TREBALL FINAL DE CARRERA

Construcció i explotació d'un magatzem de dades per modelitzar la interacció a través del mitjà de comunicació social Twitter en un determinat esdeveniment col·lectiu

Memòria

Alumne: Antonio Sánchez Navarro

Consultor: Carles Llorach Rius

TFC – Magatzem de dades
Enginyeria Tècnica d'Informàtica de Gestió
Universitat Oberta de Catalunya
Curs 2014 – 2015 – Semestre de tardor
Data de lliurament: 6 de gener de 2015

CONTROL DE CANVIS

EDICIÓ	ELEMENT AFECTAT	DESCRIPCIÓ	DATA CANVI
Primera	Document original	Primera versió del document	28/12/2014
Segona	Es reestructuren punts	Davant de noves instruccions sobre el model de memòria a lliurar, rebudes amb data 29/12/2014, es pren com a punt de partida la primera edició canviant-li l'estructura i se li afegeixen continguts	01/012015

RESUM

Aquest document consisteix en la memòria del Treball Final de Carrera (TFC) en l'àrea de magatzem de dades corresponent al títol d'Enginyer Tècnic en Informàtica de Gestió (ETIG) per la Universitat Oberta de Catalunya (UOC). El document intenta recollir els aspectes més rellevants del treball que s'ha dut a terme a mode de síntesi de la documentació que s'ha anat aportant a la llarg del projecte, atès que aquest document és el resultat de fer una suma i adequació dels documents anteriors en format PAC.

L'objectiu perseguit en aquest TFC és la construcció d'una base de dades en un Data Warehouse (DW) i la utilització profitosa de les eines disponibles actualment per poder dur-la a terme. La implementació d'aquesta base de dades s'orienta cap a un magatzem de dades físic ROLAP, el qual es caracteritza per un seguit d'elements propis com ara dimensions, fets i atributs, així com per un conjunt de característiques que el defineixen com ara des-normalització de taules, inclusió d'informació agregada, historificació de la informació i d'altres.

Per assolir aquest objectiu tan ambiciós s'intenta desenvolupar un cas pràctic que ens ha de permetre posar en pràctica els coneixements adquirits a la titulació en un escenari concret. Aquest escenari s'ha descrit de forma adequada en el document d'enunciat del projecte, on s'explica la problemàtica plantejada i els objectius (generals i específics) que s'han d'acomplir amb la construcció del magatzem de dades. A més, amb l'enunciat es proporcionen les dades necessàries per a omplir el magatzem, les quals una vegada extretes s'han de transformar i manipular per adaptar-les al model multidimensional i dotar així al producte resultant dels requeriments d'anàlisi exigits.

Per acabar aquesta secció, amb la construcció del magatzem de dades i la configuració de l'eina d'explotació de dades (desenvolupament de la ETL que permeti la càrrega de les dades) s'espera que el producte desenvolupat inclogui un conjunt prefixat d'informes que mostrin la informació sol·licitada i qualsevol altra que sigui d'interès al grup d'investigació de la UOC. Aquesta informació consisteix bàsicament en indicadors de negoci demanats en l'enunciat.

PARAULES CLAU

Data Warehouse, Business Intelligence, magatzem de dades, ROLAP, dimensió, fet, model dimensional, taula del Fet, taula de dimensions, ETL, Pentaho Spoon-ETLs, Pentaho Report Designer, MySQL, SGBD, llenguatge SQL, PostgreSQL.

ÍNDEX DE CONTINGUTS

1. Introducció

Aquest TFC en l'àrea de Magatzem de dades se centra, tal i com indica el títol de portada del treball, en desenvolupar la "Construcció i explotació d'un magatzem de dades per modelitzar la interacció a través del mitjà de comunicació social Twitter en un determinat esdeveniment col·lectiu". El producte resultant d'aquest treball ens ha de permetre formalitzar la informació relativa a un seguit d'interaccions generades via Twitter en un esdeveniment científic amb un destacat nombre de participants i extreure'n un cert grau de coneixement nou a partir d'aquestes interaccions. A tal efecte, es proporciona un conjunt prefixat d'informes on es mostra tota la informació requerida, així com la construcció del magatzem de dades en si mateix utilitzant tecnologia relacional. Per dur a terme el treball de forma exitosa, s'han fet servir eines ETL (*extract, transform and load*) a partir d'una única font de dades que ha estat la base de dades de MySQL Server accessible a través de http://nairobi.uoc.es/phpmyadmin i ha calgut transformar i carregar les dades en una base de dades en el sistema gestor PostgreSQL. El nom triat per la base de dades de PostgreSQL ha estat CONGRESS i, un cop construïda, la informació necessària per extreure el grau de coneixement requerit s'ha mostrat en informes elaborats a partir del conjunt d'eines de Pentaho.

A la part inicial d'aquest document trobareu tot el pla de treball que s'ha seguit per a realitzar aquest projecte, destacant-hi la divisió de tasques, les fites que s'han marcat i com han estat planificades, els possibles riscos que es poden haver obtingut i la planificació que s'ha emprat (incloent-hi diagrames de Gantt).

Per acabar aquesta secció, només vull assenyalar que la realització d'aquest projecte correspon a un Treball Final de Carrera (TFC) dels estudis d'Enginyeria Tècnica en Informàtica de Gestió dins l'àrea temàtica del magatzem de dades o *data warehouse*. Aquest treball és doncs la cloenda o punt i final d'uns estudis i consisteix en la construcció d'un sistema informàtic que resol un problema plantejat, de forma semblant al que ens podríem trobar en el món laboral real. Per dur-lo a terme ha calgut doncs recórrer a un conjunt d'habilitats i coneixements ja adquirits a les diferents assignatures cursades en el Pla d'Estudis de la titulació, molt especialment les de temàtica relacionada amb *Bases de dades I* o *Bases de dades II*. L'essència d'aquest treball és, per tant, pràctica i de síntesi, tot i que també contempla una part molt important de redacció i presentació de la feina duta a terme.

2. Descripció del projecte

Abordarem en aquesta secció un conjunt d'aspectes referents al projecte desenvolupat com ara la seva justificació o idoneïtat, tot preguntant-nos el per què del projecte per passar després a la seva descripció, i analitzarem quins són els seus objectius generals i específics. Vegem-ho tot seguit de manera més detallada:

2.1. Justificació del projecte (idoneïtat)

El projecte es justifica pel fet de poder treure profit en forma de nou coneixement a partir d'unes dades procedents d'una única font a les quals s'han aplicat tècniques de transformació i manipulació, proporcionant així avantatges competitius molt profitosos en un entorn d'empresa. En el cas que ens ocupa, la informació que se n'obté és útil a un grup d'investigadors de la UOC en el seu àmbit de coneixement.

2.1.1. Per què aquest projecte?

Atès que Twitter s'ha convertit en la tercera xarxa social del món, amb un creixement imparable, un grup d'investigadors de la UOC s'ha proposat aprofundir en l'anàlisi de les interaccions que es generen via Twitter en esdeveniments amb alta assistència o participació. En un principi les investigacions s'han centrat en esdeveniments tipus congrés científic i disposem de part de l'anàlisi ja fet, però no s'ha pogut continuar el projecte i doncs cal ampliar, revisar i adaptar l'univers de discurs i altres feines ja fetes.

Des d'un punt de vista més acadèmic i personal, jo vaig decidir matricular l'assignatura TFC en la seva vessant del magatzem de dades atès la bona nota que vaig treure a l'assignatura *Mineria de dades* i davant l'oportunitat de posar al dia coneixements ja adquirits del llenguatge SQL en les assignatures *Bases de dades I - Bases de dades II*. A més, sempre m'ha resultat molt interessant la idea de generar coneixement i establir regles de negoci a base de manipular i emmagatzemar grans quantitats de dades.

2.1.2. Descripció del projecte

El projecte consisteix a construir un magatzem de dades amb tecnologia relacional que formalitzi la informació relativa a les interaccions que es generen via Twitter en esdeveniments d'alta concurrència de participants, de tal manera que de la informació obtinguda se'n pugui extreure un cert coneixement. La tàctica consistirà en escollir un esdeveniment d'uns quants dies de durada i on hi participi gent de tot el món, sent la UOC qui proporcioni la informació sobre usuaris, missatges, programa de l'esdeveniment i altres qüestions.

Per deixar clar què es vol construir exactament, un magatzem de dades (de l'anglès *data warehouse*) és una base de dades amb la informació històrica d'una organització dissenyada i estructurada per dissenyar-hi consultes eficientment.

Les dades d'aquests magatzems provenen de sistemes d'informació transaccionals de les organitzacions (per exemple d'un ERP, Planificació de Recursos Empresarials, de l'anglès *Enterprise Resource Planning*). El magatzem realitza una funció d'integració de dades, ja que periòdicament es poden realitzar processos de càrrega (i refresc) d'informació des dels sistemes transaccionals fins al magatzem de dades. En aquests processos es pot realitzar una transformació o neteja de les dades i conceptualment es realitza una integració de dades de diverses fonts.

Les operacions realitzades sobre un magatzem de dades i els programes que les realitzen poden ser de diversos tipus. Les eines anomenades OLAP (*online analytical processing*) consisteixen a realitzar consultes, anàlisis, estadístiques i realitzar informes, d'una manera gràfica, multidimensional i amb operadors específics, facilitant, per tant, les consultes complexes (especialment les agregades) respecte a les eines d'informes tradicionals (generalment basades en SQL). Un altre tipus d'eines que solen anar associades al magatzem de dades són les eines de mineria de dades.

Els magatzems de dades allotgen grans quantitats de dades que poden ser agrupades en unitats conceptuals anomenades *datamarts*.

Per tal d'aconseguir un desenvolupament ben ordenat i ben documentat, així com per a seguir una bona temporització, s'ha dividit el projecte en quatre grans fites o etapes tal i com es pot apreciar tot seguit:

Etapes del TFC	Inici / Enunciat	Lliurament
PAC1 - Pla de treball	19/09/2014	01/10/2014
PAC2 - Anàlisi i Disseny	02/10/2014	05/11/2014
PAC3 - Implementació	06/11/2014	18/12/2014
PAC4 - Lliurament Final i Defensa	19/12/2014	06/01/2015

2.2. Objectius del projecte

Els objectius de tot projecte expressen la seva finalitat, els efectes i resultats que s'espera aconseguir, de tal manera que la viabilitat del projecte depèn molt de la correcta formulació dels seus objectius. Distingim dos grans tipus d'objectius, que són els generals i els específics.

• Els **objectius generals** corresponen a les finalitats genèriques d'un projecte o entitat. Expressen el propòsit central del projecte i han de ser coherents amb la missió de l'entitat.

• Els **objectius específics** es deriven dels objectius generals i els concreten, assenyalant el camí que cal seguir per tal d'assolir-los. Indiquen els efectes específics que es volen aconseguir tot i que encara no expliciten les accions directament mesurables mitjançant indicadors.

Dit això, anem a veure quins són els objectius generals i específics del projecte que ens ocupa:

2.2.1. Generals

En un context d'assignatura final de carrera, el TFC és un treball de síntesi eminentment pràctic i vinculat a l'exercici professional de la informàtica. La seva filosofia és sintetitzar uns coneixements del pla d'estudis de la titulació i enfocar-los a obtenir un producte funcional. En aquesta àrea de TFC que és el magatzem de dades, es pretén de forma genèrica oferir a l'estudiant, ja enriquit amb l'estudi de les bases de dades, l'oportunitat d'endinsar-se en aquest món dins del marc de les bases de dades.

L'objectiu essencial del projecte és aprofundir en l'anàlisi de les interaccions generades via Twitter per tal d'extreure'n informació que pugui ser útil als investigadors en el seu àmbit de coneixement. El magatzem de dades que es construeixi ha de permetre formalitzar la informació relativa a les interaccions en Twitter.

Pel que fa a l'estudiant que s'implica en aquest treball, el projecte li ha de permetre adquirir experiència en el disseny, construcció i explotació d'un magatzem de dades a partir de la informació continguda en una base de dades convencional. A més, l'estudiant tindrà ocasió de treballar les tècniques de gestió de projectes i d'analitzar els requeriments seguint les necessitats d'un client.

2.2.2. Específics

Ser capaç d'obtenir, com a mínim, el següent coneixement a partir de tècniques d'explotació de dades:

- Temes més parlats
- Els usuaris de twitter més actius de l'esdeveniment
- Els més seguits en temes concrets
- Els recursos compartits en tweets
- Els hashtag coincidents en tweets
- Les activitats dels assistents, tant si s'envien o reenvien tweets com si es segueixen a d'altres
- Evolució del nombre de tweets al llarg del temps

Ser capaç de revisar i adaptar el mapa conceptual que descriu l'univers de discurs a l'annex d'aquest enunciat de projecte.

Realitzar un primer anàlisi preliminar no detallat dels requeriments del projecte i analitzar les fonts de dades operacionals utilitzades que han de servir per carregar els diferents elements d'anàlisi.

El projecte proporcionarà també un conjunt prefixat d'informes que mostri tota mena d'informació útil als investigadors.

Documentar el projecte i redactar-ne la memòria, que inclourà tant com s'ha anat gestant el projecte com aspectes funcionals.

2.3. Enfocament i mètode seguit

L'enfocament que s'ha seguit ha estat totalment pràctic i s'ha basat en el cicle de vida clàssic o en cascada. Vegem en què consisteix:

- Anàlisi prèvia
- Definició del Pla de Treball
- Anàlisi funcional
- Disseny funcional
- Disseny tècnic
- Implementació
- Proves
- Lliurament final i defensa si s'hi escau

3. Requeriments de la solució

Abans de citar i classificar quins són els requeriments de la solució, val la pena recordar que un requeriment és una característica observable del nostre sistema que satisfà una necessitat o expressa una restricció que afecta el programari que estem desenvolupant. Un requeriment expressa les necessitats i restriccions que afecten un producte de programari que contribueix a la solució d'un problema del món real, essent útil per a delimitar quines de les possibles solucions al problema són adequades i quines no, essent els *stakeholders* els qui decideixen les restriccions que ha de complir el programari. Els *stakeholders* d'un projecte són aquelles persones i entitats que tenen un cert impacte o interès en aquest. Així, els requeriments ens diuen què és el que els diferents *stakeholders* esperen del nou sistema.

