Estudi i Implementació d'una solució Business Intelligence per a una empresa

Marc De Tébar Peralta

Resum—Actualment, les tecnologies de la informació es troben en constant evolució, i gestionar les dades de forma eficient en emmagatzematge i accés mitjançant eines especialitzades, ha passat de ser un element complementari a formar part activa i principal de les empreses, esdevenint un element fonamental per a oferir un millor servei als seus clients, i també, per a diferenciar-se de la competència. La implementació del sistema Business Intelligence, ha millorat l'organització de les dades i la seva qualitat, proveint, mitjançant les eines ETL, una neteja de les dades, i un flux visual i disseny estructurat del sistema. Els diferents elements necessaris per a dur a terme una tasca d'aquestes característiques són les dades, una base de dades en format estrella i una eina de BI, tots aquests elements han d'ésser obtinguts, creats i gestionats de la forma apropiada. Aquesta eina inclou la capacitat de realitzar i gestionar processos ETL i cubs multidimensionals. La realització d'un conjunt d'anàlisis empresarials, mitjançant la creació de diferents cubs de dades multidimensionals, ha permès assistir, agilitzar i millorar la presa de decisions executives en situacions empresarials on els factors a analitzar són molt diversos i complexos.

Paraules clau—Procés ETL, Intel·ligència de negoci, Cubs multidimensionals, cubs OLAP, Base de dades, Tecnologies de la Informació i la Comunicació, Anàlisi empresarial, Base de dades en estrella, Sistemes d'Informació.

Abstract—Nowadays, IT technologies are in constant evolution, and to handle the data in an effective way, in order to store and access it, has gone from being a complementary element to having a fundamental, active role in business. IT technologies have become the key to offer a better service to clients, and to make the difference with competitors. Business Intelligence Implementation has improved the data organization and its quality, and ETL processes have provided the cleaning of the data flow, a structured design and a visual system. The necessary elements to carry out this task are: data, star schema data base, and a Business Intelligence tool with the ability to carry out and manage ETL processes and multidimensional cubes, all these elements must be obtained, created and managed by the appropriate form. A set of business analyses were carried out through the creation of different multidimensional cubes which allowed the speeding up, attending to and improving of the decisions that must be taken in work situations where factors that need to be analysed are diverse and complex.

Index Terms—ETL Process, Business Intelligence, Multidimensional cubes, OLAP cubes, Databases, IT Technologies, Business Analysis, Star schema database, Information Systems.

1 Introducció

AQUEST treball es focalitza en la Intel·ligència de Negoci - *Business Intelligence (BI)* - , es duu a terme una implementació d'un sistema d'aquestes característiques en una empresa per assistir a la presa de decisions executives, determinant, quina és la millor solució a aplicar i el perquè. A més a més, com que en aquest cas, l'empresa és hipotètica, també es simula un entorn de treball, conformat per les dades emprades, que han estat obtingudes a partir de l'Institut d'Estadística de Catalunya (*IdesCat*) i de Dades Obertes de la Generalitat de Catalunya, i també l'entorn de la base de dades, creat en format estrella per a assistir els informes i les anàlisis que es realitzen.

L'objectiu principal d'aquest treball és el d'explicar de forma clara i entenedora quin és el paper dels BI dins dels Sistemes d'Informació, quins en són els seus principals components, tecnologies i fases, però també, que és el que poden aportar en el context empresarial, és a dir, quin és l'element de competitivitat que proveeixen i quins són els problemes que es resolen si s'implementen.

Aquests objectius doncs, es sintetitzen en explicar els passos o fases a seguir per a implementar un sistema d'aquestes característiques, així com també, les eines de què disposen, gestionar i organitzar una base de dades enfocada al treball amb sistemes de *Business Intelligence* i entendre'n les seves peculiaritats, realitzar una anàlisi de requeriments de la situació de partida de l'empresa i una comparativa de solucions BI per a determinar quina és la que s'adequa més a les necessitats requerides. Finalment, implementar la solució i utilitzar les eines *Extract – Transform – Load* (*ETL*) per a tractar les dades, i ,en darrer terme, realitzar l'anàlisi multidimensional de les dades mitjançant el que s'anomenen cubs *OLAP*, i demostrar com aquests assisteixen a la presa de decisions empresarials.

En aquest article, es segueix tot el procediment realitzat,

- *E-mail de contacte: marc.detebar@e-campus.uab.cat*

- *Menció realitzada: Tecnologies de la Informació.*
- *Treball tutoritzat per: Ramón Musach Pi (Departament d'Enginyeria de la Informació i les Comunicacions)*
- *Curs 2014/15*

de forma ordenada, així doncs les seccions s'organitzen en la Introducció, Estat de l'art, Metodologia, Desenvolupament del Treball, Resultats, Conclusió, Agraïments i Bibliografia.

2 ESTAT DE L'ART

Els BI són un conjunt d'eines que tenen la finalitat de millorar la presa de decisions d'una empresa, mitjançant el tractament i organització correcte de les dades. El concepte va ser introduït per primera vegada l'any 1958 per - *H.P. Luhn* - [1], en un article on es definien quines havien de ser les característiques dels BI, anomenat - *A Business Intelligence System* - .

La primera idea del que acabaria esdevenint un sistema de BI va sorgir l'any 1969 quan – *Codd* - [2], va crear ni més ni menys que el concepte - base de dades - . A la dècada dels 70 van implementar-se els primers sistemes de Base de Dades, però no va ser fins la dècada següent, als anys 80, quan van sorgir els primers sistemes de *reporting*.

Van haver de passar 10 anys més per a tenir un sistema de BI formalitzat, amb l'aparició de diferents aplicacions, incloent-hi les cerques multidimensionals mitjançant cubs *OLAP*. No obstant això, no va ser fins l'any 2000 quan aquests sistemes varen consolidar-se en unes plataformes concretes que encara s'utilitzen avui en dia, com ara *SAP* i *Oracle*, on es focalitza en el *Data Mining* , per a detectar patrons de comportament, una cosa inèdita fins al moment.

3 METODOLOGIA

3.1 Metodologia emprada

La metodologia emprada per a realitzar aquest treball es basa en el model en cascada, en anglès, *waterfall model*. Aquest model comprèn les fases d'Anàlisi de Requeriments, Disseny, Codificació, Testeig i Manteniment.

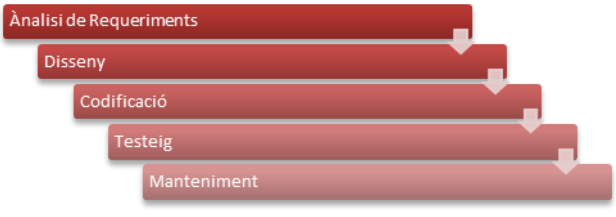

Fig. 1. Fases del Model en Cascada.

Ha estat l'escollit perquè s'adequa a les necessitats del projecte, cal tenir en compte que una modificació en el disseny prèviament realitzat, fa canviar la forma en què han estat dissenyats els cubs, i també la posterior anàlisi i el manteniment. També, canvis en l'Anàlisi de Requeriments, impliquen modificacions en el disseny, i en conseqüència, en les següents etapes.

