això que només es parla de les tres primeres, conegudes com les 3 Vs.

Diferents problemàtiques de negoci tindran associades diferents combinacions d'aquestes magnituds com es mostra a la figura 30.

Figura 30. Magnituds físiques de la dada



Font: Josep Curto, adaptat de SmartDataCollective

Hi ha un últim punt que hem de comentar respecte de la naturalesa de la dada: el **Valor**. Encara que hàgim discutit com podem entendre la complexitat de la dada, l'important no és si una organització és capaç de gestionar la dada en repòs, en moviment i/o en les seves múltiples formes i fonts. El més rellevant és com una organització és capaç de generar valor a partir de la dada i quin impacte té per al negoci i per als clients. En una primera instància, aquest valor pot prendre diferents formes:

1) Presa de decisions: l'ús de la dada ens permet prendre millors i/o més ràpides decisions, cosa que es tradueix en què l'organització és més competitiva en el seu mercat. És a dir, som capaços de prendre decisions informades.

2) Ingressos: l'ús de la dada permet millorar els ingressos en línies de negoci o n'habilita la creació de noves.

3) Costos: l'ús de la dada permet optimitzar els nostres processos de negoci tant a nivell de sistemes com de persones, fet que implica que podem fer més amb menys.

A posteriori analitzarem més en profunditat els casos d'ús per a la dada.

5.2.2. On es troben les dades rellevants per al negoci?

Hem parlat en el subapartat anterior sobre l'aspecte més important de la dada: el seu valor. O el que és el mateix, que la dada sigui rellevant per al negoci. Davant el nou context, les organitzacions necessiten treballar amb la dada deixant enrere la noció que la informació de valor es troba tan sols en el si de l'organització. Aquest fet obliga a pensar no només en les magnituds de la dada sinó també en l'origen de partida de les dades. Per tant, hem de parlar de dades internes i externes.

- **Dades internes**: fa referència a dades que pertanyen a l'organització. Dins de les dades internes tenim aquelles que ja existeixen o es creen en els propis sistemes d'informació de l'organització (com poden ser l'ERP i/o el CRM), o bé que s'estan capturant i emmagatzemant mitjançant mecanismes automàtics a través de diferents estratègies com el *crowdsourcing*, sensors i/o dispositius de monitorització (com un podòmetre amb localització geogràfica).

Crowdsourcing

Quan parlem de *crowdsourcing*, fem referència al procés d'obtenir serveis, idees i contingut a través de la participació d'una gran massa de persones.

- **Dades externes**: fa referència a dades de tercers i que s'havien d'aconseguir per a l'organització. Aquestes fonts de dades poden estar disponibles per la seva compra o ser de lliure accés. Les dades de lliure accés poden ser, al seu torn, de tres tipus: dades capturades mitjançant tècniques de *crowdsourcing*, dades de xarxes socials (com poden ser Facebook, Twitter o LinkedIn) i *open data*.

Open data

Quan parlem d'*open data*, fem referència a conjunts de dades considerades un bé comú i que, per això, són gratuïtes, accessibles i ben estructurades per descarregar-les i analitzar-les. Les tipologies són múltiples: de transport, financeres, meteorològiques, estadístiques, científiques, culturals i geolocalitzades.

5.2.3. Metadades: més enllà del valor de la dada

Cal parlar d'una última potencial font de dades que, tot i que es podria incloure dins de la categoria de dades internes, no se sol contemplar dins de les organitzacions. Estem parlant de la metadada i el valor associat a ella. Hem de definir què és la metadada.

S'entén per metadades dades estructurades i codificades que descriuen característiques d'un objecte, dada o procés de negoci.

Referència bibliogràfica

Conesa, J.; Curto, J. (2012). *Introducción al business intelligence*. Barcelona: Editorial UOC.

És a dir, no és suficient generar valor a partir de la dada, sinó també a partir de les metadades vinculades a ella. Podem parlar de tres grans categories de metadades:

- **Tècniques**: descriuen els aspectes tècnics vinculats a la dada. Com poden ser les seves magnituds o, per exemple, els drets de propietat.

- **Operacionals**: fan referència als processos de captura, transformació, emmagatzematge, anàlisi i visualització de la dada, incloent-hi les fórmules de càlcul.

- **Atributs**: fan referència als atributs que enriqueixen la informació sobre la dada. Per exemple, en una fotografia trobem aspectes com el dispositiu amb el qual es va realitzar.

La metadada obre la porta a una nova gamma d'anàlisi del valor i, sobretot, a comprendre d'una manera molt més profunda el que passa en una organització. Informació que no sempre és rellevant per a l'usuari final, però sí molt important per al sistema que gestiona la dada.

> Considerem el correu electrònic, la trucada o el missatge de text que s'usa per a la comunicació tant personal com professional. En aquest cas, les metadades són l'horari, la data en què es va enviar, la durada, els agents que participen en la conversa i la localització des d'on es va connectar l'usuari l'última vegada, entre d'altres. Aquesta informació no revela el contingut de les comunicacions, sinó de les transaccions electròniques mostrant els seus patrons, relacions i comportaments.

Després de discutir la naturalesa de la dada, una altra pregunta natural sorgeix: **Per què necessitem una nova tecnologia per analitzar la dada?**

5.3. Les limitacions del *data Warehouse*

Tradicionalment, les organitzacions han abordat la necessitat d'analitzar dades i generar valor mitjançant dos sistemes interconectats: el *data Warehouse* i la intel·ligència de negoci, o *business intelligence* (BI). Sabem que el *data warehouse* és un repositori de dades que proporciona una visió global, comuna i integrada de les dades de l'organització, mentre que la intel·ligència de negoci és un conjunt de metodologies, aplicacions, pràctiques i capacitats enfocades a millorar la presa de decisions.

El *data warehouse* ha estat el component principal per a l'emmagatzematge de dades i el BI per a la seva explotació. No obstant això, a mesura que les organitzacions han anat progressant en la seva transformació digital, la complexitat de la dada ha anat augmentant i noves necessitats han emergit. Aquestes són algunes d'elles:

- La presa decisions necessita integrar dades estructurades, semi-estructurades o no estructurades.

- La presa de decisions necessita treballar amb estructures de dades que no són persistents en el temps.

- La presa de decisions necessita considerar tota la informació associada a un procés de negoci, el que es tradueix, per a alguns d'ells, en grans quantitats d'informació, no processables de forma eficient.

- La presa de decisions s'ha de fer en temps real accelerant la captura i el consum de la dada.

- La presa de decisions s'ha de fonamentar en l'ús d'aplicacions analítiques on la metadada del procés juga un paper fonamental en la comprensió i el descobriment del que ha passat.

- La dada es reutilitzarà per diferents anàlisis i, per això, es necessita desar en brut o aplicant-hi el mínim de transformacions possible.

Una manera d'entendre aquest tipus d'escenaris és comparar els casos d'ús del magatzem de dades respecte dels nous casos d'ús, tal com es recull en la taula 3.

Taula 3. *Data Warehouse* vs. nous escenaris d'ús de la dada.

Factor	Data Warehouse	Nous escenaris d'ús de la dada
Fonts de dades	Sistemes corporatius i transaccionals	Fonts no tradicionals com: sensors, logs, vídeos, etc.
Volum	Fins a 100 Terabytes	A partir de 100 Terabytes
Velocitat	*Batch* o processos que no requereixen resposta immediata	Resposta immediata
Varietat	Principalment estructurada	De tot tipus
Veracitat	Organitzada i de qualitat	De qualitat variable
Valor	BI i analítica	*Machine Learning*, *Deep Learning* i anteriors
Objectiu	Presa de decisions	Diversa, però destaca la creació de productes i serveis de dades

La gran majoria d'implementacions de *data Warehouse* han estat creades de manera optimitzada per a la generació d'informes, quadres de comandaments, així com l'anàlisi OLAP. Escenaris enfocats a l'anàlisi de rendiment passat d'una organització i que han d'estar fonamentats en informació de qualitat.

Els nous escenaris no han estat tractats pel magatzem de dades i fins i tot no formen part de les seves capacitats; per tant, com a resposta a aquesta necessitat, ha emergit una nova generació de tecnologies i enfocaments que amplia les capacitats de la nostra organització a nous casos d'ús.

Estem parlant, per tant, de casos d'ús diferents, encara que complementaris, que cal comprendre per identificar les capacitats que s'han de desenvolupar, així com els costos associats. El *data Warehouse*, en definitiva, permet proporcionar respostes a preguntes recurrents en una organització (que són conegudes). Les noves tecnologies obren la porta a escenaris flexibles d'anàlisi en els quals l'estructura de la dada i de les preguntes que cal respondre o bé no són conegudes, o bé canvien freqüentment.