Els requeriments es poden classificar en dos grans grups: els que fan referència a les **necessitats** que ha de satisfer el sistema (**què** ha de fer) i els que expressen **restriccions** sobre el conjunt de solucions possibles (**com** ho ha de fer). Dels del primer grup en diem **requisits funcionals**, mentre que dels del segon grup en diem **requisits no funcionals**. Els requeriments funcionals fan referència a la funcionalitat que ha de proporcionar el sistema i ens indiquen quin és el comportament del sistema davant dels estímuls que li arriben de l'exterior. Els requeriments no funcionals acostumen a tenir forma de restricció i acostumen a afectar gran part del sistema, sense incloure-hi comportament.

Un cop feta aquesta exposició, esmentaré els requeriments funcionals i no funcionals que he identificat en una primera i acurada lectura de l'enunciat del TFC:

3.1. Requeriments funcionals

- El magatzem de dades que en resulti ha de permetre formalitzar la informació relativa a les interaccions i extreure'n un cert grau de coneixement, que s'especifica a l'enunciat del TFC, així com l'explotació dels seus usuaris.

- El sistema proporcionarà un conjunt prefixat d'informes on es mostri la informació sol·licitada, així com la que sigui requerida pels investigadors.

La solució informàtica, a més, constarà dels següents subsistemes:

- Una eina de càrrega de dades ETL (acrònim de l'anglès *Extract, Transform and Load*). Es tracta de l'aplicació que permet a les organitzacions moure dades des de múltiples fonts, reformatar-les, netejar-les i carregar-les en una altra base dades, *datamart* o *data warehouse* per analitzar, o en un altre sistema operacional per suportar un procés de negoci.

- Una eina d'anàlisi de dades OLAP (acrònim de l'anglès *On-Line Analytical Processing*), que és la solució utilitzada en el camp de la intel·ligència empresarial (*Business Intelligence*) amb l'objectiu d'agilitzar la consulta de grans quantitats de dades. Es fa servir en informes de negocis de vendes, *marketing*, informes de direcció, mineria de dades i àrees afins.

El projecte ha de permetre doncs a l'estudiant del TFC adquirir nous coneixements en els àmbits dels ETL i els OLAP.

L'aplicació ha de tenir almenys dos perfils d'usuari. Hi haurà un usuari consultor que accedirà a les dades, executarà consultes i emetrà informes. L'usuari administrador tindrà accés a totes les tasques anteriors i a més podrà executar les càrregues de noves dades i dissenyarà noves consultes.

3.2. Requeriments no funcionals

- El magatzem de dades ha de ser construït amb tecnologia relacional. Això vol dir que el seu model de dades ha de basar-se en la lògica de predicats i en la teoria de conjunts, essent la seva idea fonamental l'ús de les relacions.

- La informació s'ha de poder consultar de forma agregada per tipus d'usuari, característiques del missatge i etiqueta.

- Les dades han de ser processades a nivell de dia, hora i minut.

- Tot i que sempre estaran en format text, cal analitzar les fonts de dades operacionals proporcionades que han de servir per carregar cadascun dels elements d'anàlisi. La justificació d'aquest fet és que les dades en format text poden tenir diferents arxius de procedència (ODT, MDB, CSV, XLS i d'altres) i les dades es poden expressar de formes diferents per expressar un mateix concepte (hores, dates, dades geogràfiques de longitud i latitud...).

- El producte final que en resulti ha de realitzar-se sobre una màquina virtual Amazon que proporciona la UOC.

4. Funcionalitats a implementar

A més de l'eina de càrrega de dades ETL i de l'eina d'anàlisi de dades OLAP que ja he esmentat en l'apartat dels requisits funcionals (secció *3.1. Requeriments funcionals*), destaco ara que la base de dades relacional ha de formalitzar la informació relativa a les

interaccions generades via Twitter descrivint-les com a registres o files d'una taula que ens diguin que un usuari U, en un instant T, faci una acció A, sobre un recurs R, mitjançant un dispositiu D i n'obtingui un resultat R. La gestió i manipulació d'aquesta base de dades ens ha de dur a obtenir nou coneixement, que serà d'interès dels investigadors.

5. Resultats esperats

Els resultats esperats en cadascuna de les etapes del projecte són els que es mostren en el següent esquema:

ETAPA 1 – PLA DE TREBALL ☞ Presentació del document del Pla de Treball amb data límit 01/10/2014. En aquest Pla de Treball l'estudiant ha d'indicar, amb un cert nivell de detall, les tasques que s'hauran de realitzar, juntament amb una anàlisi de riscos i un diagrama de Gantt.

ETAPA 2 – ANÀLISI I DISSENY ☞ En aquest segon apartat del TFC, s'haurà de fer l'anàlisi de la problemàtica proposada a l'enunciat i s'haurà de fer el disseny de la solució generant el model multidimensional de les dades, així com tota la informació necessària per a poder assolir els objectius demanats. Data límit de lliurament 05/11/2014.

ETAPA 3 – IMPLEMENTACIÓ ☞ En aquesta fase s'haurà de crear la BDD, tal i com s'ha dissenyat a l'apartat anterior, i carregar-la amb les dades corresponents. En aquest procés de càrrega hi haurà d'haver un procés de transformació i adequació al que es demana. Un cop carregada la BDD s'hauran de generar els informes necessaris per a poder resoldre el que s'ha demanat. Aquests informes s'han de generar amb l'eina indicada pel consultor. La data límit de lliurament és 18/12/2014.

ETAPA 4 – MEMÒRIA I PRESENTACIÓ VIRTUAL ☞ Al final del semestre i, un cop lliurats els treballs anteriors, caldrà crear la memòria del TFC on s'explicarà el treball realitzat (habitualment serà una suma i adequació dels lliuraments anteriors). També s'haurà de crear una presentació virtual on s'expliqui el treball fet. Aquesta presentació serà la que permetrà la defensa del treball davant del tribunal. La data límit de lliurament és 06/01/2015.

5.1. Planificació inicial vs planificació final

A grans trets, puc afirmar que he respectat de forma prou acurada la planificació inicial i que he anat complint les fites en els terminis establerts. Per motius obvis, les fites que han requerit més dedicació i que han absorbit més temps en la realització del projecte han estat les d'implementació i elaboració de la memòria i presentació del projecte. La fita d'implementació presentava com a activitats afegides un seguit de proves unitàries, tant com per assegurar que les dades havien estat carregades de forma correcta com per a assegurar que els informes mostraven la informació requerida de forma correcta. A més, representava haver de programar en el llenguatge SQL tant la creació de les taules del magatzem co la seva càrrega. La fita consistent en l'elaboració de la memòria i vídeo de presentació ha suposat haver de dedicar un temps extra considerable a aprendre

l'ús i possibilitats de MS Power Point, així com de l'eina de presentacions d'àudio i vídeo Camtasia Studio 8. Un cop dit tot això, exposaré ara el que ha estat la Planificació d'aquest projecte contemplant-hi la seva proposta d'activitats i cronograma.

5.1.1. Proposta d'activitats i cronograma

5.1.1.1. Relació d'activitats, estimació de temps i fites a complir

La següent taula mostra la proposta d'activitats amb un cert grau de detall, contemplant-hi el llistat d'activitats, la durada en el temps de les activitats i les fites a complir. Tot seguit abordaré el tema del diagrama de Gantt a mode de cronograma que recull i situa en el temps tota la informació continguda en la taula.

ACTIVITAT	Descripció de l'activitat	Data inici	Data fi	Predecessores	Observacions
1	**TFC - Magatzem de dades**	**17/09/2014**	**23/01/2015**		L'assignatura finalitza amb el debat virtual
2	**Fase 1 - Iniciem el TFC!**	**17/09/2014**	**01/10/2014**		
3	Presentació al fòrum	17/09/2014	17/09/2014		
4	Obtenció dels recursos de l'assignatura (documentació i programari)	17/09/2014	17/09/2014		He obtingut la llicència del MagicDraw i l'he instal·lat al meu computador. El programari DW, la BBDD i la màquina virtual les instal·laré abans de la fase d'implementació
5	Lectura del Pla Docent, de la documentació i primera ullada al programari	18/09/2014	19/09/2014	4	
6	**Fase 2 - PAC1 - Pla de Treball**	**19/09/2014**	**01/10/2014**		
7	Lectura i anàlisi de l'enunciat del Projecte (PAC1)	19/09/2014	21/09/2014		
8	Consulta al fòrum sobre el SGBD PostgreSQL	21/09/2014	21/09/2014		La consulta ha estat atesa i ben tractada pel consultor de l'assignatura
9	Elaboració del Pla de Treball	19/09/2014	26/09/2014		
10	Anàlisi preliminar de requeriments	22/09/2014	23/09/2014	7	L'anàlisi preliminar de requeriments es basa en els exposats a l'enunciat del projecte
11	Rebuda conjunt de dades a tall d'exemple	27/09/2014	28/09/2014		El conjunt definitiu de dades es fa arribar al cap de tres setmanes
12	Revisió del document solució PAC1 que cal lliurar	28/09/2014	30/09/2014		Revisió dels continguts i correcció de possibles faltes d'ortografia
13	**Fita 1 - Lliurament PAC1**	**01/10/2014**	**01/10/2014**	**12**	
14	**Fase 3 - PAC2 - Anàlisi i Disseny**	**02/10/2014**	**05/11/2014**		
15	Lectura enunciat PAC2	02/10/2014	02/10/2014		
16	Estudi dels materials de l'assignatura referents al DW:	02/10/2014	07/10/2014		Es tracta d'estudiar i assimilar conceptes nous

	eines ETL i OLAP				
17	Anàlisi detallat de requeriments basat en l'anàlisi preliminar	08/10/2014	11/10/2014	7, 15, 16	Revisió obligada de l'enunciat del Projecte
18	Disseny amb la descripció del model dimensional	13/10/2014	17/10/2014	7, 17	Revisió obligada de l'enunciat del Projecte
19	Disseny dels procediments d'extracció de dades a alt nivell (eines ETL i OLAP)	18/10/2014	25/10/2014	18	
20	Disseny i nombre d'informes	26/10/2014	29/10/2014	19	
21	Instal·lació de la màquina virtual, incloent-hi el programari DW i el SGBD	30/10/2014	31/10/2014		
22	Revisió del document solució PAC2 que cal lliurar	02/11/2014	04/11/2014		
23	**Fita 2 - Lliurament PAC2**	**05/11/2014**	**05/11/2014**	**22**	
24	**Fase 4 - Implementació**	**06/11/2014**	**18/12/2014**		
25	Lectura enunciat PAC3	06/11/2014	06/11/2014		
26	Lectura de material i bibliografia complementària, si s'hi escau	06/11/2014	08/11/2014		
27	Construcció del magatzem de dades	09/11/2014	25/11/2014		
28	Configuració de l'eina d'explotació de dades (processos ETL i OLAP)	26/11/2014	10/12/2014	27	
29	Construcció dels informes i anàlisi de la informació	11/12/2014	16/12/2014	28	
30	Revisió del document solució PAC3 que cal lliurar	17/12/2014	18/12/2014		
31	**Fita 3 - Lliurament PAC3**	**18/12/2014**	**18/12/2014**	**30**	
32	**Fase 5 - Lliurament Final i Defensa**	19/12/2014	06/01/2015		
33	Lectura enunciat PAC4	19/12/2014	19/12/2014		
34	Elaboració de la memòria final	20/12/2014	04/01/2015		
35	Elaboració de la presentació virtual	26/12/2014	04/01/2015		
36	**Fita 4 - Revisió dels aspectes formals de comunicació escrita i Lliurament Final**	**05/01/2015**	**06/01/2015**	**34**	
37	Recull de dades i documentació per a confeccionar la memòria	20/09/2014	19/12/2014		

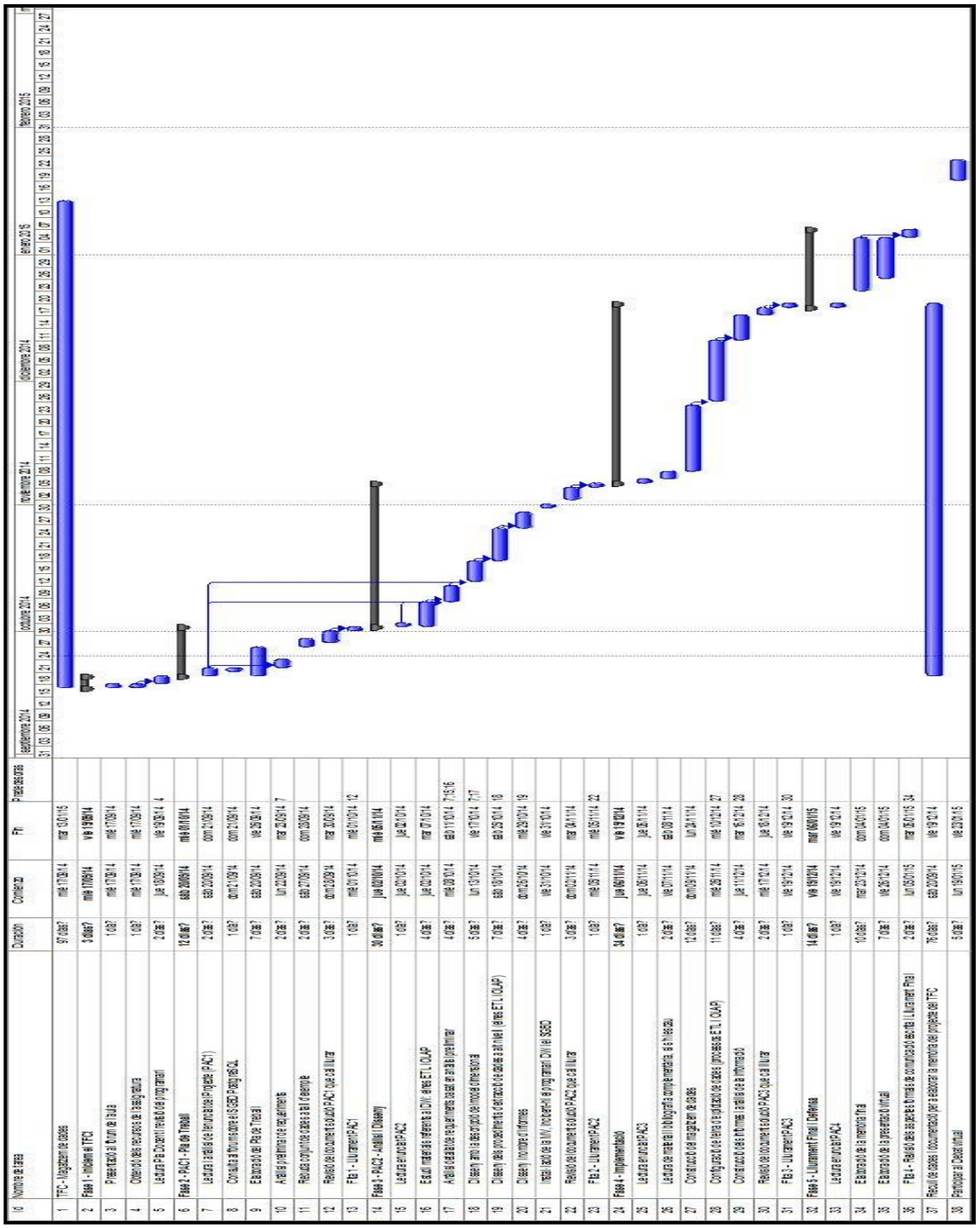
	del projecte				
38	Participar al Debat virtual	19/01/2015	23/01/2015		

Tot i que les fites a complir, així com la seva temporització, ja s'han remarcat a la taula d'abans, per facilitar-ne la informació mostro un resum detallat de les fites d'aquest TFC. A mode de conclusió, el termini de presentació de cadascuna de les fites, així com llur realització, ha estat prou ben respectat en tots els casos.