3.2 Context de l'empresa

Per a simular l'operativa d'una empresa, han estat emprades un conjunt de dades obtingudes a través de l'administració catalana, concretament de Dades Obertes de la Generalitat de Catalunya ,i de l'*IdesCat*. El fet de recórrer a les dades obertes, ha permès obtenir un gran volum de dades, necessari per a simular l'operativa de treball i el context d'una empresa real.

D'aquesta manera doncs, partint d'un context prou rellevant i del què es disposa de molta informació, com són les eleccions al Parlament de Catalunya en el període temporal del 1988 fins al 2012, la informació ha estat complementada amb fitxers addicionals de dades d'atur i població, que han permès afegir un punt de visió addicional, i acabar de conformar el que esdevé informació dispersa de l'empresa a tractar i a integrar en un sistema de *BI*.

Així doncs, s'implementa un sistema BI, que permet englobar tot aquest conjunt de dades, disperses en 18 fitxers diferents, i gestionar-les i emmagatzemar-les d'una forma més eficient i estructurada, que permet un accés més ràpid i efectiu, sense penalitzar en temps la posterior realització dels informes i les anàlisis.

3.3 Requeriments

A partir del context actual de l'empresa, i dels conceptes sobre BI, s'han establert un conjunt de requeriments, per ajudar a concretar com ha de ser aquesta eina de BI. Els requeriments es troben classificats en:

Requeriments Funcionals

- El sistema ha de proveir a l'administrador amb l'habilitat d'afegir orígens de dades per a treballar a partir d'aquestes.
- El sistema ha de proveir a l'administrador amb l'habilitat de configurar una connexió de bases de dades.
- El sistema ha de proveir a l'usuari amb l'habilitat de configurar els paràmetres necessaris per a la realització de cubs OLAP.
- El sistema ha de permetre exportar diferents informes en els formats especificats.

Requeriments No Funcionals

- El procés d'exportació dels informes realitzats no pot superar el de generació.
- L'aplicatiu hauria de poder utilitzar-se en diferents plataformes (multi-plataforma).
- L'aplicatiu hauria de poder-se personalitzar en la llengua de l'usuari.

Restriccions de sistema

- El servidor de BI, ha de funcionar en la plataforma Windows.
- El format de la base de dades, ha de ser en estrella.

3.4 Gestió de la base de dades

Per a poder gestionar i crear la base de dades s'ha emprat un servidor *MySQL* local, conjuntament amb l'aplicació *MySQL Workbench CE*.

Una base de dades, que pugui gestionar i organitzar aquest gran nombre de taules, i permeti realitzar una anàlisi *OLAP*, ha d'estar estructurada en format estrella, establint una taula de fet.

En un primer terme es disposa, per a cada un dels comicis, de dues taules: electes i participació [3] , una taula global de la població [4] i una altra, també de global, de l'atur [5] , totes en format .csv o bé .xlsx.

Per a organitzar la base de dades, finalment s'han utilitzat les dimensions: atur, població i electes, perquè permeten analitzar la informació des de diferents perspectives. A més a més, la taula de fet, conté informació relativa a la participació, perquè d'aquesta forma, aquesta informació és utilitzada per dur a terme les mètriques, en el procés de creació de cubs *OLAP*. L'estructura final de la base de dades queda de la següent forma:

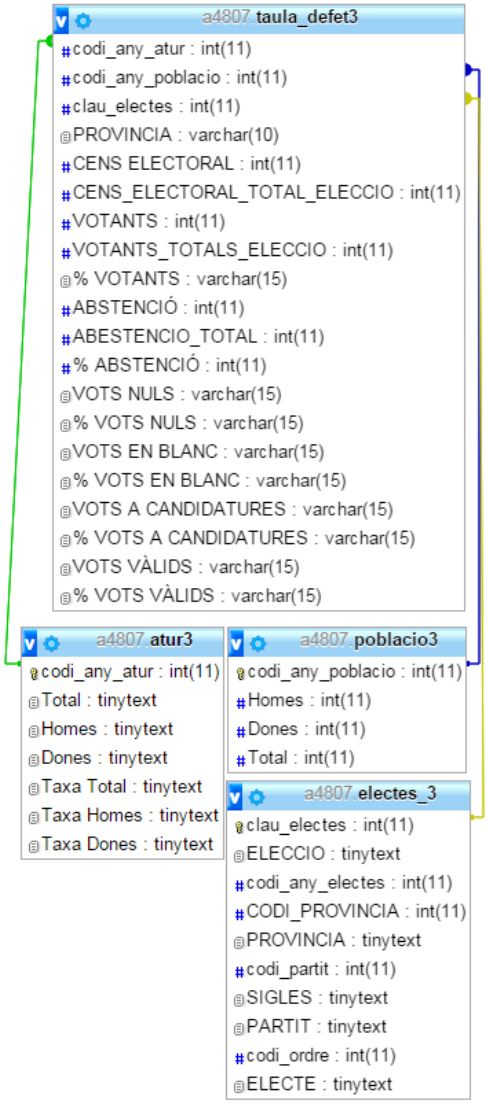

Fig. 2. Diagrama de l'estructura de la base de dades.

3.5 Comparativa d'eines

Actualment, en el mercat, hi ha un gran ventall de possibilitats pel que fa a eines de BI, algunes de les més importants són *Pentaho, Qlik View, Jedox* i *SpagoBI*. Totes tenen els seus punt forts, però fent una comparativa d'aquestes en relació als requeriments establerts en l'anàlisi de requeriments, s'observa que *Pentaho* és la que millor satisfà les necessitats de l'empresa.

Això és així perquè disposa d'eines de gestió de dades *ETL* [6] i permet realitzar i treballar amb cubs *OLAP*, i tot això en la seva versió gratuïta, anomenada, *Community Edition* [7]. En canvi, *Qlik View*, té una gran mancança, i és que no disposa d'eines *OLAP*, sinó que es basa en la lògica associativa, realitzant les anàlisis en memòria [8], i *Jedox*, en la seva versió gratuïta, no disposa d'eines *OLAP*, però tampoc permet exportar els informes realitzats [9]. En darrer terme, hi ha *Spago BI*, és molt similar a *Pentaho*, però no és tan utilitzada ni tampoc disposa de responsables oficials en cas de fallada [10].

3.6 Pentaho Business Intelligence Server CE

L'eina escollida per a dur a terme la implementació ha estat *Pentaho*, principalment perquè satisfà els requeriments establerts en l'anàlisi de requeriments, sobretot gràcies al fet que *Pentaho* disposa d'eines per a la gestió de les dades *ETL* i per a realitzar cubs *OLAP*. A més a més, es tracta d'una eina generalment utilitzada per les empreses, i les mancances que hi ha a la *Community Edition*, poden ésser cobertes, en certa mesura, per aplicacions addicionals disponibles al *Marketplace* de *Pentaho,* localització on es troben diferents eines que proveeixen funcionalitats addicionals.

La suite de *Pentaho* s'organitza en diferents aplicacions per a donar cobertura a la realització d'aquest conjunt d'operacions, les aplicacions emprades són:

- ***Business Analytic Platform Community Edition:*** És en essència el servidor de *BI*, on són publicats els cubs *OLAP* i on es realitzen els diferents informes i les diverses anàlisis de les dades.