6. Què és *big data*?

El 2009, IDC va estimar la mida de la informació digital generada i desada, a la qual va anomenar Univers Digital, en 0,8 Zettabytes (ZB) i va predir que per a l'any 2020 s'arribarien als 35 ZB. Posteriors estudis de la mateixa companyia han revisat la xifra a l'alça per a aquest any i l'han ajustada a 45 ZB, sent de 8 ZB per a l'any 2015. Aquesta revisió en les prediccions il·lustra l'acceleració fruit de l'aparició de cada vegada més fonts que produeixen i consumeixen dades, d'una major incorporació d'usuaris a Internet, del desplegament d'una major quantitat de dispositius intel·ligents i del continu desenvolupament de solucions i serveis digitals.

Lectura complementària

Gantz, J.; Riensel, Sr. (2009). *As the Economy Contracts, the Digital Universe Expands*. New York: IDC

Aquesta explosió de dades està caracteritzada per un creixement en les magnituds físiques de la dada: volum, varietat i velocitat. Es crea un major volum de dades, provinents d'una major varietat de fonts, representades en múltiples formats i que s'han de capturar i consumir a una major velocitat. Aquest nou paradigma de les dades es coneix freqüentment com ***Big data***, si bé el nom genera confusió tenint en compte la seva referència a només una de les magnituds (volum). En essència, estem parlant d'una explosió en la complexitat de la dada.

6.1. Definició de *big data*

Es considera que *big data* és un concepte novell, atès que hi ha múltiples definicions*.

*Tal com s'argumenta en el següent article acadèmic: *Undefined By Data: A Survey of Big Data Definitions*. Font: http://arxiv.org/pdf/1309.5821v1.pdf. I com també ha argumentat Timo Elliot, evangelista de SAP, al seu bloc: http://timoelliott.com/blog /2013/07/7-definitions-of-big-data -you-should-know-about.html.

Cal comentar que podem trobar referències a la problemàtica de la dada ja al 2001, quan Doug Layney** va apuntar que el creixement de les dades en volum, varietat i velocitat propiciaria la necessitat d'invertir en noves tecnologies que permetrien capturar, extreure, processar, desar i analitzar les dades en la nova era. Però els orígens del terme poden trobar-se fins i tot abans, a la dècada dels 90, en les converses dins de la comunitat de Silicon Graphics dirigida pel científic John Mashley, en què s'analitzaven les principals tendències de futur.

**En aquell moment, aquest analista pertanyia a MetaGroup, actualment pertany a Gartner.

El fet que hi hagi múltiples definicions complica la seva comprensió i la identificació d'escenaris dins de la pròpia organització. La gran majoria d'elles inclouen el que es coneix com les 3Vs del *big data*, que hem comentat en l'anterior

subapartat: volum, velocitat i varietat, que són magnituds físiques de la dada. Podem trobar, però, altres definicions que n'inclouen algunes més com, per exemple, la veracitat.

Per tal de tenir un enfocament pragmàtic, farem servir la següent definició.

S'entén per ***Big data*** el conjunt d'estratègies, tecnologies i sistemes per a l'emmagatzematge, processament, anàlisi i visualització de conjunts de dades complexes.

Cal recordar que, quan parlem d'emmagatzematge, ens referim als suports físics i de *programari* que permeten desar la dada en estructures que representen la seva complexitat; que, quan parlem de processament, ens estem referint a aquelles operacions que permeten la ingestió, la transformació i la distribució de la dada per adequar-la al consum; que, quan parlem d'anàlisi, fem referència a les tècniques aplicades per generar valor; i que, quan parlem de visualització, fem referència als mecanismes de consum d'informació.

6.2. Tipus de *big data*

La definició de *big data* emmascara, en certa mesura, les complexitats del que suposa treballar amb dades extremes en termes del seu volum, velocitat i varietat. Ja hem introduït en el subapartat 5.2. en què consisteix la nova naturalesa de la dada.

En aquest sentit i, amb l'objectiu de millorar la comprensió de la definició de *big data*, cal parlar de les diverses tipologies existents de *big data*.

6.2.1. Classificació de NIST

D'acord amb el NIST i, en particular dins del seu grup de treball de *big data**, una forma de categoritzar *big data* és mitjançant les necessitats de negoci:

- El model de negoci no es pot representar mitjançant una estructura de dades relacional (és a dir, mitjançant una base de dades relacional).
- El model de negoci necessita ser escalable pel creixement de dades respecte de la seva velocitat o volum.

*http://bigdatawg.nist.gov

NIST

NIST és l'acrònim de *National Institute of Standards and Technology*, una institució americana que estudia, defineix i promou estàndards tecnològics.

La base d'aquesta classificació és poder identificar correctament els diferents escenaris que es deriven d'aquestes condicions, ja sigui mitjançant recursos interns, ja sigui mitjançant l'ajuda d'especialistes externs. La combinació d'aquests dos aspectes ens proporciona tres tipologies de *big data* i un escenari en el qual no hi ha aquesta necessitat. Tot i que és palès el valor que aporta *big data*, no tots els problemes d'una organització són necessàriament un problema de dades. Els tipus disponibles es resumeixen a la taula 4.

Estructura de dades relacional

Quan parlem d'estructura de dades relacional, fem referència a un tipus de base de dades que permet establir interconnexions o relacions entre les dades desades en taules.

Taula 4. Tipus de *big data*. Font: NIST

Tipus	Descripció
Tipus 1	On una estructura de dades no relacional és necessària per a l'anàlisi de negoci
Tipus 2	On cal aplicar estratègies d'escalabilitat horitzontal per processar i analitzar de forma eficient el negoci
Tipus 3	On cal processar una estructura de dades no relacional mitjançant estratègies d'escalabilitat horitzontal per processar i analitzar de manera eficient el negoci

Escalabilitat

Quan parlem d'escalabilitat, fem referència a l'habilitat d'un sistema, xarxa o procés per reaccionar i adaptar-se sense perdre qualitat, o bé gestionar el creixement continu de treball de manera fluïda, o bé per estar preparat per fer-se més gran sense perdre qualitat en els serveis oferts. L'escalabilitat horitzontal està fonamentada en l'increment de nodes del sistema, procés o xarxa. Mentre que la vertical consisteix a afegir més recursos –memòria, disc dur i/o processadors–.

Així doncs, per a una determinada necessitat de negoci, és possible identificar si estem en un escenari de *big data* o no, i si és necessari aquest tipus de tecnologies, fet que cada vegada més s'erigeix com un punt rellevant i de partida per a la implementació d'aquest tipus de projectes. Això es resumeix en la taula 5.

Taula 5. Autoavaluació de l'existència de *big data*. Font: NIST

Volum	Velocitat	Varietat	Escalabilitat horitzontal	Estructura no relacional	Tipus de *Big data*
No	No	No	No	No	No
No	No	Sí	No	Sí	Sí, Tipus 1
No	Sí	No	Sí	Potser	Sí, Tipus 2
No	Sí	Sí	Sí	Potser	Sí, Tipus 3
Sí	No	No	Sí	Potser	Sí, Tipus 2
Sí	No	Sí	Sí	Sí	Sí, Tipus 3
Sí	Sí	No	Sí	Potser	Sí, Tipus 2
Sí	Sí	Sí	Sí	Sí	Sí, Tipus 3

Aquesta taula permet avaluar una necessitat de negoci. Una reflexió interessant és que, dins d'una mateixa organització, poden plantejar-se diferents escenaris de *big data* per resoldre diferents necessitats de negoci, cosa que, en definitiva, apunta que seran necessàries diferents tecnologies de *big data*.

Cas: Infojobs

InfoJobs és un dels principals portals d'ocupació a Europa a través d'una plataforma *online* que connecta empreses (que publiquen ofertes) i persones (que busquen noves oportunitats per a la seva carrera professional).

La companyia estava interessada a poder oferir recomanacions analitzant les trajectòries professionals més habituals dels seus usuaris (cosa que també es coneix com a creixement professional). Això es tradueix en analitzar una gran quantitat d'informació que ha d'estructurar-se com una xarxa que relaciona candidats (més de 4 milions), experiències professionals (més de 12 milions) i capacitats (més de 18 milions).

Usant la taula anterior, estem en una situació en què prevalen el volum i la varietat i, per tant, és un escenari *big data* de tipus 3.