Fita 1 – Lliurament PAC1	Pla de Treball i anàlisi preliminar de requeriments	01/10/2014
Fita 2 – Lliurament PAC2	Anàlisi i Disseny	05/11/2014
Fita 3 – Lliurament PAC3	Implementació	18/12/2014
Fita 4 – Lliurament PAC4	Lliurament Final i Defensa	06/01/2015

5.1.1.2. Diagrama de Gantt

El diagrama de Gantt és una eina gràfica molt útil que té com a objectiu mostrar el temps de dedicació previst per diferents tasques o activitats a la llarg d'un temps total determinat que dura un projecte. Tot i així, el diagrama de Gantt no indica les relacions que hi ha entre tasques. Hi ha força programari per treballar amb diagrames de Gantt, que pot ser lliure (*GanttProject*) o amb llicència (*MS Project 2007*). Vegem el diagrama de Gantt que en resulta de la meva planificació del TFC amb l'eina de Microsoft:



Abans de finalitzar aquesta secció, val a dir que també havia elaborat un possible pla de contingència per atendre les possibles eventualitats que poguessin afectar la temporització dissenyada inicialment per a aquest projecte. Aquest pla de contingència consisteix en una anàlisi de riscos on s'enumeren possibles incidències i on es proposen possibles solucions que ajudin a superar aquestes incidències. Vull assenyalar també que havia fixat que les setmanes que dedicaria al TFC serien de cinc dies amb horari de tardes en la franja que va de les 16:00 a les 22:00 hores, amb una hora aproximadament per sopar. Aquesta disponibilitat és resultat d'haver de compatibilitzar els estudis amb la meva feina com a professor docent de secundària en horari intensiu de matí. Atès que he

acabat matriculant un màster de Sistemes i Business Intelligence de SAP a l'escola PC Carrier, la meva disponibilitat pel TFC ha minvat entre setmana i he hagut de dedicar més hores a la causa també els caps de setmana de novembre i desembre de 2014. Vegem ara una taula on s'analitzen els possibles riscos i on s'hi proposa una solució:

5.1.1.3. Anàlisi de riscos

Incidència	Solució proposada
Avaria o fallada del maquinari emprat	Replicar les condicions de treball i fer còpies de seguretat externes amb una certa periodicitat. Un bon suggeriment és copiar diària o setmanalment en el pitjor dels casos. Reparar o substituir el maquinari avariat
Incompetència d'algun dels components del programari amb la versió del SO Windows 7 de 64 bits	Eliminem el component de software i instal·lem la versió que s'ajusti millor a les característiques del SO que estem utilitzant
En el disseny de la BDD amb PostgreSQL, una fila supera el límit de 8 k que suporta el SGBD	Augmentar el límit de 8 k a 32 k per fila del PostgreSQL, amb la considerable disminució del rendiment que això representa. En el pitjor dels casos, substitució del SGBD per un altre com ara MySQL
Es produeix una fallada durant l'execució d'una transacció en el SGBD, que la fa avortar completament	Es re-programa la part de codi que ha provocat la fallada, assumint si s'escau una més gran quantitat de recursos o una execució més lenta. Si la fallada en alguns elements de programació és inevitable usant PostgreSQL, substitució del SGBD per un altre com ara MySQL
Malaltia	Els recursos de personal en aquest projecte del TFC es limiten a l'estudiant, que assumeix tota la feina. És per això que en cap cas suspendrà la seva realització en cas d'indisposició lleu. En cas de malaltia amb incapacitat temporal moderada, l'estudiant avisarà als consultors i aquests valoraran si li poden atorgar més dies per completar la feina. Si la incapacitat s'allarga força en el temps per malaltia greu, es suspèn el TFC i l'estudiant l'ajorna al semestre de primavera de 2015
Contratemps diversos que redueixen la disponibilitat horària	Disposo d'una bona disponibilitat horària, tot i que treballo a jornada completa, atès que aquest semestre només matriculo el TFC. Si decidís matricular-me en un Màster de tecnologia SAP i ABAP IV, llavors no disposaria de tanta disponibilitat i hauria de comptar amb els caps de setmana i anar a dormir més tard en dies feiners, atès que m'hi estaria dedicant al TFC en unes hores més intempestives del que és habitual
Campus virtual de la UOC temporalment fora de	No lliurar els documents solució sobre la data límit.

servei que impideix lliurar la feina en data límit (incidència al campus)	Esperar que els responsables de manteniment solucionin la incidència i efectuar llavors el lliurament. Si no fos possible el lliurament, alertar els consultors de l'assignatura i pactar un ajornament excepcional en el lliurament

6. Anàlisi i Disseny

Abordem en aquesta secció, d'una banda, una anàlisi detallada dels requeriments funcionals i no funcionals de la problemàtica plantejada i, de l'altra, el disseny de la solució generant el model multidimensional de les dades. Vegem-ho tot seguit:

6.1. Requeriments funcionals / no funcionals

Tot i que a la secció *3. Requeriments de la solució* ja s'ha fet una primera anàlisi preliminar dels requeriments, s'afegiran ara alguns requeriments i es durà a terme un examen més exhaustiu dels mateixos atès que es considera que ja disposem de les dades a incorporar al sistema BI.

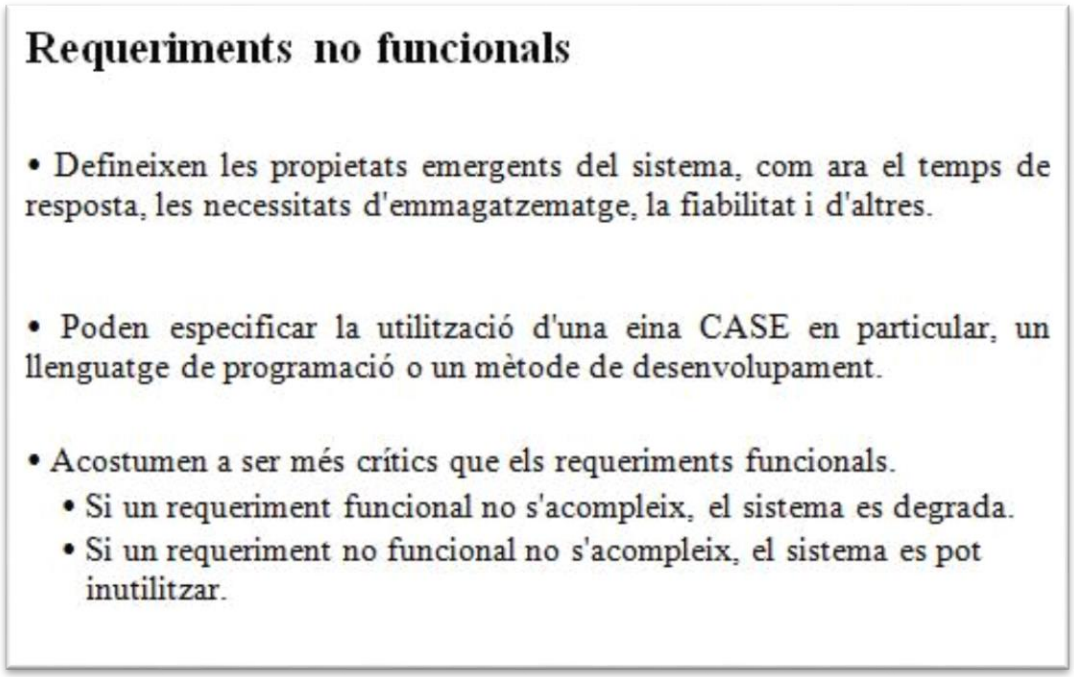
Abans d'analitzar els requeriments del sistema que s'intenta construir, insistim en la definició de requeriment com a una condició o necessitat que li cal a l'usuari per resoldre un problema o aconseguir un determinat objectiu. També es pot aplicar a les condicions que ha de tenir un sistema o un dels seus components per a satisfer un contracte, una norma o bé una especificació. Així, tenim que un requeriment es pot veure com:

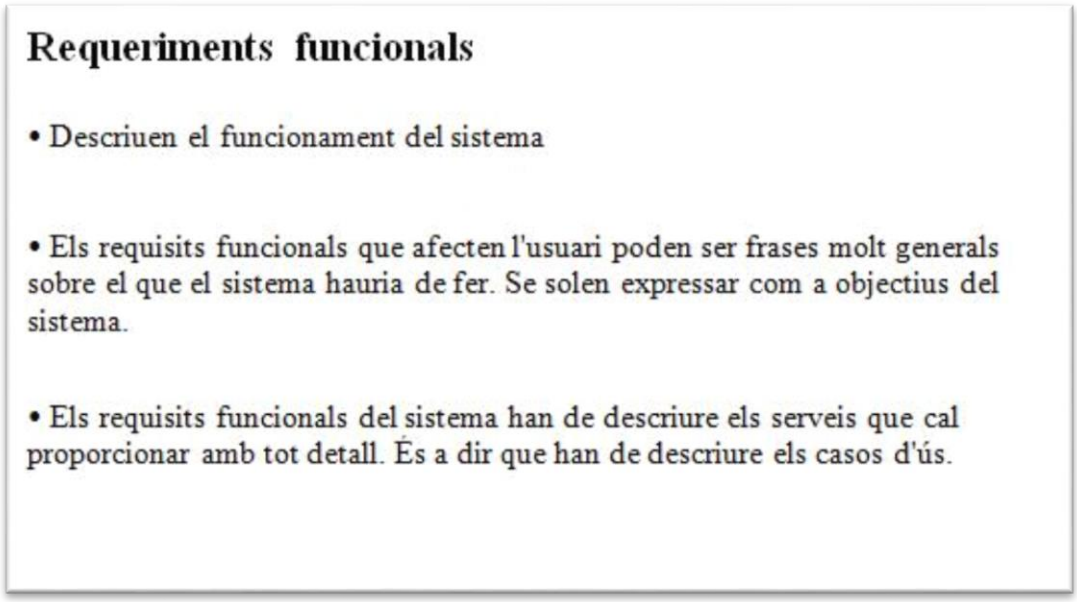
- una declaració abstracta d'alt nivell d'un servei que el sistema ens ha de proporcionar.
- una definició matemàtica detallada i formal d'una funció del sistema.

D'altra banda, els requeriments compleixen una doble funció:

- Són una oferta de contracte, la qual cosa implica que estan oberts a la interpretació.
- Són el contracte en si mateix, la qual cosa vol dir que han d'estar definits de la forma més detallada possible.

Els requeriments es poden classificar en dues grans categories, que són les de requeriments funcionals i no funcionals en funció de si ens resolen qüestions del tipus **què fa** el sistema o bé si ens resolen del tipus **com ho ha de fer** el sistema. També entren en aquesta segona categoria els requeriments que indiquen alguna restricció o impediment pel que fa al mode funcionament del sistema. Vegem-ho a les següents taules:





La plantilla Volere, a més, inclou la següent llista predefinida de requeriments no funcionals:

- Requisits de presentació
- Requisits d'usabilitat i humanitat
- Requisits de compliment
  - Velocitat
  - Latència
  - Seguretat (dany que podem provocar a altres persones)
  - Precisió
  - Fiabilitat

- Robustesa
- Capacitat (volum d'usuaris, de dades, etc.)
- Escalabilitat o extensibilitat
- Longevitat

• Requisits operacionals i d'entorn

• Requisits de manteniment i suport

• Requisits de seguretat

• Requisits culturals i polítics

• Requisits legals

Un cop feta la presentació de requeriments d'abans, estem en condicions de citar i analitzar detalladament els requeriments del nostre projecte.

6.1.1. Requeriments funcionals

Seguint l'índex que proposa la plantilla Volere, distingirem els requisits funcionals que fan referència a l'abast del projecte, els que fan referència a l'abast del producte i els requisits funcionals i de dades.

Abast del projecte:

• La solució informàtica constarà dels següents subsistemes:

- Base de dades relacional (del tipus MySQL Server o PostgreSQL), que recollirà les dades de les interaccions via Twitter en l'esdeveniment i permetrà l'explotació per part dels usuaris amb l'objectiu de crear nou coneixement.