> ***Saiku Analytics:*** Si bé no és una aplicació de *Pentaho*, és una eina disponible en el seu *Marketplace*, i permet, analitzar la informació dels cubs de forma més visual que no pas l'analitzador que duu per defecte *Pentaho Community Edition* [11].

-***Pentaho Data Integration:*** És l'eina que permet realitzar tot el procés de gestió de les dades, seguint el procediment *ETL - Extract, Transform, Load* - , s'acostuma a anomenar *Kettle*.

-***Schema Workbench:*** Aquesta eina permet, mitjançant una connexió a la base de dades, realitzar un cub *OLAP* per a publicar-lo al servidor de *BI* i poder realitzar-ne una anàlisi posterior.

4 DESENVOLUPAMENT DEL TREBALL

4.1 Pentaho Data Integration

Pentaho Data Integration és l'eina de *Pentaho* que s'empra per a poder realitzar el procés *ETL*, que comprèn les fases d'extracció de la informació, i la transformació i càrrega de les dades a la base de dades. L'entorn de treball d'aquesta aplicació permet realitzar transformacions i treballs. Si bé una transformació són el conjunt d'operacions que es realitzen, un treball permet englobar diferents transformacions, per exemple, per a executar-les totes d'una sola tirada, i no haver d'anar manualment d'una en una.

En aquest treball, s'ha decidit realitzar una transformació per cada tipus de taula, per a tenir-ho tot ben organitzat i separat en diferents fitxers, en aquestes transformacions, poden realitzar-se diferents operacions, cadascuna de les quals, fa referència a l'extracció, transformació o càrrega de les dades.

4.1.1 Extracció

Depenent de la tipologia de les dades, poden utilitzar-se diferents operacions per a extreure la informació dels fitxers. En aquest cas, d'acord amb el format dels fitxers originals, han estat utilitzades les següents operacions:

Microsoft Excel Input: Permet establir un fitxer d'entrada, en format Excel, a partir del qual són obtingudes les dades. En aquest pas, es poden eliminar columnes, es pot canviar el nom de les columnes originals, i també, el tipus de cadascun dels atributs (*Integer, Numeric, Boolean*...).

CSV file input: De la mateixa manera que l'element anterior, permet obtenir el contingut d'un fitxer, però en aquest cas el format del fitxer origen és *csv*. L'operativa de funcionament es basa sobre els mateixos principis, tot i que és més senzilla, les funcionalitats que permet, són també la capacitat d'eliminar columnes, canviar el nom de les columnes originals, i el tipus de cadascun dels atributs (Integer, Numeric, Boolean...).

4.1.2 Transformació

Poden realitzar-se diferents operacions sobre les dades obtingudes en el pas anterior, referent a l'extracció. En aquest cas, tenint en compte el context de l'empresa, les operacions de transformació que permeten dur a terme una millor gestió de les dades, són les següents:

Check if a column exist: Com que abans de realitzar la càrrega de dades, es crea l'estructura de la taula a la base de dades, el que permet realitzar aquesta operació és comprovar si una columna important per a l'estructuració o organització, existeix o no realment, en una taula de la base de dades en el moment de realitzar la transformació. Així doncs, retorna dos possibles resultats *true* o *false*, de manera que cadascun durà a un estat diferent, si és *true*, es segueix amb una nova acció, si és *false* també, però aquesta nova acció és d'error, perquè la transformació no prossegueixi.

Filter Rows: Permet realitzar comprovacions i comparacions sobre els camps, d'aquesta manera, és utilitzat per comprovar que les claus primàries no siguin nul·les, o bé comparar, en cas que dos atributs per regles d'integritat hagin de ser els mateixos, que efectivament tinguin el mateix contingut. Retorna, *true* si la condició és certa, per a seguir el progrés normal i anar a una altra operació, o *false*, estat que duu a una altra operació que llança un error.

Abort: En cas que, algunes de les operacions anteriorment exposades retorni *false*, s'enllacen a aquesta operació de transformació, que permet llançar un missatge d'error personalitzat, que es mostra en el *logging* de l'aplicació.

Add Sequence: Permet, donat un atribut, inserir el valor d'aquest per a cadascun dels registres que contingui a la taula. Aquest pas ha estat emprat amb la finalitat de crear una clau subrogada, és a dir, una clau numèrica en forma de seqüència, sense tenir en compte el contingut de la taula, aquesta operació ha estat realitzada en la transformació de la dimensió electes.

Append Streams: Aquest element s'utilitza per a unificar el contingut de dos fitxers de forma automàtica, establint en quin ordre han d'ésser unificats, és molt útil per a ajuntar els diferents fitxers d'electes de què es disposa.

4.1.3 Càrrega

Depenent de la base de dades on s'emmagatzema la informació, hi ha diverses maneres possibles de realitzar la càrrega final de les dades. En aquest cas, donat el tipus i format de la base de dades, l'operació que ha permès efectuar aquesta acció de manera efectiva, ha estat la següent:

Insert/Update: Permet inserir o actualitzar elements en una taula, establint les claus necessàries per a dur a terme l'actualització, i els camps que han d'ésser actualitzats. Si s'executa en primera instància, s'ha de clicar el botó *SQL* per a que generi l'estructura de la taula a la base de dades, en altre cas, la transformació no prossegueix.

4.2 Transformacions realitzades

A continuació, es mostren les transformacions realitzades sobre cadascuna de les taules, per a aplicar el procés *ETL*, i realitzar la càrrega d'informació a la base de dades.

4.2.1 Taula de fet

Aquesta transformació, permet extreure les dades que conformaran la taula de fet, i tractar-les, i carregar-les a la base de dades. Els passos que se segueixen per a fer-ho són, primer de tot, dur a terme l'extracció, tot seguit, determinar si diferents columnes essencials per a la taula, són realment a la l'estructura creada a la base de dades. Això és així perquè abans d'executar la transformació, en els paràmetres de l' *Insert/Update* , s'ha de clicar el botó *SQL*, per a generar i executar el codi de creació de l'estructura de la taula a la base de dades.

Posteriorment, es comprova que els atributs que contenen valors enters, no siguin nuls, i que es compleixin les igualtats corresponents.

Finalment, si no s'ha produït cap error, es procedeix a la inserció/actualització del contingut a la base de dades. Si s'hagués produït algun error, en qualsevol dels passos anteriors, s'hagués anat a parar a l'element *Error* perquè la transformació no prosseguís.

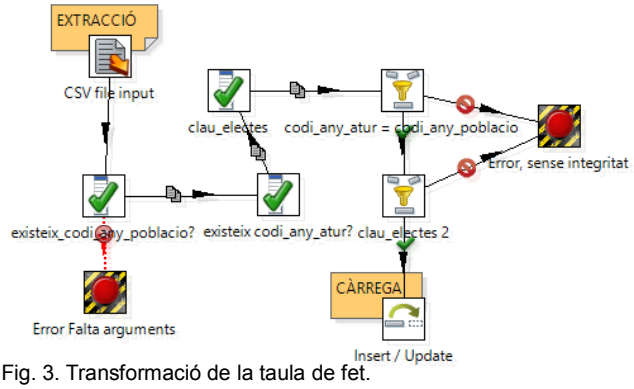

Fig. 3. Transformació de la taula de fet.