6.3. Quan és necessari *big data*?

La no existència d'una definició formal, una que permeti distingir de manera completament precisa quan una organització està en una situació de necessitat de *big data*, ha generat barreres en l'adopció d'aquest tipus de tecnologies.

Hem vist que hi ha escenaris en què no és suficient treballar amb un *data Warehouse*. Tenim també, per als tipus de *big data*, una classificació que permet dirimir escenaris genèrics. No obstant això, no és suficient en el context d'una organització on l'experimentació no té un ampli marge.

En aquesta aproximació més pragmàtica a *big data*, les organitzacions estan treballant amb cinc grans categories de casos d'ús. Aquests casos d'ús són moviments organitzacionals d'una estratègia de negoci enfocada a *big data*. Aquests casos són:

- Presa de decisions
- Operacions i intel·ligència operacional
- Validació d'hipòtesis i resolució de problemes
- Productes i serveis basats en dades
- Comerç de dades

Explicarem en detall cadascun d'aquests moviments organitzacionals.

6.3.1. Presa de decisions

El primer cas d'ús és la presa de decisions. Aquesta aproximació consisteix en l'ampliació de les capacitats tradicionals de presa de decisions mitjançant les tecnologies de *big data*. Això significa que els sistemes d'intel·ligència de negoci i magatzems de dades corporatives poden alimentar-se o combinar-se amb els repositoris de *big data*.

Cas: NH

NH és una cadena hotelera amb més 400 hotels a 25 països. Dins de l'estratègia de millora del servei per als seus clients, la companyia selecciona cada any diversos hotels sobre els quals farà millores. Les millores que es realitzen van des de l'ampliació del personal fins a la creació de noves instal·lacions. Per prendre la decisió d'"on cal invertir en aquest període", NH s'ha fonamentat tradicionalment en dues fonts de dades:

- Les dades financeres consolidades en el *data warehouse* de la companyia.
- Una sèrie d'enquestes realitzades als clients per conèixer la seva satisfacció en relació amb els serveis i instal·lacions de l'hotel. Aquestes enquestes, a causa dels costos associats a la seva realització, no són exhaustives i no cobreixen tots els hotels ni tots els clients.

En els últims anys, a NH s'han adonat que, per conèixer la satisfacció del client així com les àrees de millora, la informació rellevant es troba més enllà del perímetre de

l'organització. Els clients d'NH comparteixen les seves impressions a través de diferents canals com poden ser TripAdvisor, Yelp o Expedia. És a dir, estem parlant de fonts de dades externes a l'organització i, a més, no estructurades o amb diferents formats.

L'enfocament de l'organització ha estat complementar la informació financera en el *data warehouse* amb informació externa que s'emmagatzema i es processa amb tecnologies de *big data* i mineria de text (per extreure els comentaris rellevants per a la millora dels hotels), i que permet millorar i complementar la presa de decisions en un procés ja existent.

6.3.2. Operacions i intel·ligència operacional

El segon cas d'ús són les operacions i la intel·ligència operacional, que succeeixen en temps real. Aquesta aproximació consisteix en l'aplicació d'aquestes tecnologies en l'àmbit d'operacions tant per al control i l'anàlisi de procés de negoci com per al disseny i implementació de sistemes transaccionals. Aquest segon escenari transcendeix la presa de decisions i permet entendre per què les tecnologies *big data* estan cridades a ser molt rellevants dins de les tecnologies d'informació. És previsible que s'integrin de manera natural en múltiples aplicacions.

Estem parlant, per tant, d'una banda, de sistemes d'intel·ligència i detecció de patrons en temps real i, de l'altra, de sistemes operacionals que, o bé per les seves necessitats en escalabilitat, o per la seva complexitat en l'esquema de les dades ja no es fonamenten en tecnologies relacionals.

Cas: Santader / Caixabank

Un dels punts més rellevants per a moltes organitzacions és quan interaccionen amb els seus clients. És el que anomenem moments de la veritat. Entre aquests moments, en destaca aquell en què el client es posa en contacte amb una organització per a la resolució d'una incidència. En èpoques anteriors s'ha automatitzat el procés (mitjançant sistemes de resposta predefinida) o s'ha dividit el servei en diverses capes dins i fora de l'organització per tenir diferents nivells de servei.

En el context financer són diverses les entitats espanyoles que ja fan servir *speech analytics* per entendre les emocions dels seus clients durant les seves interaccions en una trucada. És possible, per tant, detectar quan disminuirà la satisfacció del client i actuar en conseqüència.

6.3.3. Validació d'hipòtesis i resolució de problemes

Un dels escenaris més importants és la validació d'hipòtesis i resolució de problemes. Aquest escenari consisteix a trobar solucions per a problemes de negoci que no han estat prèviament abordats en una organització i per als quals no hi ha preguntes predefinides. És a dir, es busca conèixer què ha passat, quins factors són els més rellevants i el perquè. Cal crear hipòtesis i validar-les mitjançant la tècnica més adequada i eficient. Aquest tipus d'aplicació és l'equivalent a tenir Sherlock Holmes a casa. És, en definitiva, un entorn que ha de ser prou flexible per funcionar en diferents escenaris de necessitats.

El resultat pot ser una solució puntual o una proposta que passi a convertir-se en un dels altres escenaris.

Cas: Sky

Sky és una empresa que produeix i distribueix continguts de vídeo tant en directe com sota demanda, amb presència en diversos països europeus. La distribució de continguts de vídeo es realitza a través de xarxes informàtiques conegudes com *content delivery networks*. Aquestes xarxes estan formades per diversos elements, tant pertanyents a la pròpia companyia com a tercers, que han d'assegurar que la distribució es realitza mantenint el nivell de qualitat contractat pel client. Sovint, el procés de transmissió de vídeo a través de la xarxa es controla i es mesura per assegurar el seu correcte funcionament. Aquest és el cas de Sky. No obstant això, mesos enrere va tenir una gran fallada en la seva *content delivery networks* durant la transmissió de la jornada futbolística al cap de setmana, que produïa errors d'accés al sistema, congelació de la imatge o transmissió d'imatges en baixa qualitat. Tot i tenir un sistema de monitorització, Sky no coneixia els motius pels quals havia succeït aquesta caiguda de qualitat en el servei.

L'enfocament de l'organització ha estat contractar un expert per investigar els fets. Això es tradueix en aquest cas en treballar amb milions de registres en format *log* i investigar les causes de l'error. Després d'aplicar el que es coneix com *root cause analysis* a un conjunt de dades emmagatzemades en tecnologies de *big data* per la seva grandària, es van trobar les raons de l'error i es van proposar una sèrie de millores per a la xarxa.

6.3.4. Productes i serveis de dades

El quart escenari d'ús és la creació de productes i serveis basats en dades. La dada es transforma en la peça angular per millorar l'experiència d'ús del producte i servei, o per al seu disseny i desplegament.

Estem parlant de models de negoci en què la dada i els algoritmes analítics generen valor per al client i l'organització, i per això modifiquen tots els aspectes primordials del model de negoci.

Cas: Nest

Nest és una empresa que produeix dispositius intel·ligents, en concret, termòstats, detectors de fum i càmeres de vigilància, entre d'altres. Va ser adquirida per Google el 2014.

Els productes creats per Nest són un exemple de producte basat en dades i algoritmes. Per exemple, el termòstat conté diferents tipus de sensors per detectar les persones i els animals presents a la llar, així com la temperatura de la llar. A més, va registrant les preferències de les persones que viuen a casa. És a dir, a quina hora són a casa i quina és la temperatura que prefereixen. I ho combinen amb informació contextual, com on es troba la casa i de quina època de l'any es tracta. Amb totes aquestes dades, es crea un perfil de preferències i, a partir d'un cert moment, el dispositiu comença a treballar de manera automàtica. Aquest procés automàtic permet reduir el cost energètic.

Existeix també un altre potencial beneficiari d'aquestes dades, tot i que en format agregat: les empreses productores i distribuïdores d'energia. A partir de les dades de tots els usuaris de Nest en aquelles zones geogràfiques on ofereixen el seu servei, poden conèixer les necessitats energètiques i, per tant, ajustar la demanda.

6.3.5. Comerç de dades

L'últim escenari d'ús és el comerç de dades. La dada es prepara per a la seva venda a tercers. Això pot incloure diversos processos com agregació, transformació i distribució de la dada o, en el cas de contenir informació sensible, emmascarar aquestes dades perquè el conjunt final contingui dades anònimes. Aquest tipus d'ús també pot derivar en haver de dissenyar una plataforma *ad*

hoc. La dada pot comercialitzar-se en brut o sota la forma de coneixement.