- Una eina de càrrega de dades ETL (acrònim de l'anglès *Extract, Transform and Load*). Es tracta de l'aplicació que permet a les organitzacions moure dades des de múltiples fonts, reformatar-les, netejar-les i carregar-les en una altra base dades, *datamart* o *data warehouse* per analitzar, o en un altre sistema operacional per suportar un procés de negoci.

- Una eina d'anàlisi de dades OLAP (acrònim de l'anglès *On-Line Analytical Processing*), que és la solució utilitzada en el camp de la intel·ligència empresarial (*Business Intelligence*) amb l'objectiu d'agilitzar la consulta de grans quantitats de dades. Es fa servir en informes de negocis de vendes, *marketing*, informes de direcció, mineria de dades i àrees afins.

• El projecte ha de permetre a l'estudiant del TFC adquirir nous coneixements en els àmbits dels ETL i els OLAP.

Abast del producte:

• El sistema proporcionarà un conjunt prefixat d'informes on es mostri la informació sol·licitada, així com la que sigui requerida pels investigadors. Entre la informació sol·licitada, hi figura la següent:

- Temes més parlats
- Els usuaris de twitter més actius de l'esdeveniment
- Els més seguits en temes concrets
- Els recursos compartits en tweets
- Els hashtag coincidents en tweets
- Les activitats dels assistents, tant si s'envien o reenvien tweets com si es segueixen a d'altres
- Evolució del nombre de tweets al llarg del temps

• L'aplicació tindrà almenys dos perfils d'usuari. Hi haurà un usuari consultor que accedirà a les dades, executarà consultes i emetrà informes. L'usuari administrador tindrà accés a totes les tasques anteriors i a més podrà executar les càrregues de noves dades i dissenyarà noves consultes.

• El sistema configurarà el seu API (*Application Programming Interface*) per minimitzar les piulades no relacionades amb el tema del congrés.

Requisits funcionals i de dades:

• El sistema ha de permetre tractar les piulades fins al punt de reconstruir parcialment una conversa, sabent qui és el pare d'una piulada però no pas els seus fills.

6.1.2. Requeriments no funcionals

A partir de la llista predefinida que figura a la plantilla Volere, es distingeixen els següents requeriments no funcionals:

Requisits de presentació:

• Les captures de piulades s'han de fer amb una eina a tal efecte, com ara "140dev Streaming API Framework".

Requisits d'usabilitat i d'humanitat:

• Les dades que s'emmagatzemin han de pertànyer als dies anteriors i posteriors a l'esdeveniment que s'estudia (en el nostre cas, al congrés).

Requisits operacionals i d'entorn:

• El magatzem de dades ha de ser construït amb tecnologia relacional. Això vol dir que el seu model de dades ha de basar-se en la lògica de predicats i en la teoria de conjunts, essent la seva idea fonamental l'ús de les relacions.

• El producte final que en resulti ha de realitzar-se sobre una màquina virtual Amazon que proporciona la UOC.

• El sistema desenvoluparà un entorn gràfic que permetrà consultar la informació de forma agregada per usuaris, tweets, recursos compartits, hashtags (etiquetes), paraules clau i converses.

Requisits de compliment:

• Les dades han de ser processades a nivell de dia, hora i minut.

Requisits de manteniment i suport:

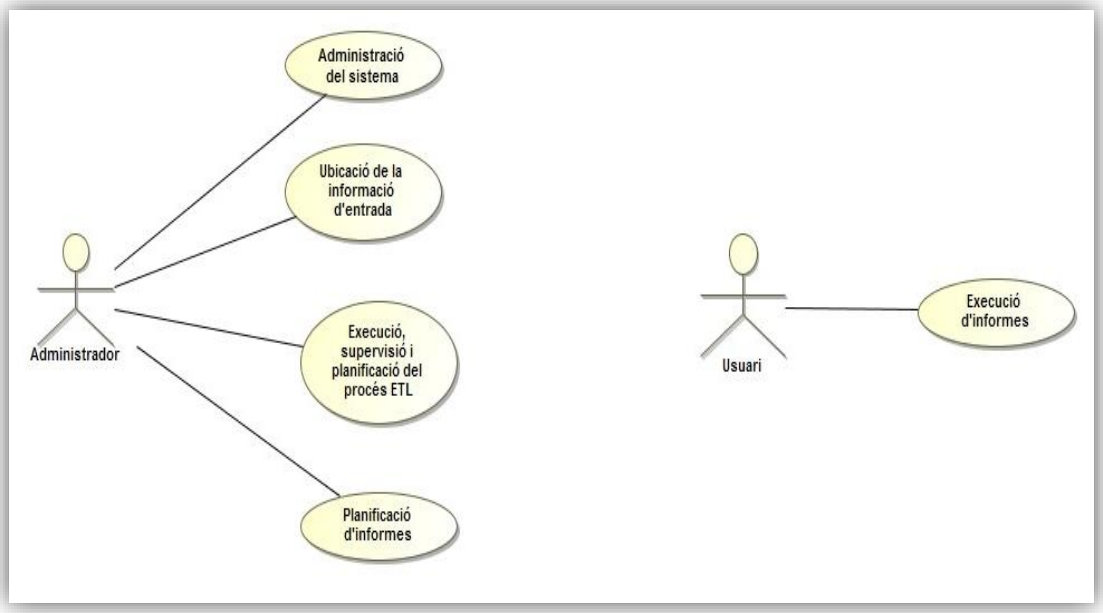
• Tot i que sempre estaran en format text, cal analitzar les fonts de dades operacionals proporcionades que han de servir per carregar cadascun dels elements d'anàlisi. La justificació d'aquest fet és que les dades en format text poden tenir diferents arxius de procedència (ODT, MDB, CSV, XLS i d'altres) i les dades es poden expressar de formes diferents per expressar un mateix concepte (hores, dates, dades geogràfiques de longitud i latitud...).

Requisits legals:

Les dades que reculli el sistema han de pertànyer a la conferència internacional duta a terme a València del 28 de setembre al 3 d'octubre de 2014, a mode d'estàndard que el nostre sistema a desenvolupar ha de complir.

6.2. Diagrama de casos d'ús amb una explicació de cada cas d'ús

Distingim dos actors amb perfils diferents al nostre sistema: l'administrador i l'usuari.



6.2.1. Perfil Administrador

• **Administració del sistema:** consisteix a administrar el sistema operatiu, l'eina d'anàlisi i el servidor de la BD, creant els usuaris i administrant els permisos.

• **Ubicació de la informació:** ubica la informació en els directoris a tal efecte per tal que es processi i es consolidi en el magatzem de dades per l'ETL.

• **Execució, supervisió i planificació ETL:** execució de forma manual o planificada dels processos ETL. Supervisió del correcte funcionament i recepció dels correus amb els resultats de l'execució del sistema.

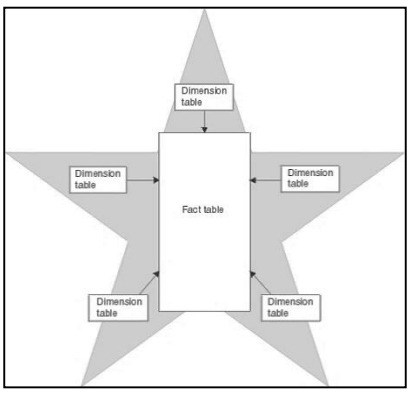
• **Planificació d'informes:** configuració dels destinataris de la distribució d'informes, que serà manual o planificada.

6.2.2. Perfil Usuari

• **Execució d'informes:** execució d'informes predefinits utilitzant els paràmetres que consideri oportuns a tal efecte. Es preveu que en futurs lliuraments els usuaris elaborin els seus propis informes.

6.3. Model conceptual

En aquest apartat ens ocuparem de fer el disseny conceptual multidimensional del projecte que hem de construir. Aquest disseny s'articula en dos components bàsics que són el Fet i el seu corresponent conjunt de Dimensions, de tal manera que l'estructura que en resulta adopta una forma característica que s'acostuma a anomenar Estrella. Aquesta Estrella també es pot representar com un paquet i el que farem en aquesta secció és ocupar-nos de dissenyar el que hi ha a dins d'aquest paquet o Estrella. Aquest disseny consisteix en un procés iteratiu de passos que descriurem tot seguit. La següent figura ens mostra doncs l'esquema global o estructura que es vol construir amb aquest tipus de disseny:



6.3.1. Tria del fet

La tria del Fet objecte d'anàlisi és el primer pas en el disseny d'una Estrella, atès que determina el tema que s'estudiarà amb l'Estrella. La concreció del Fet determina l'èxit del model proposat i la consecució dels objectius establerts. El Fet s'ha d'obtenir a

partir de l'anàlisi dels processos de negoci, entre els quals cal triar-ne el que s'adapti millor a les nostres necessitats i sigui prou significatiu. Un Fet és un conjunt d'esdeveniments amb dades numèriques associades, essent candidats oficials els processos de negoci (processos operacionals de l'organització suportats per algun sistema informàtic del qual se'n poden extreure dades). El Fet és un indicador de negoci, és a dir tota mesura o dada numèrica que hem d'incloure en el nostre sistema Business Intelligence. El Fet és la peça central de l'esquema multidimensional i ha de contenir els valors de les mesures de negoci o indicadors de negoci.

Centrant-nos en el cas que ens ocupa, és un grup d'investigadors de la UOC que participa en el projecte MAVSEL qui ens determina el Fet o objecte d'anàlisi del nostre disseny: les interaccions generades via Twitter en un esdeveniment d'àmbit científic com és la conferència internacional duta a terme a València del 28 de setembre al 3 d'octubre de 2014 sota el nom *Model Driven Engineering Languages and Systems* (MODELS). Entenem com a tweets o piulades les interaccions generades via Twitter i destaquem que el mateix grup d'investigació també ens proporciona el seguit de dades a relacionar amb les piulades: usuaris, possibles etiquetes associades als tweets, localitzadors URL, converses que s'hi associen i possibles emmagatzematges en memòria cau.

6.3.2. Tria del gràNul escaient

El gràNul és l'individu últim o tipus d'objecte concret que es vol analitzar. Els més típics són les transaccions individuals (per exemple, les piulades o tweets) o les instàncies diàries d'un estat. Acostumen a ser esdeveniments de qualsevol tipus que es donin amb freqüència relativament alta. La tria del gràNul associat al Fet és quelcom importantíssim, atès que determina la dimensió de la base de dades i té un impacte directe en la grandària del conjunt de dades. Un gràNul és més gran conforme representa més elements i un bon disseny demana triar-ne el més petit possible, per no perdre la possibilitat de calcular cap dada derivada amb el mínim error. Però cal fer la tria de forma assenyada i no excedir-se en una mida del gràNul massa petita, atès que triar un gràNul massa petit pot suposar malbaratar espai o fer inviable el projecte per excés de dades (un gràNul molt petit implica una base de dades molt gran, no sempre viable).

A partir de la informació lliurada a l'enunciat del projecte i en el document *Font de Dades*, el Fet més simple o gràNul candidat de la Nostra Estrella pot ser "tweet emmagatzemat en memòria cau associat a un cert usuari, identificat amb una etiqueta que el relaciona (o no) amb el tema del congrés i localitzat en una URL, esdevingut en un cert instant de temps o moment".

6.3.3. Tria de les dimensions que s'utilitzaran en l'anàlisi

Les dimensions representen els diferents punts de vista que s'utilitzen en l'anàlisi de les dades. Són variables independents que afecten cada observació. Les dimensions són els elements que contenen els atributs o camps que es fan servir per a restringir i agrupar les dades emmagatzemades en una taula de fets quan es realitzen consultes sobre aquestes

dades en un entorn de magatzem de dades o data mart. Les dimensions ajuden a estudiar/analitzar les dades afegint-hi informació sobre les dades de la taula de fets, de manera que es pot dir que en un cub OLAP el Fet conté les dades d'interès i les dimensions contenen metadades sobre aquest Fet.

A partir de la mateixa definició del gràNul que hem fet abans, podem establir el següent primer conjunt de Dimensions d'anàlisi: Moment, Emmagatzematge, Usuari, Etiqueta i Localitzador. A partir d'aquest primer conjunt inicial de Dimensions, observem que una combinació de les instàncies d'Usuari i de Localitzador permet afegir una nova Dimensió, que és el cas de Conversa. Vegem-ne per sobre cadascuna d'elles per separat:

- Moment. Aquesta dimensió és gairebé omnipresent en qualsevol model multidimensional i pot ser útil en l'elaboració d'informes que responguin a preguntes que requereixen localitzar tweets esdevinguts en un interval de temps. Atès que el seu comportament és similar al d'una dimensió tipus data, considerarem que només conté un sol atribut i a l'anàlisi físic formarà part de la taula del Fet.

- Emmagatzematge. Aquesta dimensió recull informació sobre aspectes de memòria cau referits als diferents tweets, com ara dades en format d'imatge i elements de programació.

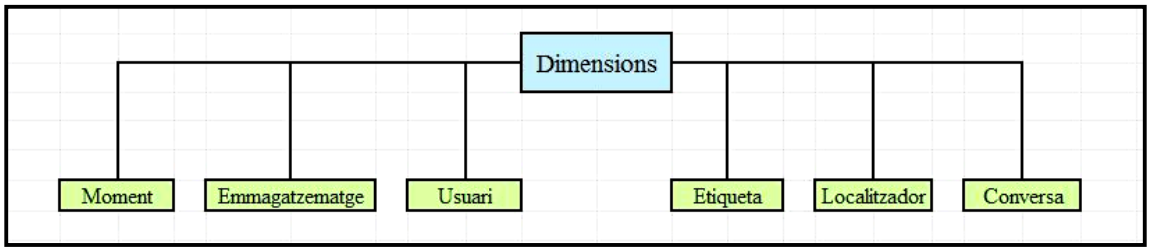
- Usuari. Una de les dimensions més importants, atès que recollirà dades interessants per a majoria d'informes relatives a l'autoria, referències i mencions a les piulades.

- Etiqueta. Es tracta d'una dimensió que se n'ocupa d'associar totes les piulades relacionades amb el congrés, així com marcar les que no en tenen res a veure.

- Localitzador. Aquesta dimensió es refereix al tipus de localitzador URL associat a un tweet i pot esdevenir útil com a criteri per a consolidar la informació dels mateixos en el moment d'elaborar informes.