4.2.2 Atur

En la transformació Atur, també ben diferenciada en els tres passos que determina el procés ETL, primer de tot, es realitza una extracció de les dades a partir d'un fitxer obtingut a través de l'IdesCat, prèviament modificat per a obtenir només els elements que interessen, amb la finalitat d'evitar la sobrecàrrega del servidor.

A l'etapa de transformació, es comprova que els diferents valors de la clau primària i dels atributs no siguin nuls, i en cas de ser-ho, s'associa un element perquè si la comprovació retorna *false*, no es prossegueixi amb l'execució.

Finalment, es realitza la càrrega de la informació a la base de dades.

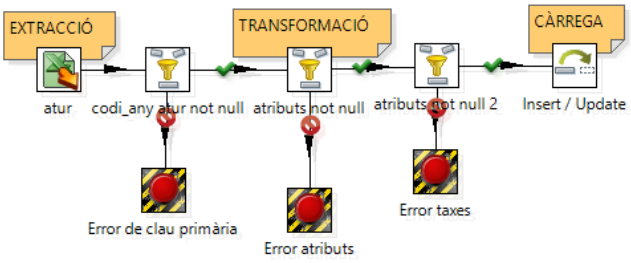

Fig. 4. Transformació de la dimensió atur.

4.2.3 Electes

A la primera columna, es troben localitzades totes les operacions d'extracció, de manera que s'obté la informació que acabarà conformant la taula de la dimensió electes, des de diferents fitxers origen.

Tot seguit, a la transformació, el que es fa amb *Append Streams* és unificar el contingut, perquè estigui tot en un mateix fitxer ajuntat. A continuació, es realitzen operacions de comprovació de columnes a la base de dades i que els atributs numèrics no siguin nuls. També però, en el darrer pas abans de la càrrega, es genera la *clau_electes*, l'atribut que esdevé clau primària subrogada, un enter autoincremental per a cadascuna de les files. Finalment, es procedeix a la Inserció/Actualització de la informació a la base de dades, en la darrera etapa, anomenada càrrega.

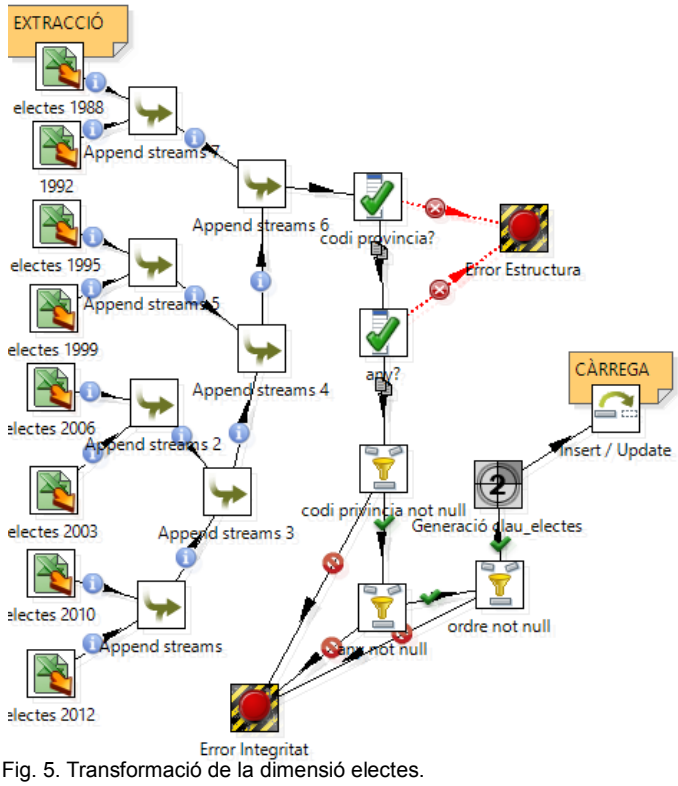

Fig. 5. Transformació de la dimensió electes.

4.2.4 Població

En la transformació Població, primer de tot en el pas referent a l'extracció, es duu a terme l'obtenció de la informació a través del fitxer de dades poblacionals, obtingut mitjançant l'IdesCat.

Tot seguit, es comprova que la clau primària no és nul·la, sinó s'associa l'element perquè salti un error i no es prossegueixi amb la transformació.

Finalment, si no hi ha cap error, es prossegueix amb la càrrega de la informació a la base de dades. Cal recordar, que abans d'executar la transformació, de la mateixa forma que en el cas anterior, s'executa l'*script* SQL d'aquest darrer element.

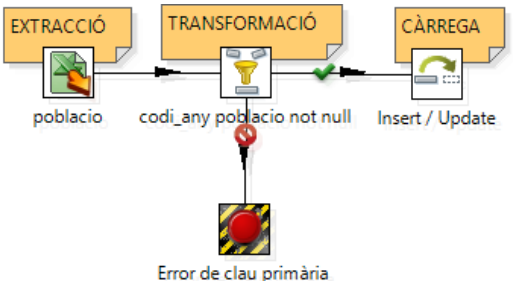

Fig. 6. Transformació de la dimensió població.

4.3 Schema Workbench

Aquesta eina, permet realitzar un cub *OLAP*, i publicar-lo posteriorment, al servidor de *Business Intelligence*. Per a dur a terme aquest procés, s'ha establert una connexió amb la base de dades, des d'on s'obté la informació ne-

cessària per a realitzar els diferents elements dels què es composa el cub.

El cub *OLAP*, s'organitza, a nivell arrel, en un s*chema* o esquema, en el *schema* se li associa un cub, i en aquest cub, alhora també se li associa, en primer lloc, la taula de fet i un conjunt de dimensions, i en segon lloc, les mètriques que seran emprades.

Cada dimensió disposa d'una o més jerarquies, i aquestes alhora disposen d'una taula, que és a partir de la qual s'extreu la informació necessària per aquesta dimensió, i un o més nivells, que són els elements de la jerarquia avaluats individualment.

Les mètriques de què disposa cada cub, fan referència a un conjunt d'atributs de la taula de fet, que han d'ésser mesurats, mitjançant operacions d'agregació, és a dir, operadors de *suma, count, distinct-count...*

S'han realitzat dos cubs, el primer d'ells és AturAnyEleccionsnb, i el segon, VotantsProvinciaGener.

4.3.1 Cub OLAP 1

En aquest primer cub, es pretén obtenir informació de les taules *atur* i *població*, en relació amb la taula de fet. Per a aquest propòsit s'han creat dues dimensions, una per a l'*atur* i una altra per a la *població*.

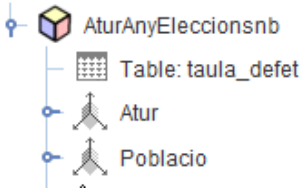

Fig. 7. Dimensions del cub AturAnyEleccionsnb.

La dimensió *atur*, conté una sola jerarquia amb dos nivells, l'*anyAtur* i *atur*, el primer, simplement és per a saber l'any en què es vol consultar l'atur, i el segon, és el valor de l'atur, en milers de persones.