Cas: Vodafone / TomTom

Vodafone és una coneguda companyia que proporciona serveis de telecomunicacions a nivell mundial. TomTom és una companyia que ofereix productes i serveis de GPS.

Els serveis de TomTom permeten conèixer la ruta òptima a un conductor. La qualitat d'aquest servei depèn de treballar amb dades actualitzades, incloent-hi accidents o embussos de trànsit. Per això, TomTom compra dades de tercers a, per exemple, Vodafone.

En aquest cas particular, Vodafone comercialitza les dades agregades, anònimes i geolocalitzades dels usuaris de la seva xarxa. En el cas de tenir una gran acumulació d'usuaris en un mateix lloc (i ser aquest indret en una carretera), això es tradueix en una situació d'embús o accident. Per la qual cosa TomTom pot fer servir aquesta informació per proposar una ruta alternativa i oferir una millor experiència al client.

7. Tecnologies de *big data*

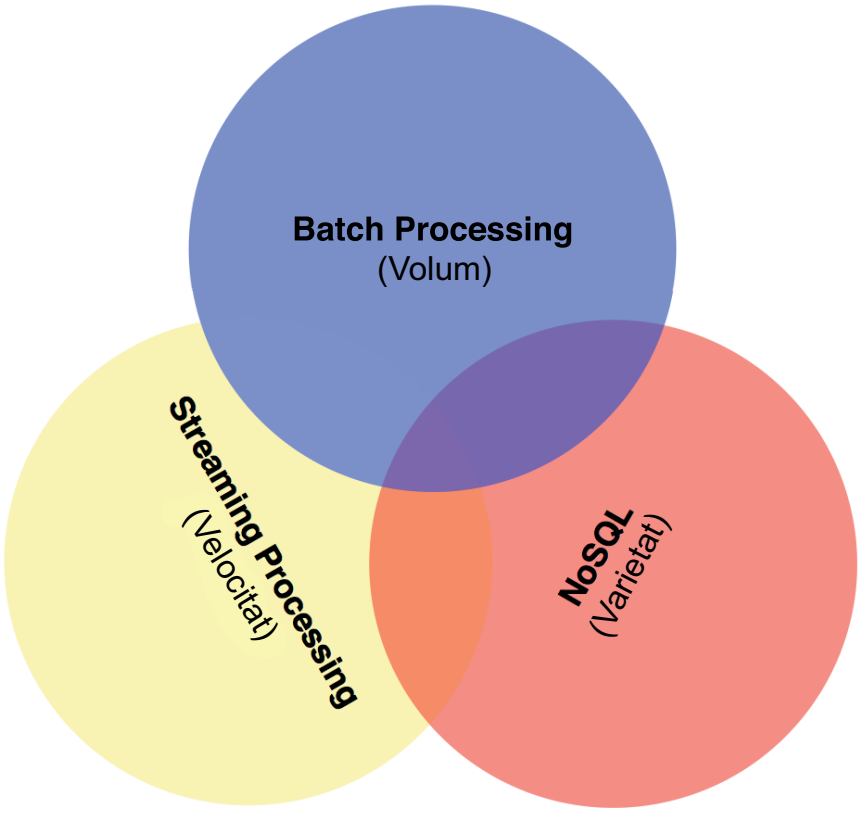
Un primer enfocament per pensar en les tecnologies de *big data* és recuperar les 3Vs presentades en el subapartat 5.2. Diferents problemàtiques de la dada necessiten diferents paradigmes, tal com apunten Casado i Younas.

Lectura complementària

Casado, R.; Younas, M. (2015). "Emerging trends and technologies in big data processing". *Concurrency and Computation: Practice and Experience*, (nº 27 (8), pàgs. 2078-2091).

Quan el volum és la principal problemàtica, s'usen les tecnologies de *batch processing*; per a la velocitat, les de *streaming processing*; i per a les de varietat, *NoSQL*, com s'il·lustra a la figura 31.

Figura 31. Tecnologies de *big data*



Font: Casado, R.; Younas, M.

NoSQL

Quan parlem de NoSQL, fem referència a bases de dades no relacionals. NoSQL és l'acrònim de *Not Only SQL*. SQL és l'acrònim de *Structure Query Language* i fa referència al llenguatge de consultes de bases de dades relacionals.

Hi ha una correspondència directa entre l'explosió a la problemàtica en la dada i l'emergència d'una determinada tecnologia. D'aquesta manera, tenim:

- **Tecnologies de processament per lots o *batch processing***: Permeten resoldre problemes vinculats amb el volum de la dada.

Criteo* és una companyia que ofereix solucions per a màrqueting digital basades en dades. Aquesta organització fa servir les tecnologies de processament per lots per consolidar dades i optimitzar els seus algoritmes d'anàlisi de campanyes de màrqueting.

*http://www.criteo.com

- **Tecnologies de processament en flux o (*streaming processing*)**: Permeten resoldre problemes vinculats a la velocitat de la dada.

Capital One** és una entitat bancària que ofereix productes i serveis financers a consumidors. Aquesta organització fa servir les tecnologies de processament en flux per monitoritzar l'activitat dels seus clients en temps real.

**http://www.capitalone.com

- **NoSQL**: permeten resoldre problemes vinculats a la varietat de la dada.

Metlife*** és una entitat asseguradora amb presència internacional. Aquesta organització fa servir les tecnologies NoSQL per integrar totes les referències de client en un únic punt d'accés i tenir una visió de 360 graus.

***http://www.metlife.com

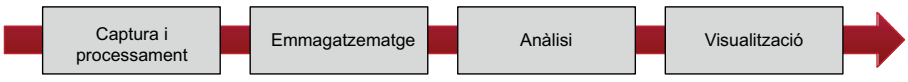
Aquesta aproximació a les tecnologies de *big data* està principalment centrada en dos punts: l'emmagatzematge i el processament. En aquest material, a ampliarem els punts que tractarem afegint-hi l'anàlisi i la visualització. El motiu darrere d'aquest enfocament és que els canvis en les capes de processament i emmagatzematge influeixen en la resta.

Quan parlem de tecnologies de *big data*, ens estem referint, en realitat, a una col·lecció de components, plataformes i solucions que cobreixen les diferents necessitats envers la dada. Aquestes necessitats són:

- **Emmagatzematge**: permetre l'emmagatzematge de la dada d'acord amb les necessitats de negoci.
- **Processament**: permetre la captura, la transformació i el moviment de la dada d'acord amb les necessitats de negoci.
- **Anàlisi**: permetre la generació de valor per al negoci a partir de la dada.
- **Visualització**: permetre la presentació i comunicació dels resultats d'acord amb les necessitats de negoci.

Aquestes necessitats es combinen seguint un flux com representa la figura 32.

Figura 32. Flux de *big data*



Font: Josep Curto

Moltes de les tecnologies de *big data* tenen origen *codi obert* per accelerar la innovació, la qual cosa vol dir que podem tenir accés a una versió *community* i, al mateix temps, diversos fabricants ofereixen una plataforma de pagament amb diferents components integrat i preparat a nivell empresarial.

7.1. Emmagatzematge

En les últimes dècades, les bases de dades relacionals han estat l'opció d'emmagatzematge *de facto* per als sistemes d'informació. En alguns contextos amb grans necessitats d'emmagatzematge i processament, com pot ser la meteorologia, s'ha treballat amb sistemes combinats de *maquinari* i *programari* optimitzats per a tasques intensives en la dada, i coneguts com *High Performance Computing* (HPC). L'enfocament d'HPC s'ha fonamentat principalment en l'escalabilitat vertical.

HPC

Quan parlem d'HPC, fem referència a la pràctica d'afegir capacitat de computació de manera que millora el rendiment d'una estació de treball i fa possible abordar problemes complexos en la ciència, enginyeria i/o negocis.

Amb l'emergència de *big data* això està canviant de forma significativa, principalment per diversos motius:

- La tecnologia relacional no és escalable per suportar el volum de dades en el context de *big data*.
- La tecnologia relacional és incompatible amb les dades no estructurades, que cada vegada són més rellevants per al negoci.
- La nova tecnologia no necessita HPC per executar-se, sinó que pot treballar amb xarxes d'ordinadors treballant de manera combinada amb prestacions de computació menors individualment, però majors col·lectivament.