- Conversa. Aquesta dimensió no és de trobada immediata, sinó el resultat de combinar Usuari i Localitzador. Constitueix una de les dimensions més importants, atès que recollirà dades que es poden agregar per una bona part dels informes referents a interaccions que entre tweets, com ara mencions, referències o respostes.

De tot el que s'ha dit es deriva el següent esquema que recull les dimensions del model conceptual:



6.3.4. Identificació dels atributs de les taules de Dimensions

Havent determinat les dimensions que s'utilitzaran en l'anàlisi, ara cal establir els atributs i les jerarquies, si s'hi escau, per cadascun dels registres de la taula de Fets. Els atributs que pertanyen a les Dimensions són útils per a triar i descriure l'espai d'anàlisi. Es tracta de seleccionar qualsevol atribut que considerem útil per a seleccionar, agrupar o simplement posar com a capçalera d'un informe. Els atributs de Dimensions es defineixen sobre un domini discret, són generalment descriptius, fàcils de recordar i entenedors a primer cop d'ull. Amb tota aquesta informació a mode de recolzament i suport teòric, els atributs que en resulten per les diferents Dimensions són els que presento tot seguit:

• Dimensió Moment. Aquesta dimensió consta d'un sol atribut. És el que s'anomena una dimensió degenerada. Les dimensions degenerades no existeixen físicament i s'acaben convertint en una columna de la taula de fets. Aquesta dimensió és bàsica per qualsevol model, atès que el temps (o un moment en el temps) sempre és una de les perspectives requerides per analitzar la informació.

• Dimensió Emmagatzematge. Dimensió contenidora d'elements o aspectes a tenir en compte de memòria cau referents a l'emmagatzematge de les piulades. Establirem els següents atributs:

- tweet_id. De tipus Enter i clau primària compartida. És també una clau forana que fa referència a la clau primària tweet_id de la taula de Fet Tweet.
- cache_id. De tipus Enter i clau primària compartida amb tweet-id. Es tracta d'una clau auto-generada que va augmentant el seu valor d'un en un conforme es va omplint la taula a partir de la informació de les fonts.
- raw_tweet. De tipus Text i no nul. Atribut referit a un format d'imatge.

• Dimensió Usuari. S'utilitza per a identificar els usuaris que d'una forma o altra s'involucren amb les piulades. Aquesta dimensió es carregarà en el moment de crear el model en el motor de la base de dades. Presenta un ventall d'atributs considerable:

- usuari_id. Tipus Enter i clau primària. Identifica l'usuari de forma única i els diferencia dels altres.
- àlies. Tipus Text. Exerceix de nom d'usuari amb el qual s'identifica un usuari en

la piulada. Pot ser un camp interessant a l'hora de generar informes si volem distingir o classificar tweets per àlies o nom d'usuari.

- nom. Tipus Text. Recull el nom complet de l'usuari, ja sigui una persona o bé una institució. És un camp molt interessant per seleccionar o agrupar tweets sota el nom complet d'un usuari, fet que esdevé molt més formal que seleccionar-los per àlies d'usuari.

- imatgePerfilURL. Tipus Text. Camp que fa referència a com veuen els altres usuaris a un de donat en una piulada, a partir d'una imatge o fotografia fixa que s'associa a l'usuari.

- lloc. Tipus Text. Algunes files de la taula Usuari presenten valors nuls en aquest atribut. Pot generar informes molt interessants respecte als usuaris que es localitzen en una ciutat i han participat al congrés intervenint a les piulades.

- url. Tipus Text. Pot donar-se el cas que presenti valors nuls. Es tracta de l'identificador de recursos uniforme associat a cada usuari, format per una seqüència de caràcters d'acord amb un format modèlic i estàndard que designen recursos en una xarxa, com ara Internet.

- descripció. Tipus Text. Pot presentar-ne també valors nuls. Es tracta d'un text que ens presenta i descriu breument l'usuari (càrrec, professió...).

- recompteSeguidors. Tipus Enter. Indica quants usuaris són seguidors dels que ens ocupa. Informació valuosa per fer informes estadístics amb perfil *rànking*.

- recompteAmics. Tipus Enter. Atribut de característiques similars al d'abans, però Ara indica quants usuaris són amics d'un de donat objecte d'estudi.

- recompteEstats. Tipus enter. Atribut de característiques semblants als d'abans, tot i que ara s'indica els estats de compte d'un cert usuari. També pot ser útil per generar informes estadístics amb perfil de *rànking*.

- zona_Horària. Atribut de tipus Text. Camp que pot presentar valors nuls a la taula de Dimensió Usuari. Conté la referència del lloc (ciutat) on s'hi mesura l'hora.

• Dimensió Etiqueta. Aquesta dimensió pot ser de gran utilitat per elaborar informes sobre piulades que hagin de ser referides pel seu sobrenom o àlies en comptes de pel seu identificador. Podríem tenir interès a conèixer quantes piulades responen a l'etiqueta de "models14" o "AirFrance", per citar un parell d'exemples. Aquesta dimensió consta dels següents atributs:

- tweet_id. De tipus Enter, clau primària de la taula de Dimensió Etiqueta i clau forana que referencia a la clau primària del mateix nom de la taula de Fets de les

piulades o tweets. Identifica l'etiqueta de forma única.

- nomEtiqueta. De tipus Text. És el camp que conté l'etiqueta associada al tweet o piulada.

• Dimensió Localitzador. Aquesta dimensió esdevindrà útil per reportar informes estadístics sobre piulades que es vulguin estudiar sota la perspectiva del seu localitzador URL. Els atributs d'aquesta dimensió que poden ser útils per posar-los com a capçalera d'informes són els següents:

- tweet_id. Tipus Enter i clau primària. Clau forana que referencia a la clau primària d'idèntic nom a la taula de Fets de les piulades o tweets. Identifica una instància de Localitzador de manera única i exclusiva.

- url. Tipus Text. És l'atribut que aporta el localitzador URL a la piulada i dóna un sentit propi a la Dimensió Localitzador.

• Dimensió Conversa. Ja hem comentat abans que aquesta dimensió és el resultat de combinar Usuari i Localitzador, o també Usuari i Etiqueta. La seva importància a l'hora d'elaborar informes és cabdal, atès que recull tot el ventall d'interaccions entre piulades com ara mencions, referències o respostes. Vegem quins són els atributs que esdevenen útils per agrupar, seleccionar o per posar com a capçalera d'informes:

- tweet_id. Tipus Enter i clau primària compartida juntament amb els dos atributs que mencionaré tot seguit. Clau forana que referencia a tweet_id de la taula de Fet tweets.

- fontUsuari. Tipus Enter i clau primària compartida. Clau forana que fa referència a l'atribut usuari_id de la taula de fet tweet. La informació que conté és la identificació d'usuari del tweet al qual en fa menció.

- usuari_Referenciat. Tipus enter i clau primària compartida. Clau forana que fa una referencia a usuari_id de la taula de Dimensió Usuari.

En aquesta Dimensió Conversa he decidit agafar tots tres atributs com a clau primària, atès que considero que tota conversa queda del tot identificada pel fet d'anar referida a un sol tweet que ha estat elaborat per un cert usuari, però que l'usuari que el comenta no ha de ser necessàriament (de fet, no ho serà en la majoria de casos) el mateix que ha creat el tweet. Estem en el cas d'una clau primària composta de tres atributs que identificarà de forma inequívoca cada registre de la taula.

6.3.5. Decidir quines són les mesures que interessen

Les Mesures són atributs numèrics que es poden sumar i normalment additius, de tal manera que com més Mesures contingui un Fet més útil esdevé l'Estrella de la qual en formi part. Poden haver-hi també Fets sense Mesures molt interessants pel nostre estudi. Les Mesures que consideraré per a respondre a una part de les necessitats del grup d'investigació de la UOC que participa en el projecte MAVSEL són les següents:

• Recursos compartits a la piulada. Es processarà la variable nombre de recursos compartits al tweet, en resposta a un dels requeriments respecte al coneixement que s'ha d'extreure de la interacció entre piulades. Aquests recursos tanmateix poden consistir en una etiqueta com una imatge del tweet o bé dades de memòria enregistrades.

• Temps transcorregut des de l'inici de la conferència (primera piulada enregistrada). Es processarà el temps mesurat en minuts que ha transcorregut des de la primera piulada enregistrada al Congrés fins a la data de creació del tweet objecte d'estudi. L'objectiu d'aquesta mesura és fer possible els informes adients referents a l'evolució del nombre de tweets al llarg del temps. Es tracta doncs d'una Mesura derivada, resultat de fer una diferència entre el moment de creació de la piulada i el moment en què es va enregistrar la primera piulada.

Aquestes Mesures es veuran reflectides a la taula de Fets en forma d'atributs. Llur combinació seguint diferents criteris de selecció i fent les agrupacions adients hauríem de ser capaços d'elaborar una part dels informes que planteja el grup d'investigadors de la UOC participant al projecte MAVSEL. Com que es tracta en tots dos casos d'atributs numèrics i de caire additiu que facilita derivar-ne els respectius valors totals, la informació requerida pot ser mostrada en format de gràfic, complementant-ne així la informació que s'acostuma a presentar en formats més tradicionals d'informes o fulls de càlcul.

6.3.6. Identificació dels atributs de la taula de Fets

Havent-hi determinat les Dimensions del model conceptual, la taula de Fets ens ha de proporcionar la informació referent a cadascuna de les combinacions dels atributs de la clau primària de l'esmentada taula, considerant que l'atribut temporal, en el cas que ens ocupa "Hora exacta d'un moment en què es crea un tweet" ha d'exercir de pivot, de manera que si no s'inclou aquesta informació en una determinada consulta aquesta omissió pot ocasionar resultats degradats o erronis. En aquest sentit, l'exemple de consulta més representatiu que ha de ser contestat amb la taula de fets és l'evolució del nombre de tweets al llarg del temps, que ja s'havia comentat abans.

Hi ha però tot un seguit de consultes que tenen sentit i són del tot legítimes sense haver de considerar de forma explícita la dimensió temporal, algunes de les quals s'indiquen a l'enunciat mateix del projecte:

• Els temes més parlats

• Els usuaris de Twitter més actius de l'esdeveniment
• Un cop escollit un tema (havent triat un *hashtag*), entre tots els usuaris que han "usat" aquest tema (tweet, retweet o reply) quins tenen més seguidors
• Els recursos compartits en tweets i els seus totals
• Els *hashtag* coincidents en tweets
• Els tweets coincidents en localització geogràfica (latitud, longitud o totes dues)
• Les activitats dels assistents, tant si envia / reenvia tweets i / o segueix a d'altres.
• El total d'activitats de cada tipus que s'ha citat abans (enviament, reenviament o seguiment), així com el total d'activitats de cap d'aquests tipus

En funció de la informació demanada pel grup d'investigadors de la UOC la taula de Fets contindrà la informació necessària per l'elaboració dels informes requerits. La taula de Fets serveix per maximitzar el rendiment de les consultes realitzades i d'acord amb la bibliografia bàsica presenta els següents avantatges:

- Millora del rendiment eliminant unions entre taules, atès que la unió entre taules és l'operació que presenta un cost més gran en l'àmbit de les bases de dades.
- El seu ús intuïtiu, atès que per un usuari sense perfil de programador la informació bàsica es troba en una sola taula i això li permet oblidar-se'n de les relacions i restriccions d'un model relacional tradicional.

Com a conseqüència de tot el que s'ha dit, la meva proposta d'atributs per la taula de Fets Piulada és la que mostro tot seguit a la taula següent:

Tipus	Atribut	Descripció de l'atribut
PK	dataCreacio	Descriu el moment exacte que la piulada ha estat creada
	tweet_id	Identificador únic de la piulada
	cache_id	Identificador únic d'emmagatzematge cau de la piulada
	usuari_id	Identificador únic de l'usuari que interactua amb la piulada
	latitud	Conté la latitud geogràfica de l'usuari emissor de la piulada
	longitud	Conté la longitud geogràfica de l'usuari emissor de la piulada
	nombreRecursosCompartits	Indica quants recursos comparteix la piulada
	/totalMinutsTranscorreguts	Temps transcorregut, mesurat en minuts, des de la primera piulada enregistrada fins al moment de creació de la piulada
	is_retweet	Indicador de si el tweet "retweeteja" o no un altre
	is_reply	Indicador de si el tweet replica o no un altre
	is_not_retweet/reply	Indicador de si el tweet entra o no en cap de les categories "retweet" o "reply"

Una vegada assignats els atributs de la taula de Fets, és el moment de comentar un parell d'aspectes molt interessants del disseny conceptual com són la definició de cel·les adients i explicitar les restriccions d'integritat. Vegem-ne cadascun per separat:

6.3.7. Definició de Cel·les

En el conjunt de Mesures triat abans s'ha assignat una granularitat de l'ordre del minut a l'atribut "totalMinutsTranscorreguts". De tot el conjunt de possibles cel·les amb les seves granularitats hem de desestimar les que no són interessants pel model, d'acord amb les seves especificacions, i per tant no han de ser emmagatzemades en ser derivades de les principals. Per la solució proposada contemplem una única cel·la que es refereix al fet especificat i que es referirà a les dades de les piulades enregistrades en un moment determinat, associades a un usuari i emmagatzemades i localitzades al sistema sota una determinada etiqueta o *hashtag* i un localitzador URL.

6.3.8. Explicitar les restriccions d'integritat

Expressar les restriccions d'integritat és l'últim pas del disseny conceptual, essent en aquest punt especialment important destacar les Bases. En aquest sentit, el conjunt inicial de Dimensions – Moment, Emmagatzematge, Usuari, Etiqueta i Localitzador – establert en definir el gràNul ja constitueix en si mateix una Base. Teòricament, substituint Dimensions en aquesta Base segons les dependències funcionals que hi ha amb la resta de Dimensions, es podrien trobar les altres bases de l'espai. A la pràctica, però, si ens plantegem afegir la Dimensió Conversa a la base de "Piulada Atòmica" {Moment, Emmagatzematge, Usuari, Etiqueta i Localitzador} ens adonem que aquesta no forma part de la Base, atès que queda completament determinada per les Dimensions Moment, Usuari i Localitzador.