Pel que fa a la dimensió *població*, està conformada també per una sola jerarquia, *poblacioAny*, que té dos nivells, *població* i *any*. El primer fa referència a la població real que viu a Catalunya, encara que no tinguin dret a votar, i el segon, fa referència a l'any en què aquesta població és consultada.

En darrer terme, hi ha les mètriques, en aquest cas, s'han creat 4 mètriques:

1. *VotantsAnyEleccio:* Mesura el conjunt de votants en uns comicis electorals determinats.
2. *Provincies :* Mesura el conjunt de províncies conformades quan van fer-se els comicis.
3. *Poblacio_dretavot:* Mesura el cens electoral que hi ha en uns comicis.
4. *Abstenció:* Mesura l'abstenció, en nombre de persones, que hi ha en uns comicis.

4.3.2 Cub OLAP 2

En aquesta ocasió, aquest cub està format per dues dimensions, *ElectesxProvincia* i *AnyPoblació*. La dimensió *ElectesxProvincia*, fa referència a la taula Electes, i està formada per una sola jerarquia, *ElectesProvincia*, que alhora es conforma per quatre nivells, *Electes, Província, Any* i *Partit.*

Electes fa referència al nom de l'electe, *Província*, a la circumscripció a la qual es presenten, *Any*, a l'any en que es realitzen els comicis i *Partit*, al partit amb el qual es presenten.

La segona dimensió, treballa amb la taula *Població*, que també es compon d'una sola jerarquia, i en aquest cas, de dos nivells, el primer permet saber-ne l'any, i el segon, la població.

En darrer terme, hi ha les mètriques del cub, en aquest cas, han estat realitzades 2 mètriques:

1. *Cens:* Mesura el conjunt de votants en uns comicis determinats.
2. *Votants:* Mesura el cens que hi ha en uns comicis electorals.

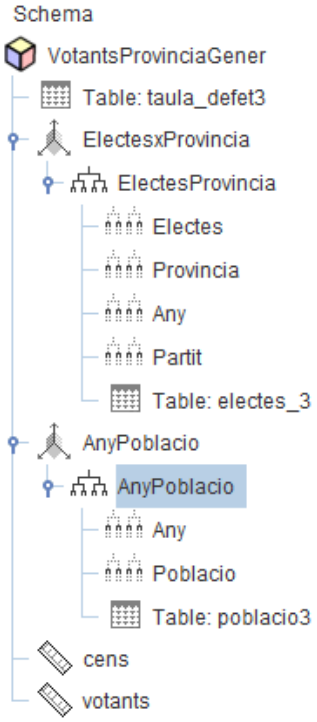

Fig. 8. Estructura del cub VotantsProvinciaGener.

5 RESULTATS

Analitzats els diferents conceptes teòrics i el funcionament de les diferents eines utilitzades, es proposen un conjunt de preguntes complexes de la situació actual en el context en què es troba l'empresa, on gràcies al *BI*, s'assisteix a la presa de decisions. El procés de realització dels cubs analitzats en aquest apartat, així com també la realització dels gràfics mitjançant *Saiku Analytics* i el conjunt de dades obtingudes de la base de dades, han estat realitzats seguint el procediment mostrat en els apartats anteriors.

Així doncs, es proposen un seguit de tres contextos empresarials, d'acord a l'operativa de l'empresa fictícia, extrapolables a una empresa real, adaptant-los a la seva operativa i context.

5.1 Context empresarial 1

En els darrers anys, el nombre de votants s'ha incrementat notòriament, no obstant això aquest increment podria ésser a causa de diferents circumstàncies, o bé a que el cens electoral s'hagi vist incrementat o que percentualment l'abstenció s'hagi vist reduïda. El que quedaria per veure també, és si aquest increment de la participació, i per tant de vots emesos, realment són vots a candidatures o si simplement són votants en nul.

5.1.1 Anàlisi

Per a poder analitzar la situació anteriorment exposada, s'utilitza el cub *AturAnyEleccionsnb*, i les mètriques que fan referència a *VotantsAnyEleccio*, *poblacio_dretavot* i *Abstenció*, com a nivell s'utilitza *anyAtur*, de la dimensió *Atur*. Tota aquesta informació pot ésser més contextualitzada tenint en compte la mètrica d'*abstenció*, i que permet saber si aquest increment es deu a un augment notori del cens electoral o bé a un declivi abstencionista.

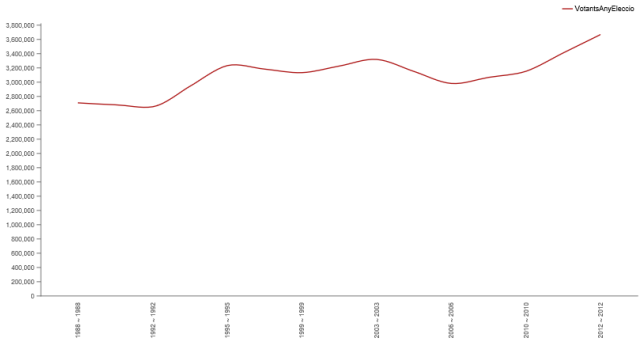

Fig. 9. Gràfic que mostra l'evolució de votants.

Gràcies al fet d'afegir aquesta nova mètrica, pot veure's la relació directa que hi ha entre l'augment o no de l'abstenció i el reflex que això té en el nombre de votants. Així doncs, hi ha una correlació directa, tot i que per a saber del cert si només és aquest el factor influent, ha d'ésser tingut en consideració també el cens electoral.

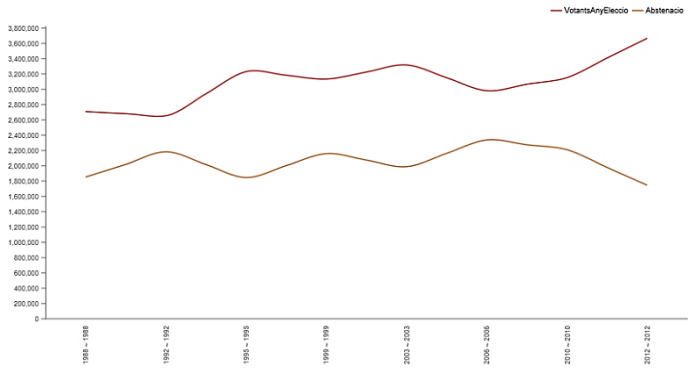

Fig. 10. Gràfic que mostra l'evolució de votants i de l'abstenció.

Amb aquesta nova mètrica, referent al cens electoral, s'observa que si bé aquest s'ha vist incrementat, un augment de la participació tan notori, no pot ser a causa només de l'increment censal, i encara menys en el període electoral que comprèn des de l'any 1999 fins al 2012, ja que aquest augment del cens ha estat, en termes absoluts, de 121.211 persones.