En el context d'un projecte de *big data*, existeixen diferents tecnologies d'emmagatzematge que habiliten estratègies eficients i escalables tant en cost com en resposta a les necessitats de la naturalesa de la dada. Una de les característiques d'aquest tipus de sistemes és que proporcionen alta disponibilitat (*High Availability* o HA) i/o tolerància a fallades (*Fault Tolerance* o FT). Encara que similars, no són el mateix. D'una banda, HA implica tenir un esquema en què els temps de caigudes s'han de mantenir molt curts en un període anual. D'altra banda, FT fa referència a un sistema on no hi ha la possibilitat de perdre ni un sol minut de treball en producció, fet que implica tenir una infraestructura totalment redundant.

Una de les tècniques utilitzades per a l'alta disponibilitat és la replicació que habilita la còpia i el manteniment dels objectes en una base de dades distribuïda. També es coneix com *sharding*. Farem servir indistintament una paraula o una altra.

La taula 6 resumeix les diferents opcions disponibles i també què aporta cadascuna d'elles.

El sistema d'arxius distribuït també ha estat adoptat per les bases de dades relacionals, cosa que permet poder treballar en paral·lel, procés que es coneix com *Massive Parallel Processing* (MPP). Tenim exemples com: Teradata*, IBM Netezza, Pivotal Greenplum** o Oracle Exadata.

***www.teradata.com**
****http://greenplum.org**

Taula 6. Tecnologies de l'emmagatzematge.

Tecnologia	Descripció	Característiques	Productes	Cas d'ús
Sistema d'arxius distribuït	Sistema que proporciona emmagatzematge basat en la divisió de les dades en fitxers i servidors.	Proporciona redundància i alta disponibilitat per replicació. Accés seqüencial de dades. Per minimitzar les lectures de cerca a disc, així com el processament de molts fitxers, aquest tipus de sistemes agreguen les dades en fitxers de mida més gran.	Apache HDFS, Amazon S3 o Google File System	Arxivat de conjunts de dades. Emmagatzematge de dades en brut. Emmagatzematge de baix cost per a llargs períodes.
NoSQL	Sistema que proporciona emmagatzematge basat en una ordenació / representació no relacional.	En general complex: escalat horitzontal en lloc de vertical; alta disponibilitat, consistència eventual; BASE, no ACID i *actesharding*. Persistència políglota. Consultes distribuïdes.	MongoDB, Apache Cassandra, Riak, Redis, Neo4j o CouchDB	El model de negoci no pot representar-se de manera relacional. El model de negoci evoluciona ràpidament i necessita una base de dades flexible en el seu model.
NewSQL	Sistema NoSQL que combina propietats ACID.	A més de les característiques de NoSQL, inclou suport per a SQL i l'ús d'estructures relacionals.	VoltDB, NuoDB, Google Spanner o CockroachDB	Sistemes OLTP amb alt volum de transaccions. Analítica en temps real.
In-Memory	Ús de la memòria del processador per a l'emmagatzematge de dades.	Redueix la latència d'accés i de càlcul. Pot basar-se en *grid* o base de dades.	HazelCast, Pivotal Gemfire, Aerospike, MemSQL o Altibase HDB	Analítica operacional. BI Operacional. *Streaming Analytics*.

Dins de NoSQL, hi ha principalment quatre tipus de bases de dades:

Graf

Quan parlem de graf, fem referència a un conjunt d'objectes (anomenats vèrtexs o nodes) units per enllaços (anomenats arestes o arcs). Un graf permet estudiar les interrelacions entre els seus nodes.

- ***Key-Value Store***: l'emmagatzematge es fonamenta en l'ús de parelles clau-objecte en les quals no hi ha cap esquema. Exemples:

  – Apache HDFS,
  – Riak (http://basho.com),
  – Voldemort (http://www.project-voldemort.com),
  – Redis (http://redis.io),
  – RocksDB (http://rocksdb.org) o
  – Amazon DynamoDB.

- **Bases de dades orientades a columnes**: l'emmagatzematge de la dada es realitza per columnes, no per files. Exemples:

  – Apache Hbase (http://hbase.apache.org),
  – Apache Cassandra (http://cassandra.apache.org),
  – MonetDB (http://www.monetdb.org),
  – Druid (http://druid.io/),
  – Vertica,
  – Sybase IQ,
  – LucidDB o
  – Amazon SimpleDB.

- **Bases de dades de grafs**: utilitza nodes i vèrtexs per representar dades. Exemples:

  – Neo4J (http://neo4j.com),
  – HyperGraphDB (http://hypergraphdb.org),

– ArangoDB (http://www.arangodb.com),
– Ontotext GraphDB (http://ontotext.com) o
– OrientDB (http://orientdb.com).

- **Bases de dades orientades a documents**: l'emmagatzematge de la dada es realitza com si fos un document semi-estructurat. Exemples:

– MongoDB (https://www.mongodb.org),
– CouchDB (http://couchdb.apache.org) o
– MarkLogic (http://www.marklogic.com).

Per a algunes de les opcions disponibles, les distincions entre les diferents bases de dades s'estan diluint, ja sigui perquè una mateixa base de dades passa a ser multi-NoSQL (suportant-ne més d'un tipus), o perquè pertany a diverses categories al mateix temps. Exemples: ArangoDB combina grafs, documents i *key-value*; OrientDB combina grafs, documents, objectes i *key-value*. En general, estem parlant d'una alta especialització en el cas d'ús i, per tant, d'un escenari políglota en l'emmagatzematge de la dada.

7.2. Processament

La necessitat de processar dades no és un aspecte nou per a les organitzacions. En el passat, això s'ha abordat usant tècniques d'integració de dades, o *data integration*, com hem explicat en anteriors apartats. Encara que existeixen moltes tècniques d'integració, el processament de *big data* es fonamenta principalment en ELT (*Extract, Load, Transform*). És a dir, es posa el focus en guardar la dada en brut, amb el menor nombre de canvis, i el procés de transformació s'executa en cadascuna de les bases de dades (sigui quina sigui la seva tipologia).

Data integration

Quan parlem de *data integration*, fem referència al conjunt d'aplicacions, productes, tècniques i tecnologies que permeten una visió única i consistent de les nostres dades de negoci.

En línia amb els sistemes d'emmagatzematge, les principals aproximacions per al processament són:

- **Processament de dades en paral·lel**: vol dir que un procés es divideix en diverses tasques que s'executen en paral·lel. Tradicionalment, aquest enfocament s'ha realitzat amb una única màquina amb múltiples processadors o nuclis.

- **Processament de dades distribuïdes**: significa que el procés es divideix en múltiples tasques que s'executen en un clúster de màquines connectades en xarxa seguint la filosofia "divideix i venceràs".

Clúster

Quan parlem de clúster, fem referència al conjunt d'ordinadors connectats en xarxa que treballen de manera conjunta. Cada ordinador del clúster és anomenat node. Si els ordinadors són heterogenis, realitzen tasques independents o no estan en la mateixa localització, parlem de *grid*.

En el context de *big data*, per poder abordar les necessitats de treballar amb grans volums de dades i/o de capturar-les i consumir-les a diferents velocitats (des d'hores fins a fraccions de segon), han emergit diferents aproximacions:

- **Processament en mode *batch*, o per lots**: la dada es processa en mode *offline*. La seva latència pot anar des de minuts fins a hores. Abans de ser processada, la dada s'ha emmagatzemat prèviament. Apache MapReduce i Spark, aquest últim amb millors prestacions en termes de velocitat, permeten aquest tipus de processament.

  Hulu* és un servei de vídeo en *streaming* amb més de 5.5 milions de subscriptors i més de 20 milions de visitants únics per mes. Aquesta companyia fa servir MapReduce per processar els *logs* resultat de la visualització de més de 400 milions de vídeos al mes. L'objectiu és poder oferir un servei de *streaming* amb un nivell de qualitat consistent. És a dir, sempre disponible, des de qualsevol dispositiu i amb el nivell de qualitat de vídeo adequat al dipositiu.

  ***http://hulu.com**

- **Processament en mode *real time*, o en temps real**: la dada es processa en mode *online*. La seva latència està en el rang des de menys d'un segon fins al minut. Per això, la dada se processa en memòria en el moment de la seva captura, abans d'emmagatzemar-la. N'hi ha de dues menes: processament en flux (*stream*), en què la dada arriba de manera continuada, i processament per intervals o esdeveniments (*event*). Apache Storm, Apache Flink i Spark permeten aquest tipus de processament.