A més, les restriccions han de tenir present les dimensions que hem establert de manera que les consultes que es realitzin siguin coherents i ens aportin el coneixement requerit. En particular, la restricció que hem de considerar és sobre la Dimensió Etiqueta, en el sentit que s'ha intentat minimitzar les piulades no relacionades amb la temàtica del congrés. Així, encara s'hi detecten piulades que no tenen res a veure amb la temàtica desitjada, degut a una elecció poc afortunada en el nom del *hashtag* "models14". També s'apliquen restriccions a la Dimensió de Conversa, en el sentit que no es pot reconstruir en la seva totalitat una conversa perquè se sap qui és el pare d'una piulada, però no pas els seus fills.

6.3.9. Estudi de la viabilitat

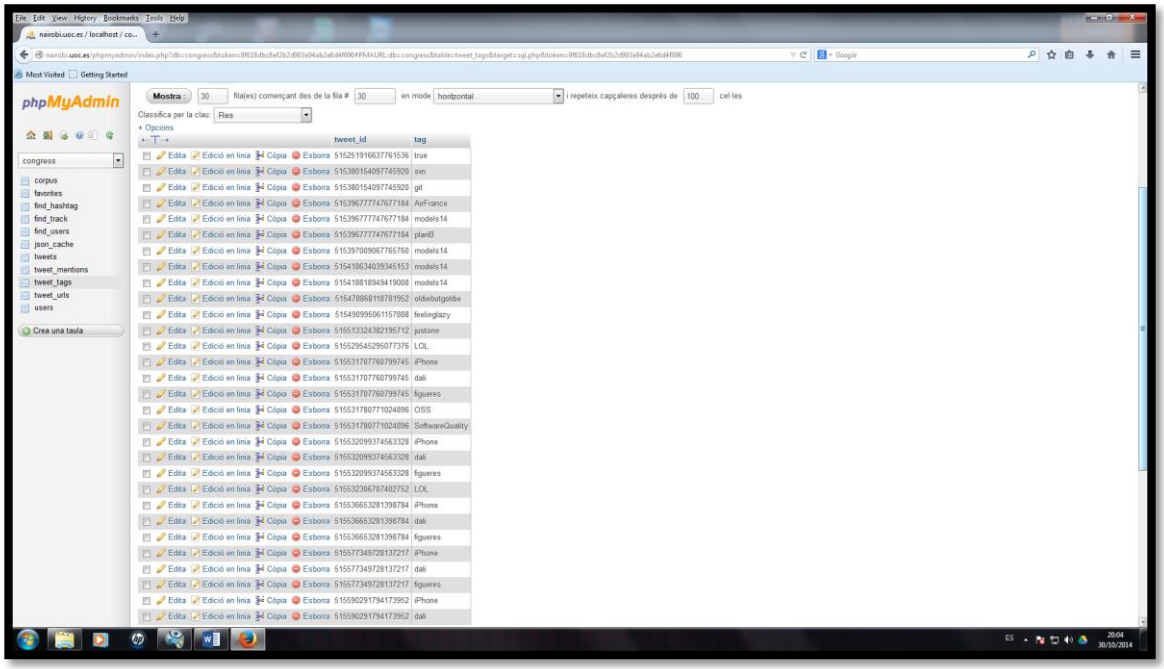
Arribats en aquest punt, ja hem deixat enrere el disseny conceptual i es tracta de determinar si l'Estrella que en resulta es pot implementar o no. Estudiar la viabilitat és estimar l'espai que ocupen totes les dades. Per a saber aquest espai, el més realista és mirar les dades que conté el nostre sistema operacional o font de dades, però això no sempre és possible. Una possible alternativa és fer una estimació a partir del nombre d'instàncies del Nivell atòmic de cada Dimensió i la mida de la cel·la. És a dir que el que fem és considerar només el que ocupa emmagatzemar les instàncies del Fet, atès que el volum de dades que ocupen les Dimensions és insignificant comparat amb el del Fet.

En el nostre cas particular, la base de dades MySQL accessible a través de l'adreça http://nairobi.uoc.es/phpmyadmin ens aporta la següent informació:

• Tenim 2136 instàncies diferents d'Emmagatzematge (taula *json_cache* a la base de dades) tal i com mostra la següent imatge:

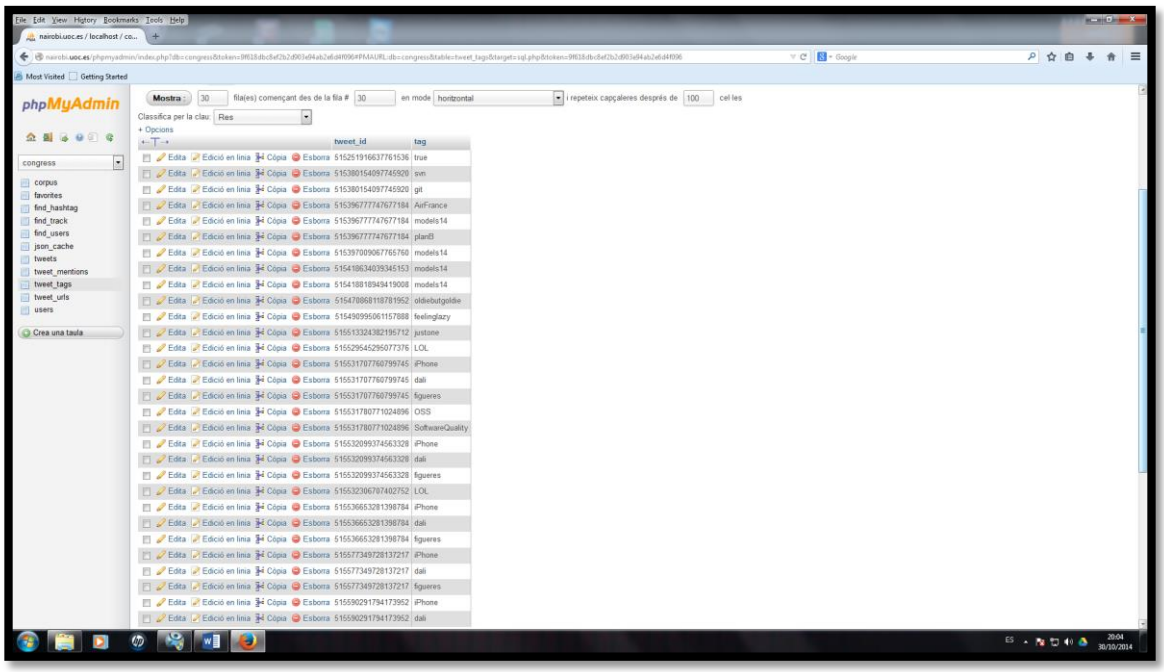
• Tenim 2276 instàncies diferents d'Etiqueta (taula *tweet_tags* a la base de dades), tal i com es mostra a la imatge:

La següent imatge en mostra els 30 primers registres:



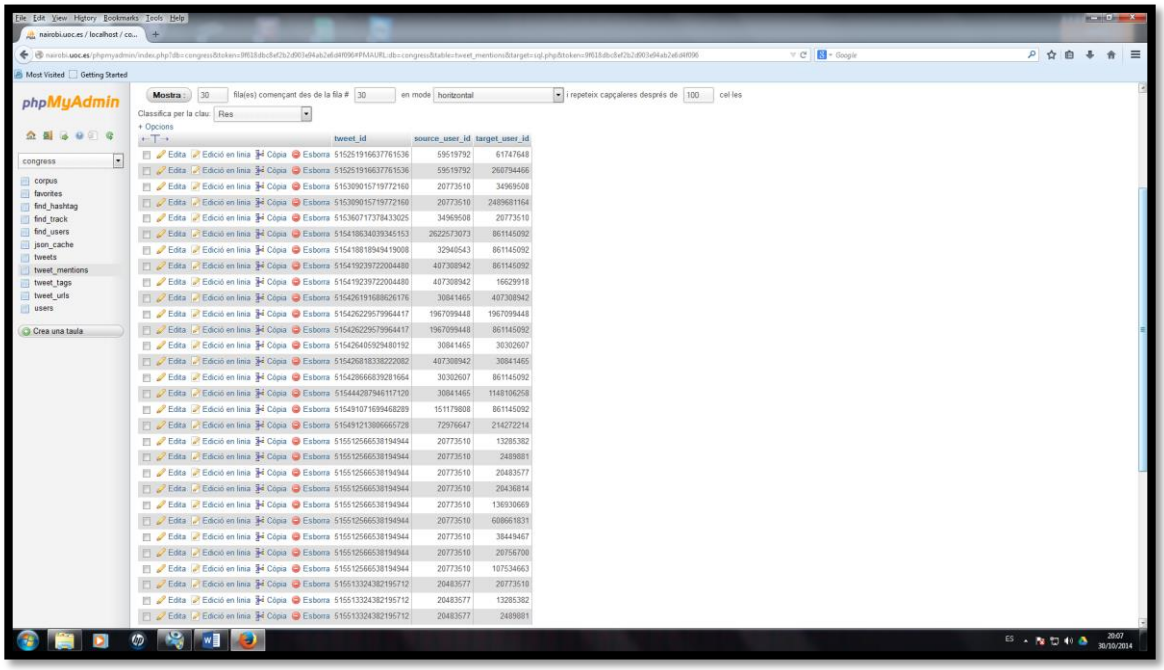
• Tenim 341 instàncies diferents de Localitzador (taula *tweet_urls* a la base de dades), tal i com mostra la següent imatge:

I vet aquí una mostra dels 30 primers registres i de com queden distribuïts a la taula:



• Tenim 1454 instàncies diferents de conversa (*tweet_mentions* a la base de dades), tal i com es mostra a la següent figura:

Vet aquí una imatge dels 30 primers registres a la taula:



Prenent la Base {Emmagatzematge, Usuari, Etiqueta, Localitzador} obtenim un espai de 2136 · 446 · 2276 · 341 = 739.371.564.096 possibles cel·les.

Un cop sabem quantes cel·les tenim i a efectes d'obtenir una estimació del que ocuparà la nostra Cel·la, multipliquem aquesta xifra pel nombre de bytes que ocuparà cada cel·la (obtingut a partir dels bytes que ocupen les Mesures i dels bytes que ocupen els identificadors de les 5 Dimensions que he citat abans). Per calcular el nombre de bytes que ocupa cada cel·la, considero el següent:

- L'identificador de la Dimensió Moment ocupa 8 bytes, atès que és de tipus *datetime*.
- Els identificadors de la Dimensió Emmagatzematge ocupen 8 i 4 bytes, atès que són del tipus *bigint* i *int, integer* respectivament. En total, 12 bytes.
- L'identificador de la Dimensió Usuari ocupa 8 bytes, atès que és de tipus *bigint*.
- L'identificador de la Dimensió Etiqueta ocupa 8 bytes, atès que és de tipus *bigint*.
- Els identificadors de la Dimensió Conversa ocupen tots tres 8 bytes, atès que són tots tres del tipus *bigint*. En total en resulten 24 bytes.
- La Mesura latitud ocupa 1 byte, atès que és de tipus decimal (10,5).
- La Mesura longitud ocupa 1 byte, atès que és de tipus decimal (10,5).
- La Mesura de total de minuts transcorregut ocupa 4 bytes, atès que la podem considerar de tipus *float.*
- Les Mesures *is_retweet*, is*_reply* i *is_not_retweet/reply* ocupen cadascuna y byte, atès que són totes tres del tipus *tinyint*. En total, això representa 3 bytes.

Amb tota aquesta informació, la nostra Cel·la ocuparà:

(8 + 12 + 8 + 8 + 24 + 2 +4 +3) · 739.371.564.096 bytes = 69 · 739.371.564.096 bytes =

= 51.016.637.922.624 bytes = 47.512,95 Gbytes.

Aquest xifra indica un volum de dades impossible de manejar en el nostre sistema, atès que la velocitat de resposta seria extremadament lenta i seria molt difícil carregar dins la finestra d'actualització totes les dades necessàries per actualitzar-lo. La solució passa per triar un gränul no tan fi i descartar algunes Mesures que no són massa essencials, com ara la latitud i la longitud geogràfiques. Si considerem la base {Usuari, Etiqueta, Localitzador} (gränul tipus "tweet associat a un cert usuari, identificat amb una etiqueta que el relaciona amb el tema del congrés i localitzat en una URL, esdevingut en un cert instant de temps o moment"), l'espai de possibles cel·les resulta ser 446 · 2276 · 341 = 346.147.736 cel·les. Atès que la taula *tweets* presenta molts registres sense dades de longitud i latitud (els hi atorga el valor zero), podem prescindir d'aquestes mesures i llavors cada cel·la ocupa 67 bytes. Amb aquesta nova informació, ens resulta una Cel·la de mida:

67 · 346.147.736 bytes = 23.191.898.312 bytes = 21,6 Gbytes

Aquesta darrera mida és molt més raonable i ens fa pensar que la podrem emmagatzemar al nostre sistema.

Com a darrer exemple, si considerem un gränul encara menys fi de Base {Usuari, Etiqueta} ("tweet associat a un cert usuari, identificat amb una etiqueta que el relaciona amb el tema del congrés esdevingut en un cert instant de temps o moment"), l'espai de possibles cel·les resulta ser de 446 · 2276 = 1.015.096 cel·les. I la mida de la Cel·la que en resulta és:

67 · 1.015.096 bytes = 68.011.432 bytes = 64,86 Mbytes

Aquesta darrera mida és encara més tractable i s'acosta més al valor de 13,2 Mbytes que indica la base de dades de MySQL. La viabilitat és un punt que posa de manifest la importància de l'elecció del gränul escaient, atès que l'elecció d'un gränul massa fi pot fer inviable tot un projecte per problemes de malbaratament d'espai o excés de dades.