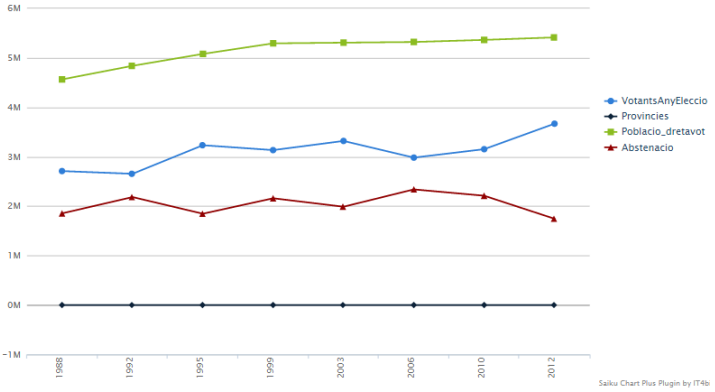

Fig. 11. Gràfic que mostra l'evolució de votants, províncies i de l'abstenció en comparació amb la població amb dret a vot.

Un altre gràfic que serveix per a veure de forma clara i precisa els diferents factors avaluats anteriorment, és el de barres percentuals al 100%, ja que permet, de les diferents mètriques avaluades, veure'n les diferències entre els diversos anys. Avaluant-lo s'extreu informació rellevant pel que fa als votants, que conformen el major nombre de tots, la reducció considerable de l'abstenció l'any 2012 i en darrer terme, el manteniment més o menys igual de la població amb dret a vot.

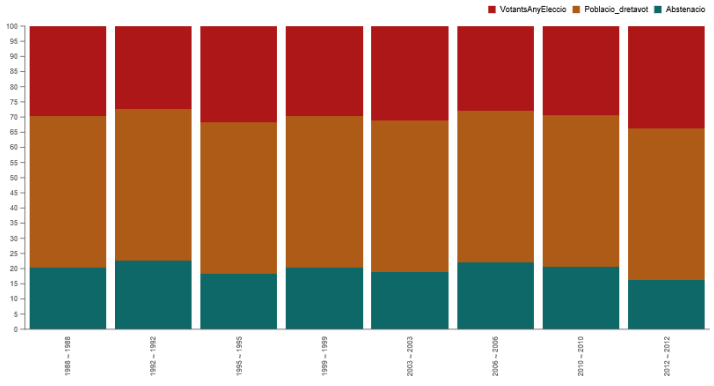

Fig. 12. Gràfic percentual al 100% que mostra l'evolució de votants i de l'abstenció en comparació amb la població amb dret a vot.

5.1.2 Contextualitzant

Tot i que el cens electoral s'ha vist incrementat, en termes absoluts de l'any 1999 al 2012, es dedueix que no conforma una xifra significativa, així doncs, l'increment dels vots emesos és a causa d'altres circumstàncies addicionals que han propiciat que l'abstencionisme es reduís de forma considerable.

5.1.3 Aspectes tècnics de l'anàlisi

Aquesta anàlisi, realitzada mitjançant un cub *OLAP* i amb *Saiku Analytics*, ha permès interpretar la informació de forma diversa, atenent-se a les característiques més adients en cada camp, per a fer comparatives.

També, ha permès la realització d'una idea més acurada i precisa de la informació subjecte de l'anàlisi, i fer-ne una traçabilitat temporal per a avaluar-ne el progrés de forma dinàmica. Emprant només les típiques taules, això no seria pas possible, ja que si bé aquestes conformen un bon element per a treballar amb la informació, la capacitat de representació que tenen, no té ni punt de comparació amb els gràfics.

Com a mancances, hi ha el fet que la personalització dels gràfics no pot realitzar-se sense tenir la versió de pagament de *Saiku Analytics*, així doncs, ni en la capçalera d'aquest, ni tampoc en els eixos de coordenades poden establir-se camps de text que permetin una major comprensibilitat.

5.2 Context empresarial 2

En els darrers anys a Catalunya s'ha viscut una situació inèdita, i és que per primera vegada, el nombre de població total esta disminuint a causa de l'emigració d'immigrants o bé de joves que marxen a treballar a fora. Aquest augment de la participació, podria deure's al fet que gent que anteriorment tenia dret a vot, ara ja no el té?

5.2.1 Anàlisi

Per a avaluar aquesta qüestió, es té en compte la població total de Catalunya en els comicis avaluats, i també, la població amb dret a vot de cadascun d'aquests. Per a la realització de la taula que es mostra a continuació són utilitzades la dimensió *Atur* i *Població*, de la jerarquia *AnyAtur*, s'empra el nivell *AnyAtur*, i de la jerarquia *PoblacioAny*, s'empren els nivells *població* i *any*, pel que fa a mètriques emprades, hi ha la *poblacio_dretavot*.

TAULA 1

EVOLUCIÓ DE VOTANTS I DE L'ABSTENCIÓ EN COMPARACIÓ AMB LA POBLACIÓ AMB DRET A VOT.

AnyAtur	Poblacio	Any	Poblacio_dretavot
1988	6041469	1988	4.564.389
1992	6096899	1992	4.839.071
1995	6104729	1995	5.079.981
1999	6174547	1999	5.293.657
2003	6693297	2003	5.307.837
2006	7146734	2006	5.321.274
2010	7501853	2010	5.363.688
2012	7478968	2012	5.413.868

D'aquesta taula, s'extreu molta informació rellevant, i és que si bé la població s'ha vist reduïda l'any 2012, paradoxalment, la població amb dret a vot ha continuat augmentant, així doncs, aquesta relació directa d'estudi queda descartada, però per a fer una comparativa i conclusió més complexa, es procedeix a avaluar també totes aquestes dades d'acord amb l'atur en cada cas. Així doncs, es realitza una nova taula, tenint en compte les dimensions *Atur* i *Població*, utilitzant la jerarquia *Atur*, del nivell *AnyAtur*, i la mètrica *poblacio_dretavot*.

TAULA 2

EVOLUCIÓ DE VOTANTS I DE L'ABSTENCIÓ EN COMPARACIÓ AMB LA POBLACIÓ AMB DRET A VOT I L'ATUR REGISTRAT.

AnyAtur	Atur	Poblacio	Any	Poblacio_dretavot
1988	0.0	6041469	1988	4.564.389
1992	0.0	6096899	1992	4.839.071
1995	0.0	6104729	1995	5.079.981
1999	225.1	6174547	1999	5.293.657
2003	279.6	6693297	2003	5.307.837
2006	260.7	7146734	2006	5.321.274
2010	562.7	7501853	2010	5.363.688
2012	647.0	7478968	2012	5.413.868

Dels anys 1988, 1992 i 1995, no es disposa de dades disponibles. Pel que fa al període de l'any 1999 al 2012, en milers de persones, fins l'any 2006 s'observa una situació bastant estable, tot i que la població continuava augmentant, i el nombre de població amb dret a vot també.

No és fins l'any 2010, amb la xifra de població més elevada a Catalunya, quan el nombre de persones a l'atur es duplica respecte l'any 2006, i finalment, l'any 2012 encara és major. A causa d'aquesta situació, un percentatge de la població comença a emigrar. No obstant això, el conjunt de població amb dret a vot, probablement a causa de diferents factors administratius, segueix augmentant.

5.2.2 Contextualitzant

Tot i que és cert que el nombre de població s'ha vist reduït en els darrers comicis, i que l'atur d'ençà l'any 2010 no para de pujar, no hi ha cap mena de relació directa en el cas avaluat, i això a causa del fet que la població amb dret a vot ha continuat augmentant ja que s'han anat consolidant els requisits de la població que encara resideix a Catalunya, per a poder participar en unes eleccions. D'aquesta manera es conclou que, la disminució de la població que si deixa de residir a Catalunya no podrà votar des de l'estranger, i que per tant, reduiria el cens, en termes generals, no ha estat el suficientment elevada com per a frenar aquesta tendència.