  MyFitnessPal* és un servei que permet conèixer el nombre de calories consumides i dóna suport al tractament de dietes. Aquesta companyia fa servir Spark per netejar, millorar i complementar les dades específiques de menjar introduïdes pels usuaris amb l'objectiu de tenir una base de dades de menjar / calories de màxima qualitat en temps real. És important per a aquest servei que sigui el més còmode i menys intrusiu per a l'usuari. A més, també s'aprofita de les capacitats de Spark per fer recomanacions.

  ***http://www.myfitnesspal.com**

El processament per intervals no és nou. Els sistemes CEP (*Complex Event Processing*) s'han fet servir des de fa anys en sectors com Banca o Energia per resoldre aquesta necessitat. En aquest tipus de sistemes, el rellevant no és processar tot el flux de dades sinó detectar aquells subconjunts que compleixen un patró. El sistema monitoritza el flux de dades i el compara amb els patrons definits. Per exemple, per detectar el frau en entitats financeres, el que és rellevant és detectar que un determinat client està realitzant un conjunt d'operacions sospitoses de cometre un frau.

CEP

Quan parlem de CEP, fem referència al processament d'esdeveniments en temps real que combina diverses fonts i que s'usa per inferir esdeveniments o patrons que suggereixen situacions complicades com oportunitats i/o amenaces.

En l'àmbit del processament fem referència al paradigma SCV. La relació dels tres components de SCV es mostra a la figura 33 i significa:

SCV

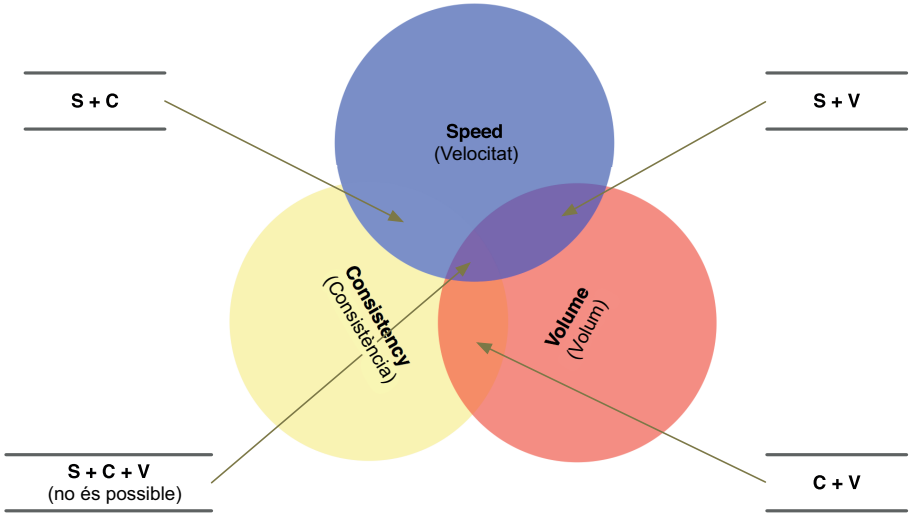
Quan parlem de SCV, fem referència a *Speed*, *Consistency* i *Volume*. És a dir, a la velocitat de processament, a l'exactitud de la dada i a la quantitat de dades processades.

- Si es necessita S i C, no és possible processar grans volums de dades perquè retarden el processament.

- Si es necessita C i V, no és possible treballar a una gran velocitat perquè el processament a gran velocitat requereix menors quantitats de dades.

- Si es necessita V i C, es consideren mostres (en lloc de treballar amb tot el conjunt de dades), fet que en reduirà la consistència.

Figura 33. SCV



Font: Josep Curto

Revisem un exemple que necessita dos processaments.

> Quan una empresa s'enfronta al frau, com pot ser en el sector de les finances, energètic o *retail*, té diverses necessitats. D'una banda, necessita fer una anàlisi forense de tot l'històric de transaccions per poder detectar nous patrons de frau. Aquesta necessitat pot considerar-se com un problema en el qual preval la capacitat de treballar amb tot l'historial i no la velocitat. Estem davant d'una necessitat que pot cobrir-se amb el processament i emmagatzematge *batch*.
>
> D'altra banda, també hi ha una altra necessitat un cop s'han reconegut els patrons. Aquesta necessitat, consisteix a analitzar el flux de transaccions en temps real i detectar si es compleix algun dels patrons. Aquí preval la velocitat de detecció de l'esdeveniment. Estem davant d'un escenari de processament en *streaming*.

7.3. Anàlisi

La creixent complexitat en la dada ha traspassat la capa de l'anàlisi, cosa que implica ajustar i modificar els diferents tipus d'anàlisi a la nova naturalesa de la dada.

Per distingir clarament els canvis en processament i emmagatzematge, hem separat el *data warehouse* i la integració de dades de la intel·ligència de negoci, malgrat que, en general, aquesta mena de sistemes no es conceben sense aquests components. No obstant això, *big data* obre la porta a una nova combinació i d'aquí la separació que estem considerant, ja que l'arquitectura per a l'emmagatzematge i el processament de dades pot arribar a ser més complexa del que era abans. Cal recordar els diferents components d'anàlisi de la intel·ligència de negoci:

- **Informes**: documents a través dels quals es presenten els resultats d'un o diversos processos de negoci, que es poden distribuir o simplement estar disponibles per al seu accés. Solen contenir text acompanyat d'elements com taules o gràfics per agilitzar la comprenssió de la informació presentada.

- **OLAP (*OnLine Analytical Processing*)**: mètode per organitzar i consultar dades sobre una estructura multidimensional.

- **Quadres de comandament (o *dashboard*)**: sistema que informa de l'evolució dels paràmetres fonamentals de negoci d'una organització o d'una de les seves àrees a través de components visuals integrats.

- ***Scorecards***: tipus de quadre de comandament format només per llistes d'indicadors. A vegades també pren la forma d'informe.

- **Consultes *ad hoc***: mètode que ofereix autoservei i exploració de dades a usuaris finals basats en metadades de negoci.

- **Alertes i monitorització automàtica**: sistema per crear, gestionar i distribuir alertes crítiques basades en indicadors clau de negoci amb focus en la gestió d'excepcions.

Els components anteriors suporten l'enfocament de planificació i control estratègic.

- **Quadre de comandament integral (o *balanced scorecard*)**: mètode de planificació estratègica, basat en mètriques i processos, ideat pels professors Kaplan i Norton, que relaciona factors mesurables de processos amb la consecució d'objectius estratègics.

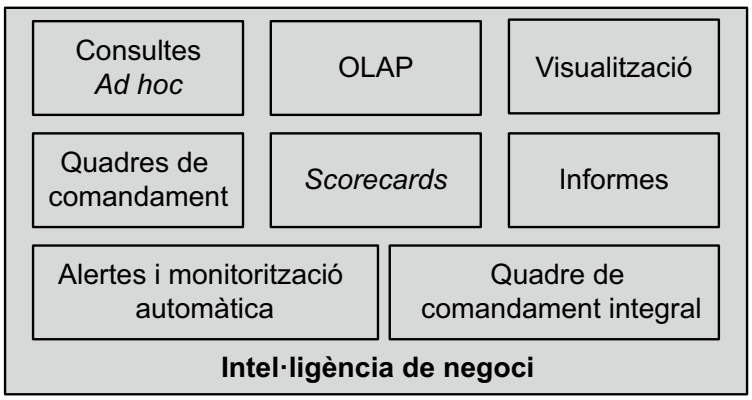
Una solució d'intel·ligència de negoci pot tenir un o diversos components. Les solucions més madures de mercat solen tenir-los tots en format modular. La implementació d'un o més components en una organització ha de dependre de les necessitats de negoci en l'organització i no de la plataforma, del proveïdor seleccionat o de les preferències de l'usuari de negoci o departament.

Lectura complementària

Kaplan, Robert S.; Norton, D. P. (1996). *The Balanced Scorecard: Translating Strategy into Action*. Boston, MA.: Harvard Business School Press.

La figura 34 il·lustra els components d'anàlisi de la intel·ligència de negoci.