Per acabar aquesta secció de la viabilitat, en atorgar valors a cadascun dels identificadors i Mesures s'han utilitzat les següents taules provinents de la font http://dev.mysql.com/doc/refman/5.0/es/storage-requirements.html:

Requerimientos de almacenamiento para tipos numéricos

Tipo de columna	Almacenamiento requerido
`TINYINT`	1 byte
`SMALLINT`	2 bytes
`MEDIUMINT`	3 bytes
`INT`, `INTEGER`	4 bytes
`BIGINT`	8 bytes
`FLOAT(p)`	4 bytes si 0 <= **p** <= 24, 8 bytes si 25 <= **p** <= 53
`FLOAT`	4 bytes
`DOUBLE [PRECISION]`, objeto `REAL`	8 bytes
`DECIMAL(M,D)`, `NUMERIC(M,D)`	Varía; consulte la siguiente explicación
`BIT(M)`	aproximadamente (**M**+7)/8 bytes

Requerimientos de almacenamiento para tipos de fecha y hora

Tipo de columna	Almacenamiento requerido
`DATE`	3 bytes
`DATETIME`	8 bytes
`TIMESTAMP`	4 bytes
`TIME`	3 bytes
`YEAR`	1 byte

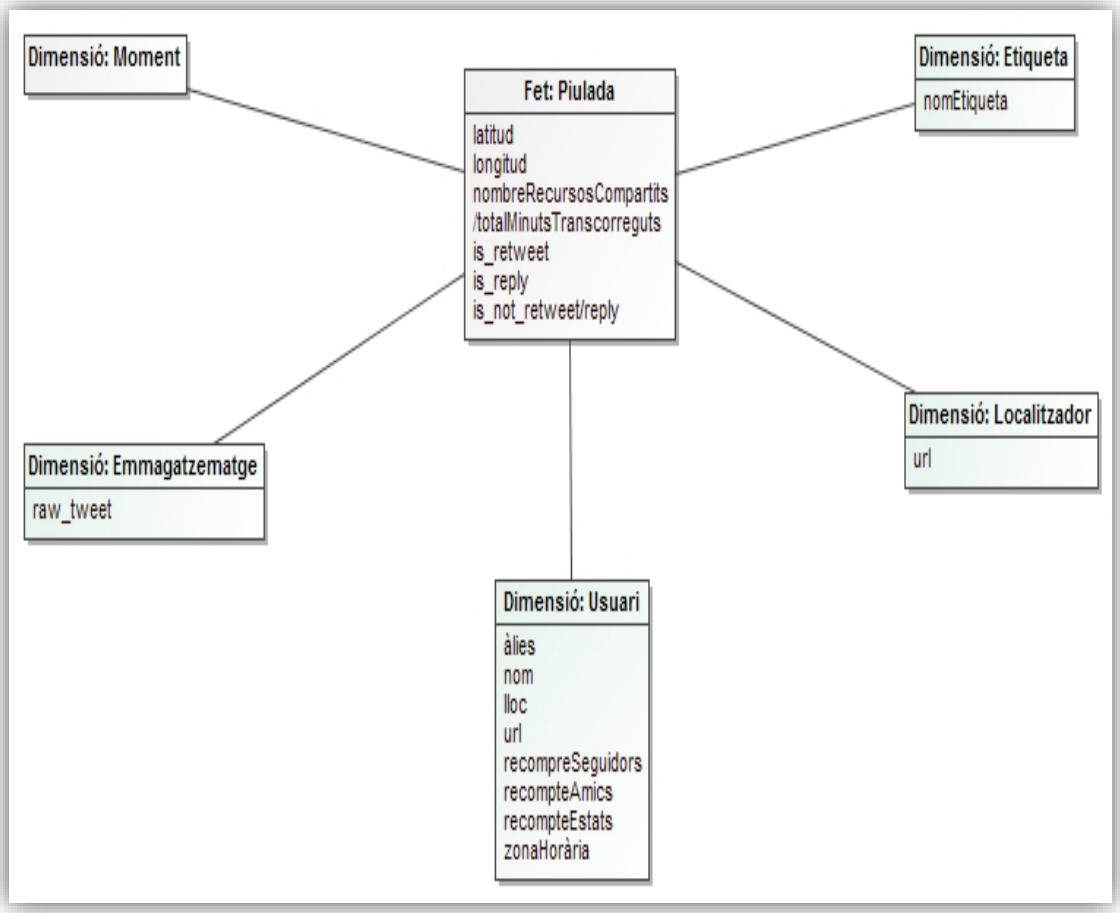
Requerimientos de almacenamiento para tipos de cadenas de caracteres

Tipo de columna	Almacenamiento requerido
`CHAR(M)`	`M` bytes, 0 <= `M` <= 255
`VARCHAR(M)`	`L`+1 bytes, donde `L` <= `M` y 0 <= `M` <= 255
`BINARY(M)`	`M` bytes, 0 <= `M` <= 255
`VARBINARY(M)`	`L`+1 bytes, donde `L` <= `M` y 0 <= `M` <= 255
`TINYBLOB`, `TINYTEXT`	`L`+1 byte, donde `L` < 2^8
`BLOB`, `TEXT`	`L`+2 bytes, donde `L` < 2^16
`MEDIUMBLOB`, `MEDIUMTEXT`	`L`+3 bytes, donde `L` < 2^24
`LONGBLOB`, `LONGTEXT`	`L`+4 bytes, donde `L` < 2^32
`ENUM('value1','value2',...)`	1 o 2 bytes, dependiendo del número de valores de la enumeración (65,535 valores como máximo)
`SET('value1','value2',...)`	1, 2, 3, 4, o 8 bytes, dependiendo del número de miembros del conjunto (64 miembros como máximo)

6.3.10. Esquema conceptual

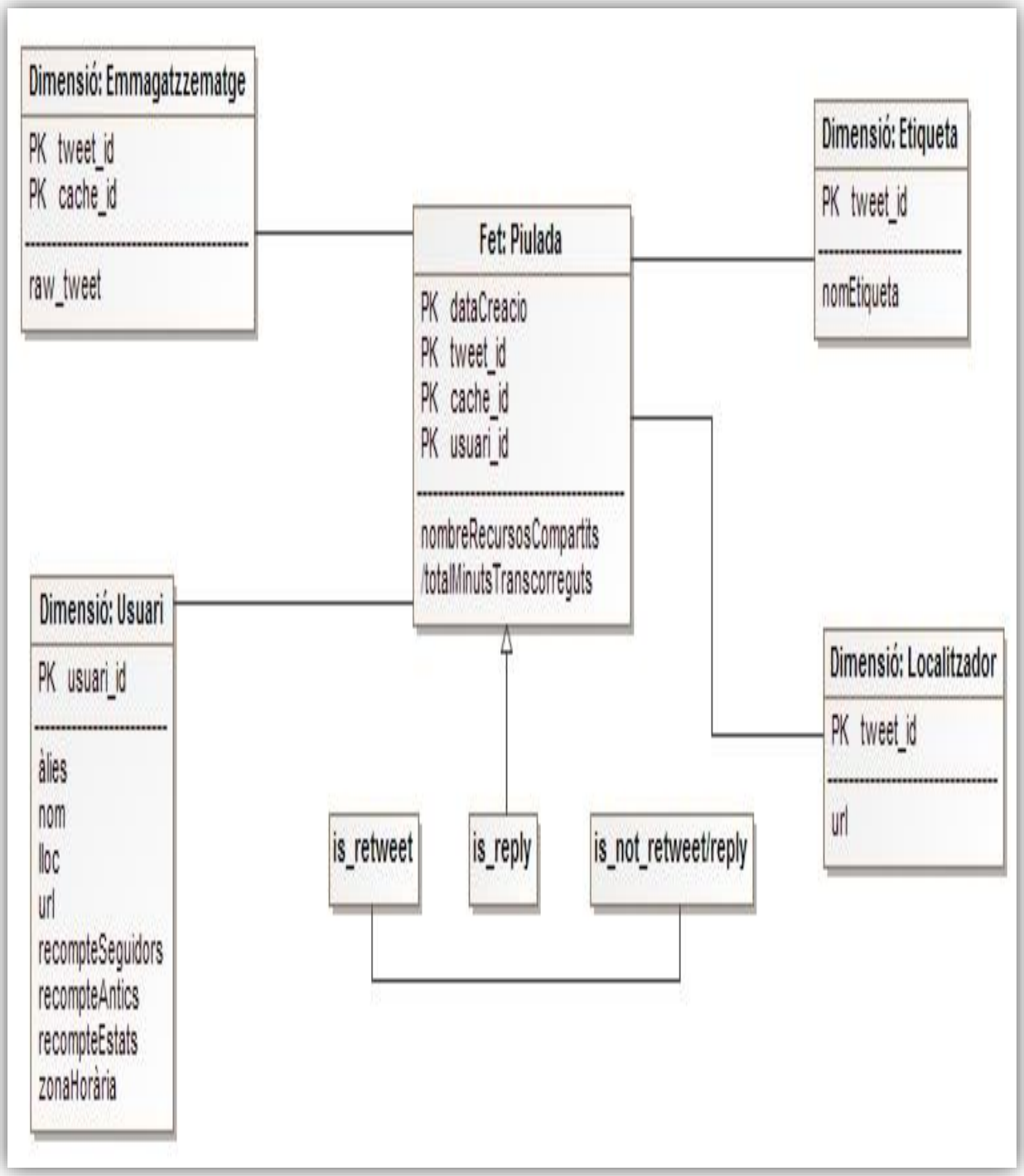
L'esquema conceptual és una representació simplificada de les característiques del model a implementar, on s'hi estableixen les relacions entre les taules i les seves dependències. Mitjançant l'esquema conceptual es pot realitzar una descripció d'alt nivell de la futura base de dades.

En l'esquema conceptual proposat en una primera instància, considerant un primer conjunt de dimensions o Base {Moment, Emmagatzematge, Usuari, Etiqueta, Localitzador} resultant de la primera definició del gràNul, es diferencia la dimensió temporal Moment (en el cas que ens ocupa un instant que correspon a la creació de la piulada) com una entitat independent. A la pràctica, el que faré serà incloure-la com un atribut independent de la taula de Fets. També es decidirà en una segona versió, més refinada, considerar que la taula de Fets inclogui només Mesures o atributs numèrics. Una primera versió de l'esquema conceptual és el que es mostra tot seguit:



A partir de l'anàlisi de la versió anterior de l'esquema conceptual podem eliminar la Dimensió moment, atès que és una dimensió degenerada (només disposa d'un sol atribut, és a dir, és un camp que es fa servir com a criteri d'anàlisi amb el mateix nivell de granularitat que les dades de la taula de Fets). També podem considerar, a efectes que la taula de Fets Piulada només tingui Mesures o atributs additius, que els atributs

latitud i longitud són negligibles atès el seu poc interès a l'hora de generar informes i que la majoria de tuples o registres de la taula *users* a la base de dades no disposen d'aquestes dades i se'ls hi atorga el valor per defecte de zero. A més, els atributs *is_retweet*, *is-reply* i *is_not_retweet/reply*, que són de tipus numèric però ho podrien ser de tipus booleà, es poden considerar especialitzacions de Piulada.



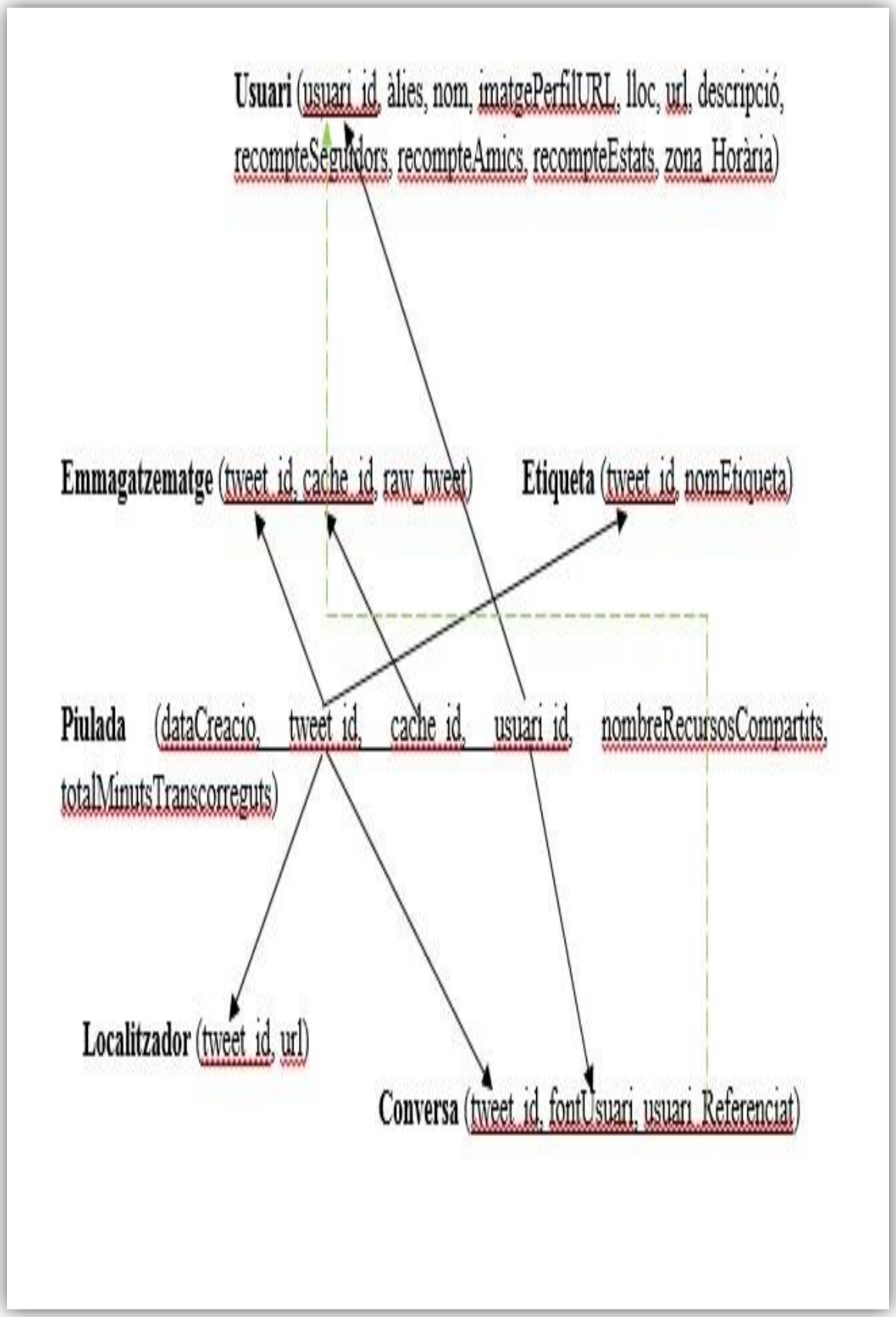
6.4. Disseny de la BD / Diagrama E-R

El disseny lògic o diagrama entitat relació és l'eina per la modelització de dades que permet representar les entitats rellevants d'un sistema d'informació, així com les seves interrelacions i propietats. Estem referint-nos doncs a l'eina que té per objectiu establir relacions entre les diferents taules que formen el nostre model.