Per concloure, aquest augment de la participació, és a causa d'altres factors externs, que han propiciat que l'electorat se senti identificat en un projecte, que fins aquest moment no creien possible o viable, i aleshores, ha permès reduir el nombre d'abstencionistes en aquest sector.

5.2.3 Aspectes tècnics de l'anàlisi

Per a dissenyar aquesta anàlisi, han estat utilitzades les dimensions *Població* i *Atur*, no obstant això, només la mètrica *poblacio_dretavot* ha estat emprada. Les taules són útils en aquests casos en què es tracta amb poca informació, però hagués estat millor, també, la utilització de gràfics de suport. Això no ha estat possible, però, amb el disseny de la base de dades realitzat, ja que l'element subjecte de ser avaluat en el gràfic són les mètriques, si bé les dimensions formen part dels eixos de coordenades.

Per tant, per a produir un gràfic que permeti millorar aquesta anàlisi, es requereixen noves mètriques que permetin avaluar la població total i l'atur total. Aquests nous elements, haurien de formar part de la taula de fet, i en efecte, representa en el seu conjunt, una millora a realitzar en el futur.

5.3 Context empresarial 3

Es vol estudiar la regeneració democràtica al país, per a tal efecte, són considerades les províncies de Catalunya de Barcelona i Girona. Vol avaluar-se si els electes es presenten constantment, o si en altre cas, van donant pas a la regeneració dels partits, i també es vol estudiar si el fet de què els electes canviïn de província, constitueix una pràc-

tica comuna a l'hora d'intentar obtenir més vots, o bé, assegurar-se'n el lloc.

5.3.1 Anàlisi

Primer de tot, és necessari remarcar, que les dades dels electes i els partits polítics utilitzats per a realitzar aquesta anàlisi, han estat anonimitzats per a no contravenir la privacitat individual dels subjectes.

Per a tal efecte, s'empra un nou cub, anomenat, *Votants-ProvinciaGener*, que té en consideració les taules *electes*, *població*, i òbviament, la pròpia *taula de fet*, contemplant dues dimensions, una per a cadascuna de les taules electes i població. Es procedeix doncs a analitzar aquest cub, amb el *Pentaho BI Server*, i l'eina *Saiku Analytics*, i es comença a treballar amb la dimensió *ElectesxProvincia*, emprant els nivells *Electes*, *Provincia*, *Any* i *Partit* de la jerarquia *ElectesProvincia*. Un exemple, del resultat d'aquesta operació en format de taula, és el següent:

TAULA 3
INFORMACIÓ RELATIVA ALS ELECTES, EN PROVÍNCIA, ANY, PARTIT, I CENS.

Electes	Província	Any	Partit	Cens
Electe 1	BCN	1988	Partit 1	4.564.389
Electe 2	BCN	2012	Partit 2	5.413.868
Electe 3	BCN	2012	Partit 3	5.307.837

S'observen el conjunt d'electes, la província a la que estan inscrits, l'any dels comicis, el partit al que pertanyen, i el cens electoral de l'any corresponent. Els resultats estan ordenats per ordre alfabètic dels electes.

5.3.2 Contextualitzant

Dels electes avaluats, els casos més significatius són, l'electe 33, en el sentit que forma part de les llistes electorals des de l'any 1992 fins a l'actualitat, o bé l'electe 57 que es presenta a la circumscripció de Girona, des de l'any 1988, també fins a l'actualitat. Un darrer exemple podria ser el de l'electe 2, que porta presentant-se des de l'any 1995. No obstant això, ha d'ésser tingut en consideració, que no són avaluades les eleccions municipals, i que per tant, el període complert dels electes en política, podria variar. Pel que fa al canvi de circumscripció dels electes, en el fragment avaluat, només se n'observa un cas, el de l'electe 57, que des de l'any 2010 ho fa a Barcelona, i abans ho feia a Girona.

5.3.3 Aspectes tècnics de l'anàlisi

En aquesta anàlisi, ha estat utilitzada la dimensió *ElectesxProvincia*, en relació amb la població. Com pot observar-se en la taula generada, la mètrica emprada és *cens electoral*. L'altra informació que forma part de la taula, no prové d'una mètrica, sinó que forma part de les dimensions de les respectives taules. Això dificulta l'anàlisi de la informació de manera gràfica, tot i que seria bo analitzar per electes les mètriques de les que es disposa, el gran nombre d'electes disponibles - 1080 - , i el fet de què formin part, en una representació gràfica, de l'eix de coordenades, fa que el gràfic realitzat, no sigui interpretable per la seva gran magnitud.

Aquest element, constitueix un factor a tenir en compte, i en futures línies de millora, ha de fer replantejar l'organització de la base de dades, o bé simplement, el fet de tenir informació redundant a la taula de fet, obtinguda a partir de la dimensió, per a constituir una nova mètrica.

6 Conclusió

Primer de tot, ha estat primordial aprofundir en la definició i explicació dels Sistemes *BI*, per a comprendre quines són les seves fases d'implementació, així com també, la funcionalitat i la finalitat que proveeixen en un entorn empresarial. Una vegada realitzat aquest primer pas, un dels altres pilars fonamentals ha estat la simulació de l'operativa de treball d'una empresa hipotètica, a partir d'un conjunt de dades obtingudes mitjançant el portal de Dades Obertes de la Generalitat de Catalunya, i addicionalment, també des d'altres portals públics oficials, com l'*IdesCat*.

Les dades amb les quals s'ha treballat, sobretot en un context real, són d'important rellevància, ja que qualsevol implementació d'un sistema de *BI*, ha de permetre adaptar-se i cobrir les necessitats de l'entorn empresarial. A partir d'aquí, s'ha de generar un procés *ETL* complex que sigui capaç de gestionar els inputs de l'entorn i l'estructura de dades de l'empresa, amb l'objectiu de generar uns outputs amb un nivell qualitatiu elevat, per tal de què les anàlisis generades tinguin la màxima correctesa i detall possibles.

Mitjançant la realització d'una anàlisi de requeriments adaptada a les necessitats de l'empresa, i també, una realització de comparativa d'eines *BI* més populars, s'ha constat que *Pentaho*, és una de les eines que millor permet assolir el conjunt de requeriments establerts, i que per tant, permet cobrir el conjunt d'objectius especificats. *Pentaho* és una eina molt potent, que proveeix moltes capacitats d'anàlisi de dades gràcies a la seva modularitat en forma de diferents aplicacions per a generar els cubs, per a la integració de les dades i el servidor, i a més a més, permet pal·liar, en certa mesura, les mancances que poden haver-hi a la *Community Edition*, gràcies al *Marketplace*.