Figura 34. Intel·ligència de Negoci



Font: Josep Curto

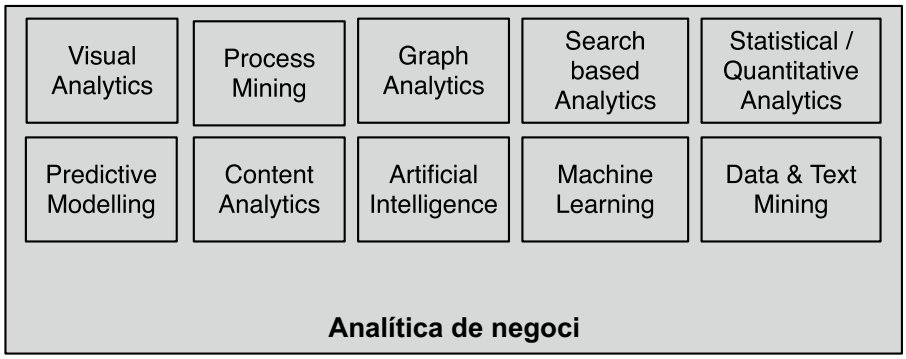
Considerem un exemple de com es combina amb *big data*.

Jagex* és una companyia de videojocs per a mòbils que suporten milions d'usuaris jugant al mateix temps. Per a aquesta companyia és absolutament primordial comprendre els seus clients: aquells que paguen, aquells que es donen d'alta, quins productes virtuals compren i fan servir en cadascun dels videojocs, i poder analitzar aquesta informació tant temporalment com geogràficament. Per fer això, s'han combinat les capacitats d'emmagatzematge i processament de *big data* amb les capacitats d'anàlisi en format quadre de comandament i informes de la intel·ligència de negoci per tenir control del negoci.

***http://www.jagex.com**

En l'analítica de negoci, hem introduït els diferents tipus d'anàlisi existents resumits en la figura 35, que il·lustra els components en l'analítica de negoci.

Figura 35. Analítica de Negoci



Font: Josep Curto

Considerem un altre exemple de com es combina amb *big data*.

Supercell* és una companyia de videojocs per a mòbils que suporten milions d'usuaris jugant al mateix temps. Entre els seus èxits destaca *Clash of Clans*. Aquesta companyia fa servir la combinació de tecnologies d'emmagatzematge de *big data* i analítica de negoci per validar hipòtesis de negoci. Un dels test A / B realitzats ha estat per decidir si valia la pena afegir la connectivitat de Facebook, en esbrinar si els usuaris fan servir aquesta possibilitat tant per convidar els seus amics com per compartir els seus èxits, i si això incideix en la retenció de l'usuari.

***http://www.supercell.com**

Les taxonomies presentades tenen una raó de ser. La principal diferència de la intel·ligència i l'analítica de negoci tradicionals pel que fa a *big data* és que cada component s'ha hagut d'adaptar. D'una banda, en el context de la intel·ligència de negoci això succeeix:

- A través de connectors per a l'ús dels sistemes d'emmagatzematge i processament de *big data*, com ara fonts de dades per al sistema d'intel·ligència de negoci.

- A través de l'adaptació de la tecnologia a la complexitat de la dada. Tenim, per exemple, Apache Kylin*, creat per eBay, que proporciona OLAP per *big data*; Hue**, creat per Cloudera, que permet visualitzar consultes *ad hoc* sobre Hadoop; o Caravel*** d'Airbnb amb focus en l'exploració de dades.

***http://kylin.apache.org**
****http://gethue.com**
*****http://github.com/airbnb/caravel**

Els fabricants tradicionals com IBM, Microsoft, Microstrategy, Oracle o Information Builders també s'estan posicionant en aquest mercat, creant la seva

pròpia proposta integrada i/o mitjançant connectors específics per a la seva plataforma.

D'altra banda, en el context de l'analítica, tenim també que les solucions i llibreries ja existents s'estan adaptant d'una manera semblant a la intel·ligència de negoci. Adaptar, en aquest cas, es tradueix en crear noves versions de l'algorisme que encapsula una certa tècnica perquè pugui aplicar-se a un conjunt de dades complexes, pugui escalar i, sobretot, tingui sentit des d'un punt de vista estadístic i matemàtic. Per això, el gran canvi resideix en l'aparició de noves llibreries de *machine learning*, *graph analytics* i *deep learning* adaptades a *big data*.

Lectura complementària

Leskovec, J.; Rajaraman, A.; Ullman, J. (2016). *Mining of Massive Datasets segona edició*.

7.4. Visualització

Tradicionalment, els components d'intel·ligència de negoci, com els quadres de comandament, informes i/o vistes OLAP, s'han fet servir per presentar el resultat de l'anàlisi de la informació. Amb l'adveniment de *big data* i la combinació de tecnologies, aquest enfocament ja no és suficient. Dues disciplines han emergit per ajudar en la visualització de la informació: *Data Visualization* (Visualització de dades) i *Data Storytelling* (Històries fonamentades en dades). Hem de comprendre primer aquests conceptes.

S'entén per *Data Visualization* la representació de dades que explota les habilitats visuals per amplificar els processos cognitius.

Data Visualization persegueix incrementar les capacitats exploratòries i explicatives, representar grans volums de dades i comprendre les relacions ocultes en les dades de forma visual. Ha aparegut una gran col·lecció de llibreries especialitzades en aquest àmbit*. Entre aquestes llibreries i eines destaquen:

***http://selection.datavisualization.ch**

- D3.js (http://d3js.org),
- Polimaps (http://polymaps.org),
- Processing.js (http://processingjs.org),
- Grafana (http://grafana.org),
- Tableau (http://www.tableau.com),
- Qliksense (http://www.qlik.com),
- CartoDB (http://cartodb.com) o
- Yellowfin (http://www.yellowfinbi.com).

Lectura complementària

Few, S. (2009). *Now You See It: simple Visualization Techniques for Quantitative Analysis*, Analytics Press.

D'altra banda, s'entén per *data storytelling* el mètode visual de presentar informació per fer-la més comprensible i fàcil de comprendre.

A l'actualitat, algunes eines propietàries, com les que ofereixen Tableau, Qliksense, Quadrigram*, Miso**, TimelineJS*** o Yellowfin, capaciten les organitzacions per a l'ús de *data storytelling*, si bé també és possible crear-lo de forma programàtica.

*http://www.quadrigram.com
**http://misoproject.com
***http://timeline.knightlab.com

Aquestes tècniques no només tracten d'usar la millor representació per explicar el que passa, sinó que, a més, han de poder-se connectar amb els components de processament i emmagatzematge de *big data*. No només es tracta de tenir la tecnologia adequada (escalable i adaptable a *big data*), sinó de dominar la comunicació de la informació. Com comenta Stephen Few, les capacitats per mostrar i explicar la informació de manera efectiva no són intuïtives i cal aprendre uns nous principis:

Lectura complementària

Segel, E.; Heer, J. (2010). *Narrative Visualization: Telling Stories with Data*, IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis).

- Conèixer l'audiència de la visualització. Això inclou factors com el paper, el flux de treball, el coneixement tècnic i de negoci de l'audiència.

- Determinar el valor que es vol proporcionar a l'audiència. En aquest sentit tenim dues grans opcions. Tenim ja una pregunta per respondre, o bé estem fent una anàlisi exploratòria. Això inclou identificació del que és rellevant, establiment de metes i expectatives.

- Seleccionar la visualització correcta. Això inclou l'elecció del gràfic i/o la representació*, l'abast, l'horitzó temporal i el tipus de decisions.

*http://extremepresentation.com

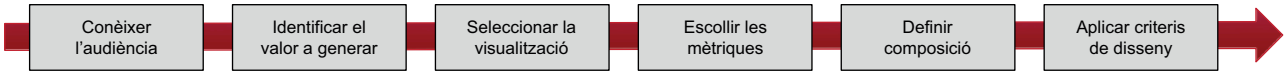
- Escollir les mesures adequades que han d'ajudar sempre a prendre decisions.

- Creació / Composició de la visualització, que ha de tenir en compte la forma, l'estructura, la funcionalitat i els principis de disseny.

- Ús de criteris de disseny i presentació d'informació, com l'elecció de colors i tipografia.

Aquests principis s'il·lustren a la figura 36.

Figura 36. Principis visualització



Font: Josep Curto

Resum

En aquest mòdul didàctic hem presentat els conceptes de *business intelligence*, *business analytics* i *big data*, que fonamentalment habiliten les organitzacions per generar valor a partir de les seves dades.

Pel que fa a *business intelligence*, hem presentat la seva definició, quan cal emprar-lo, quins beneficis aporta i les tecnologies que formen part d'aquesta estratègia. Hem entrat en detall en com es gestiona i s'explota la dada.

Quant a *business analytics*, hem presentat la seva definició i una taxonomia de les tecnologies que formen part d'aquesta estratègia.

Pel que fa a *big data*, hem presentat la seva definició, els tipus que existeixen, quins beneficis aporta, quan és necessari aplicar-la i les tecnologies que formen part d'aquesta estratègia. I sobretot s'ha posat de manifest com de diferent és i per què complementa la intel·ligència de negoci i l'analítica de negoci.

També hem anat introduint exemples. Tal com s'ha mostrat en aquests materials, hi ha moltes organitzacions que ja han desplegat aquest tipus de sistemes d'informació i han aconseguit rendiments de l'explotació de conjunts de dades complexes.

Glossari

ACID Estàndard *de facto* de les bases de dades relacionals. És l'acrònim d'*Atomicity* (Atomicitat), *Consistency* (Consistència), *Isolation* (Aïllament) i *Durability* (Durabilitat).

API Conjunt de subrutines, funcions i procediments (o mètodes, en la programació orientada a objectes) que ofereix certa biblioteca per ser utilitzat per un altre *programari* com una capa d'abstracció.

BASE Estàndard per a les tecnologies *big data*. És l'acrònim de *Basically Disponible* (bàsicament disponible), *Soft state* (estat tou) i *Eventual consistency* (consistència eventual).

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

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

byte Unitat de mesura d'informació digital.

CAP Estàndard *de facto* dels sistemes distribuïts. És l'acrònim de *Consistency* (consistència), *Availability* (disponibilitat) i *Partition Tolerance* (tolerància a la partició).

CEP Processament d'esdeveniments en temps real que combina múltiples fonts i que s'usa per inferir esdeveniments o patrons que suggereixen situacions complicades, com oportunitats i/o amenaces.

clúster Conjunt d'ordinadors connectats en xarxa que treballen de forma conjunta. Cada ordinador del clúster és anomenat node.

CRM Acrònim de *Customer Relationship Management*; fa fa referència a la gestió de la relació amb clients.

crowdsourcing Procés d'obtenir serveis, idees i contingut a través de la participació d'una gran massa de persones.

data integration Conjunt d'aplicacions, productes, tècniques i tecnologies, que permeten una visió única i consistent de les nostres dades de negoci. També denominada com integració de dades.

data warehouse Repositori de dades que proporciona una visió global, comuna i integrada de les dades de l'organització, independentment de com s'utilitzin posteriorment pels consumidors o usuaris; amb les propietats següents: estable, coherent, fiable i amb informació històrica.

equacions diferencials Equació matemàtica que relaciona una funció i les seves derivades.

edge analytics Aplicacions analítiques per IoT en què certs algoritmes s'executen en els nodes de la xarxa i no només al centre de dades.

ERP Acrònim d'*Enterprise Resource Planning*; fa referència a la gestió dels recursos d'una organització.

escalabilitat Habilitat d'un sistema, xarxa o procés per reaccionar i adaptar-se sense perdre qualitat, o bé gestionar el creixement continu de treball de manera fluïda, o bé per estar preparat per fer-se més gran sense perdre qualitat en els serveis oferts.

escalabilitat horitzontal Escalabilitat fonamentada en l'increment de nodes del sistema, procés o xarxa.

escalabilitat vertical Escalabilitat fonamentada en afegir més recursos –memòria, disc dur i/o processadors–.

estructura de dades relacional Tipus de base de dades que permet establir interconnexions o relacions entre les dades desades en taules.

grid Conjunt d'ordinadors connectats en xarxa que treballen de forma conjunta, però, a diferència del clúster, els ordinadors són heterogenis, realitzen tasques independents o no estan en la mateixa localització.

HPC Pràctica d'afegir capacitat de computació de manera que millora el rendiment d'una estació de treball i fa possible abordar problemes complexos en la ciència, enginyeria i/o negocis.

Internet de les Coses Interconnexió digital d'objectes quotidians amb Internet. Ens referirem a ell pel seu acrònim en anglès IoT, *Internet of Things*.

latència Suma de retards temporals en la captura, emmagatzematge, processament i anàlisi de la dada.

metadades Dades estructurades i codificades que descriuen característiques d'un objecte, dada o procés de negoci.

NIST Acrònim de *National Institute of Standards and Technology*, una institució americana que estudia, defineix i promou estàndards tecnològics.

NPL Acrònim de *Natural Processing Language*. Fa referència a un camp de les ciències de la computació, intel·ligència artificial i lingüística que estudia les interaccions entre els ordinadors i el llenguatge humà.

NoSQL Acrònim de *Not Only SQL*. Fa referència a bases de dades no relacionals.

OLAP Mètode per organitzar i consultar dades sobre una estructura multidimensional. És l'acrònim d'*Online Analytical Processing* o procés analític en línia.

open data Conjunts de dades considerades un bé comú i que, per això, són gratuïts, accessibles i ben estructurats per a la seva descàrrega i anàlisi.

OWL Acrònim de *Web Ontology Language*. Fa referència a un estàndard per al disseny d'ontologies de models de dades.

PMML Acrònim de *Predictive Model Markup Language*. Fa referència a un estàndard per a l'intercanvi de dades entre organitzacions.

quadre de comandament Sistema que informa de l'evolució dels paràmetres fonamentals de negoci d'una organització o d'una de les seves àrees.

RIF Acrònim de *Rule Interchange Format*. Fa referència a un estàndard per a l'intercanvi de dades entre organitzacions.

SCV Fa referència a *Speed*, *Consistency* i *Volume*. És a dir, a la velocitat de processament, a l'exactitud de la dada i a la quantitat de dades processades.

SLA Acord que estipula el nivell de servei, el suport, possibles penalitzacions, el nivell d'alta disponibilitat, tant de *maquinari* com de *programari*, i el preu.

SQL Acrònim de *Structure Query Language*; fa referència al llenguatge de consultes de bases de dades relacionals.

taxonomia Classificació o ordenació en grups de coses que tenen unes característiques comunes.

UIMA Acrònim d'*Unstructured Information Management Architecture*. Fa referència a un estàndard que permet la interoperabilitat analítica de dades en informació no estructurada.

variabilitat Fa referència al fet que els fluxos de dades poden tenir comportaments erràtics o inconsistents en certs períodes.

velocitat Fa referència tant al processament de dades com a la seva latència.

varietat Fa referència tant a la quantitat de fonts diferents que s'han de combinar com a l'heterogeneïtat de la dada.

veracitat Fa referència a la incertesa en la dada producte de la seva baixa qualitat, l'ambigüitat en la seva definició o simplificacions en la seva modelització.

vinculació Dificultat de relacionar diferents i diverses fonts de dades.

volum Mida del conjunt de dades creat diàriament.

XBRL Acrònim d'*eXtensible Business Reporting Language*. Fa referència a un estàndard per a informes financers.

Bibliografia

Corr, L.; Stagnitto, J. (2011). *Agile Data Warehouse Design: Collaborative Dimensional Modeling, from Whiteboard to Star Schema*. Leeds: DecisionOne Press

Davenport, T.H. (2014). *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*. Boston: Harvard Business Review Press.

Davenport, T.H.; Harris, J.G. (2007). *Competing on Analytics: The New Science of Winning*. Nova York: Harvard Business Press.

Davenport, T.H.; Kim, J. (2013). *Keeping Up with the Quants: Your Guide to Understanding and Using Analytics*. Boston: Harvard Business Review Press.

Erl, T.; Khattak, W.; Buhler. P. (2015). *Big Data Fundamentals: Concepts, Drivers & Techniques*. New Jersey: Prentice Hall

Fisher, T. (2009). *The Data Asset: How Smart Companies Govern Their Data for Business Success*. New Jersey: Wiley

Foreman, J.W. (2013). *Data Smart: Using Data Science to Transform Information into Insight*. New Jersey: Wiley.

Howson, C. (2013). *Successful Business Intelligence, Second Edition: Unlock the Value of BI & Big Data*. New York: McGraw-Hill Education

Kimball, R.; Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. New Jersey: Wiley.

Malcolm, F.; Roehrig, P.; Pring, B. (2014). *Code Halos: How the Digital Lives of People, Things, and Organizations Are Changing the Rules of Business*. New Jersey: Wiley

Redman, T. C. (2008). *Data Driven: Profiting from Your Most Important Business Asset*. Boston: Harvard Business Review Press.

Schmarzo, B. (2016). *Big Data MBA: Driving Business Strategies with Data Science*. New Jersey: Wiley.

Schmarzo, B. (2013). *Big Data MBA: Understanding How Data Powers Big Business*. New Jersey: Wiley.