La base de dades, o *data warehouse*, és un element essencial de qualsevol projecte de BI. La definició del seu disseny i la posterior implementació d'aquest, han estat aspectes crucials pel correcte funcionament del sistema, així com també per a garantir-ne l'eficàcia. Per a dur a terme el disseny, s'ha implementat una base de dades en format estrella, que permet disposar d'una taula de fet, i de diferents dimensions, per a analitzar la informació i generar-ne unes anàlisis multidimensionals. Per tant, i com a possible línia de millora en aquest aspecte, podria implementar-se la base de dades en un servidor extern, sempre que el lloc de treball permeti la connexió remota de les aplicacions. D'aquesta manera es resoldrien problemes referents a la càrrega del sistema, i els càlculs podrien realitzar-se de forma més ràpida. Tot i això, aquest fet dependrà, en certa mesura, del context de l'empresa.

Un altre aspecte molt important és el procés *ETL*, realitzat amb l'eina *Kettle* , proveïda per *Pentaho*, que ha permès obtenir dades de diverses fonts d'informació rellevants, i que aplicat al conjunt d'aquestes, ha constituït un procés

molt complex i continu en el temps, atès que s'ha hagut d'adaptar a canvis produïts en les necessitats dels fitxers o bé a canvis de disseny en la base de dades. No obstant això, i tenint en compte que el procés *ETL* no ho pot automatitzar tot, els fitxers també són gestionats en certa mesura, manualment. En alguns elements a analitzar a l'hora d'inserir codis que garanteixin identificadors únics, la decisió pot dependre d'altres factors externs, donat el context de l'empresa, en aquest cas, un exemple seria quan diferents partits es fusionen o bé quan es creen coalicions, el fet de decidir si l'índex ha de ser nou o s'ha de mantenir, s'ha de decidir manualment. Així doncs, s'han assolit els objectius marcats pel que fa a realitzar i explicar el funcionament del procés *ETL*.

En relació, encara amb aquest procés, les *primaryKey* i les *foreignKey*, de les diferents taules creades a través del *Kettle*, no han pogut ésser establertes de forma directa i automàtica. Per tant, han hagut d'ésser indicades a partir de la interfície del *MySQL Workbench* ,amb un mini-script, o bé de forma manual per interfície gràfica. Com a possible millora en futures línies de treball en aquest camp, s'ha evidenciat que seria molt interessant automatitzar aquest procés, ja que permetria una realització de les tasques de forma més eficient.

El *Schema Workbench*, ha estat molt útil per a la realització dels cubs *OLAP* emprats, a partir dels diferents orígens de dades, gràcies al fet d'establir dins d'aquests, dimensions, jerarquies i nivells, però sobretot al fet que es publiquin directament al servidor de l'eina *BI*, i per tant, al fet de poder-se utilitzar de forma immediata. Han estat emprats dos cubs que han permès analitzar la informació de les diferents dimensions de la base de dades, en línies de treball futures, podrien ésser millorades les mètriques que aquests cubs utilitzen, o bé afegir-ne de noves, sempre però adaptant la base de dades, i en concret la taula de fet, a aquestes noves característiques, ja que la taula de fet, és la que conté les mètriques que després són utilitzades. *Saiku Analytics*, ha servit com a mesura pal·liativa per a resoldre les mancances de la versió de *Pentaho* utilitzada, i és que aquesta eina, ha permès realitzar una anàlisi de les dades, gràficament, molt superior a la que permet fer l'eina proveïda per defecte, no obstant això, com que *Saiku Analytics* també disposa de versions gratuïta i de pagament, l'eina utilitzada, ha tingut limitacions en aspectes secundaris, però no menys importants, com ara de personalització.

En definitiva, la totalitat d'eines emprades, interaccionen de la forma adient entre elles per a garantir el correcte funcionament del procés realitzat. Això permet explotar de forma eficient el potencial que tenen per a gestionar grans volums de dades, comprovar-ne la seva estructura o unificar els diferents fitxers que acaben formant part d'una taula per a carregar-la al *data warehouse*. A continuació, amb el servidor *Pentaho BI* i *Saiku Analytics*, són observades el conjunt de dades en correlació amb les preguntes detallades en els diversos contextos empresarials avaluats, i gràcies a la diversitat de gràfics disponibles, s'escull aquell que garanteix una comprensió més eficient i directa de la situació.

En darrer terme, és d'especial èmfasi remarcar l'extrapolació que el procés aplicat en aquest treball té en l'entorn empresarial, ja que el benefici d'haver utilitzat una eina amb una capacitat de dades tan gran i freqüentment utilitzada en les empreses i institucions, permet que en el cas de dur a terme la implementació o gestió d'un *BI* en una empresa real, el conjunt de les fases a realitzar siguin les mateixes que les efectuades en aquest treball.

AGRAÏMENTS

M'agradaria agrair a en Ramón, el meu tutor, tot el suport que m'ha donat al llarg de la realització del treball.

BIBLIOGRAFIA

[1] H.P. Luhn, A Business Intelligence System,IBM Journal of Research and Development, vol. 2, no. 4, pp. 314, 1958. doi:10.1147/rd.24.0314

[2] E.F.Codd , A Relational Model of Data for Large Shared Data Banks , Communications of the ACM, Vol. 13, No. 6, pp. 377-387, 1969. doi:10.1145/362384.362685

[3] GENERALITAT DE CATALUNYA (2014). Dades obertes gencat [en línia]. Barcelona: Generalitat de Catalunya. [consultat: 26 desembre 2014] Disponible a Internet: http://dadesobertes.gencat.cat/ca/cercador/cerca-cataleg?q=eleccions+al+parlament+de+catalunya

[4] INSTITUT D'ESTADÍSTICA DE CATALUNYA (2014). Població. Províncies [en línia]. Barcelona: Institut d'Estadística de Catalunya. [consultat el: 26 desembre 2014] Disponible a Internet: http://www.idescat.cat/pub/?id=aec&n=245

[5] INSTITUT D'ESTADÍSTICA DE CATALUNYA (2014). Atur registrat. Per sexe i grups d'edat [en línia]. Barcelona: Institut d'Estadística de Catalunya. [consultat el: 26 desembre 2014] Disponible a Internet: http://www.idescat.cat/economia/inec?tc=5&id=0607&dt=201411

[6] PENTAHO (2014). Data Integration – Kettle [en línia]. Orlando: Pentaho. [consultat el: 26 desembre 2014] Disponible a Internet: http://community.pentaho.com/projects/data-integration/

[7] PENTAHO (2014). Introducing Pentaho Community Edition 5.2 [en línia]. Orlando: Pentaho. [consultat: 26 desembre 2014] Disponible a Internet: http://community.pentaho.com/

[8] QLIK (2014). Visión General de QlikView [en línia]. Radnor: Qlik. [consultat: 25 novembre 2014] Disponible a Internet: http://www.qlik.com/es/explore/products/qlikview

[9] JEDOX (2014). Download a free, fully functioning Jedox trial [en línia]. Freiburg im Breisgau: Jedox. [consultat: 25 novembre 2014]. Disponible a Internet: http://www.jedox.com/en/product/free-software-trial/

[10] Carolina Duque i Andrés Eduardo (2014).SPAGOBI [en línia] [consultat: 25 novembre 2014] Disponible a Internet: https://prezi.com/rf7x4mjnlulv/spagobi/

[11] T. Barber (2014). Saiku [en línia] London: Saiku. [consultat el: 26 desembre 2014] Disponible a Internet: http://wiki.meteorite.bi/display/SAIK/Saiku