Tal i com es va fer en el disseny conceptual, per passar al model lògic ens hem de fixar en l'Estrella amb una sola taula de Fets: la piulada o tweet. A més, en aquest apartat considerem que el Fet conté només una sola Cel·la. L'estratègia a seguir per a implementar una Estrella és fer servir una taula per al Fet (en què cada fila representa una cel·la de l'espai multidimensional) i una taula més per cadascuna de les Dimensions. Les jerarquies d'agregació queden implícites en els valors dels atributs de les taules de Dimensió, de tal manera que no els explicitem amb taules diferents. Llavors, lligarem la taula de Fet per claus foranes amb cadascuna de les taules de Dimensions, de tal manera que la clau primària de la taula del Fet sigui la concatenació de les claus foranes corresponents a una Base del Fet. La idea és que la taula del Fet contingui els atributs necessaris per a contenir les diferents mesures, mentre que les taules de dimensions contindran un seguit d'atributs descriptius de les operacions de la taula a la qual pertanyin. I tots els atributs de la taula de Fets, llevat de *dataCreacio*, són claus alienes de les taules de Dimensions. Aquest darrer cas no es considerarà com una taula d'un sol atribut, atès que el propòsit que s'està perseguint es minimitzar les consultes entre les taules per tal de millorar l'eficiència en el disseny. Als efectes d'accelerar al màxim l'execució de les consultes totes les claus seran de tipus numèric i adaptades a la mida més adient, d'acord amb l'anàlisi de viabilitat que s'ha fet en una secció anterior. Del que es tracta és d'obtenir unes consultes el més eficient possible.

Amb totes aquestes premisses i consideracions, el model lògic i l'esquema E / R que en resulten són els següents:

- **Piulada** (dataCreacio, tweet_id, cache_id, usuari_id, nombreRecursosCompartits, totalMinutsTranscorreguts)
- **Emmagatzematge** (tweet_id, cache_id, raw_tweet)
- **Usuari** (usuari_id, àlies, nom, imatgePerfilURL, lloc, url, descripció, recompteSeguidors, recompteAmics, recompteEstats, zona_Horària)
- **Etiqueta** (tweet_id, nomEtiqueta)
- **Localitzador** (tweet_id, url)
- **Conversa** (tweet_id, fontUsuari, usuari_Referenciat)


Usuari (usuari_id, àlies, nom, imatgePerfilURL, lloc, url, descripció, recompteSeguidors, recompteAmics, recompteEstats, zona_Horària)
Emmagatzematge (tweet_id, cache_id, raw_tweet)
Etiqueta (tweet_id, nomEtiqueta)
Piulada (dataCreacio, tweet_id, cache_id, usuari_id, nombreRecursosCompartits, totalMinutsTranscorreguts)
Localitzador (tweet_id, url)
Conversa (tweet_id, fontUsuari, usuari_Referenciat)

6.5. Model multi-dimensional

Arribats a un punt en què ja tenim les relacions que componen la nostra base de dades i ja sabem quin tipus de consultes volem executar, el que ens falta per fer és el disseny físic tenint molt present que desitgem un bon temps de resposta a les consultes. El disseny físic d'una base de dades és una descripció de la implementació de la base de dades en memòria secundària, tot descrivint les estructures d'emmagatzematge i els mètodes d'accés a les dades. El disseny físic tradueix el disseny lògic en una solució que es pot implementar i econòmica. En aquest sentit, suposa la transcripció del disseny lògic en taules, tipus de dades, relacions entre taules, restriccions i altres components que seran interpretats i manipulats per un sistema gestor de bases de dades (SGBD). El component és la unitat de construcció elemental del disseny físic, essent les seves característiques:

- Està definit segons com interactua amb els altres
- Encapsula les seves funcions i les seves dades
- És reutilitzable a través de les aplicacions
- Es pot veure com una "caixa negra"
- Pot contenir altres components

Les consideracions que he pres per realitzar el disseny físic han estat avaluar el nombre de registres que hi ha en el model per la definició dels tipus de dades enteres i prendre en consideració el valor més gran entre els que hi ha a la taula per la definició de la mida dels atributs alfanumèrics, si no és el cas que ja ve indicat per la base dades associada al Congrés.

Amb aquesta informació, el model físic proposat presenta la següent estructura:

Taula d'Emmagatzematge

Atributs	Tipus	Mida	És Clau Primària?	És Clau Aliena?	NULL?
tweet_id	BIGINT	20	Sí	No	No
cache_id	INTEGER	10	Sí	No	No
raw_tweet	VARCHAR	30	No	No	No

Taula d'Usuari

Atributs	Tipus	Mida	És Clau Primària?	És Clau Aliena?	NULL?
usuari_id	BIGINT	20	Sí	No	No
àlies	VARCHAR	20	No	No	No
nom	VARCHAR	20	No	No	Sí
imatgePerfilURL	VARCHAR	200	No	No	Sí
lloc	VARCHAR	30	No	No	Sí
url	VARCHAR	200	No	No	Sí
descripció	VARCHAR	200	No	No	Sí
recompteSeguidors	INTEGER	10	No	No	Sí
recompteAmics	INTEGER	10	No	No	Sí
recompteEstats	INTEGER	10	No	No	Sí
zona_Horària	VARCHAR	40	No	No	Sí

Taula d'Etiqueta

Atributs	Tipus	Mida	És Clau Primària?	És Clau Aliena?	NULL?
tweet_id	BIGINT	20	Sí	No	No
nomEtiqueta	VARCHAR	100	No	No	No

Taula de Localitzador

Atributs	Tipus	Mida	És Clau Primària?	És Clau Aliena?	NULL?
tweet_id	BIGINT	20	Sí	No	No
url	VARCHAR	140	No	No	No

Taula de Conversa

Atributs	Tipus	Mida	És Clau Primària?	És Clau Aliena?	NULL?
tweet_id	BIGINT	20	Sí	No	No
fontUsuari	BIGINT	20	Sí	No	No
usuari_Referenciat	BIGINT	20	Sí	No	No

Taula de Fets: Piulada

Atributs	Tipus	Mida	És Clau Primària?	És Clau Aliena?	NULL?
dataCreacio	DATETIME	14	Sí	No	No
tweet_id	BIGINT	20	Sí	Sí	No
cache_id	INTEGER	10	Sí	Sí	No
usuari_id	BIGINT	20	Sí	Sí	No
nombreRecursosCompartits	INTEGER	10	No	No	Sí
totalMinutsTranscorreguts	DECIMAL	10,10	No	No	No

A diferència del disseny original i per facilitar la resposta a consultes com ara recursos compartits o activitats dels assistents (enviament/reenviament o seguiment de tweets), tot i no ser Mesures de la taula de Fets, s'han mantingut els camps *is_retweet, is_reply* i *retweet_reply_tweet_id* a la taula Piulada en fer la seva implementció. D'altra banda, el camp *data_creacio* s'ha mantingut del disseny original i ha estat útil per estudiar l'evolució del nombre de tweets a la llarg del temps.

7. Desenvolupament

En aquesta secció, centrada en el desenvolupament del projecte, abordarem aspectes com ara els components de programari i maquinari (SW / HW) requerits i els resultats obtinguts com a resultat d'implementar el magatzem de dades, incloent-hi la impressió de tots els informes o *reports* amb una breu descripció per cadascun. Vegem cadascun d'aquests aspectes per separat:

7.1. Components SW / HW

Cas del desenvolupador del TFC:

- Requisits de Programari o Software

  - Sistema operatiu Windows 7 Home Premium Service Pack 1.
  - Màquina virtual Amazon proporcionada per la UOC per instal·lar el projecte de magatzem de dades.
  - Sistema Gestor de Base de Dades (SGBD) PostgreSQL, atès que és un sistema de base de dades de més nivell que MySQL presentant-hi tot aquest llistat d'avantatges:

    - La seva arquitectura de disseny escala molt bé en augmentar el nombre de CPU i la quantitat de RAM.
    - Suporta transaccions i, des de la versió 7.0, claus alienes (amb comprovacions d'integritat referencial).
    - Presenta millor suport per a triggers i procediments en el servidor.

    - Suporta un subconjunt de SQL92 més gran que el que suporta MySQL. A més, té algunes característiques orientades a objectes.
    - És el SGBD recomanat a les assignatures *Bases de Dades I* - *Bases de Dades II*.

  - Llenguatge de programació PL/SQL de Oracle.
  - Diagrames amb MS Visio 2007.

- Eines de planificació de projectes Microsoft Office Project 2007 i GanttProject.
- Ús de gràfics amb MS Excel 2010.
- Manuals d'ús online de Microsoft Office Project 2007.

- Requisits de Maquinari o Hardware

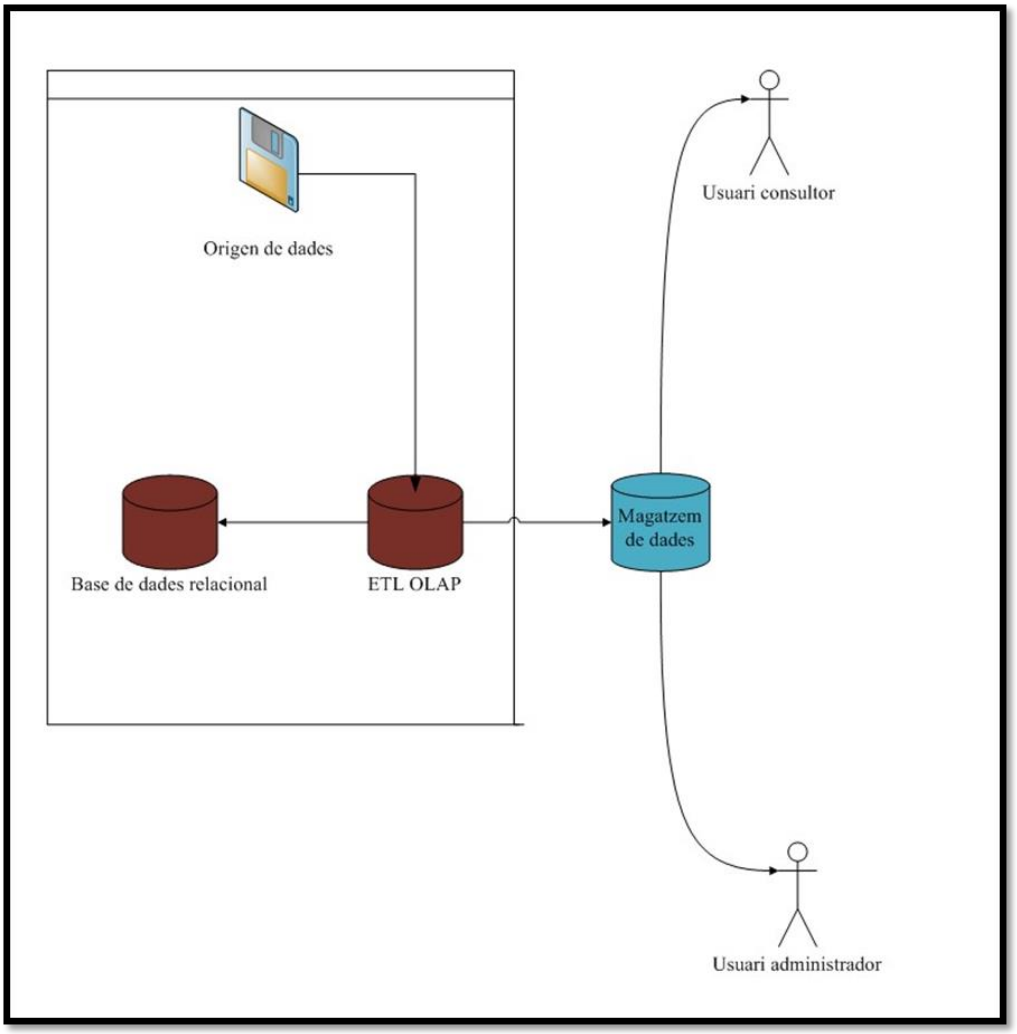
- PC del fabricant Hewlett-Packard Company.
- Model CQ5315Es.
- Processador Pentium(R) Dual-Core CPU E5300 @ 2,60 GHz 2,60 GHz
- Memòria instal·lada (RAM): 4,00 GB.
- Sistema operatiu de 64 bits.
- Disc dur amb sistema d'arxius NTFS de 500 GB.

Cas del client:

En cap cas l'enunciat del projecte no especifica el nivell tecnològic del client, per la qual cosa dono per suposat que el client té al seu abast una infraestructura de maquinari prou solvent com per a instal·lar la solució informàtica i que posseeix les llicències de programari necessàries com per tenir el magatzem de dades en les tecnologies de la solució informàtica que s'està proposant. Si l'adquisició de llicències implica el registre com a usuari a les multinacionals que lliuren el programari per fer les descàrregues adients, dono per suposat que el client ja ha superat aquests esculls.

7.2. Arquitectura del projecte

A partir del que ja s'ha comentat pel que fa a subsistemes de la solució informàtica en la secció de requeriments funcionals, un possible esquema que ens il·lustra l'arquitectura del projecte realitzat amb l'eina MS Visio 2007 és el que es mostra tot seguit:



7.3. Tecnologies a utilitzar

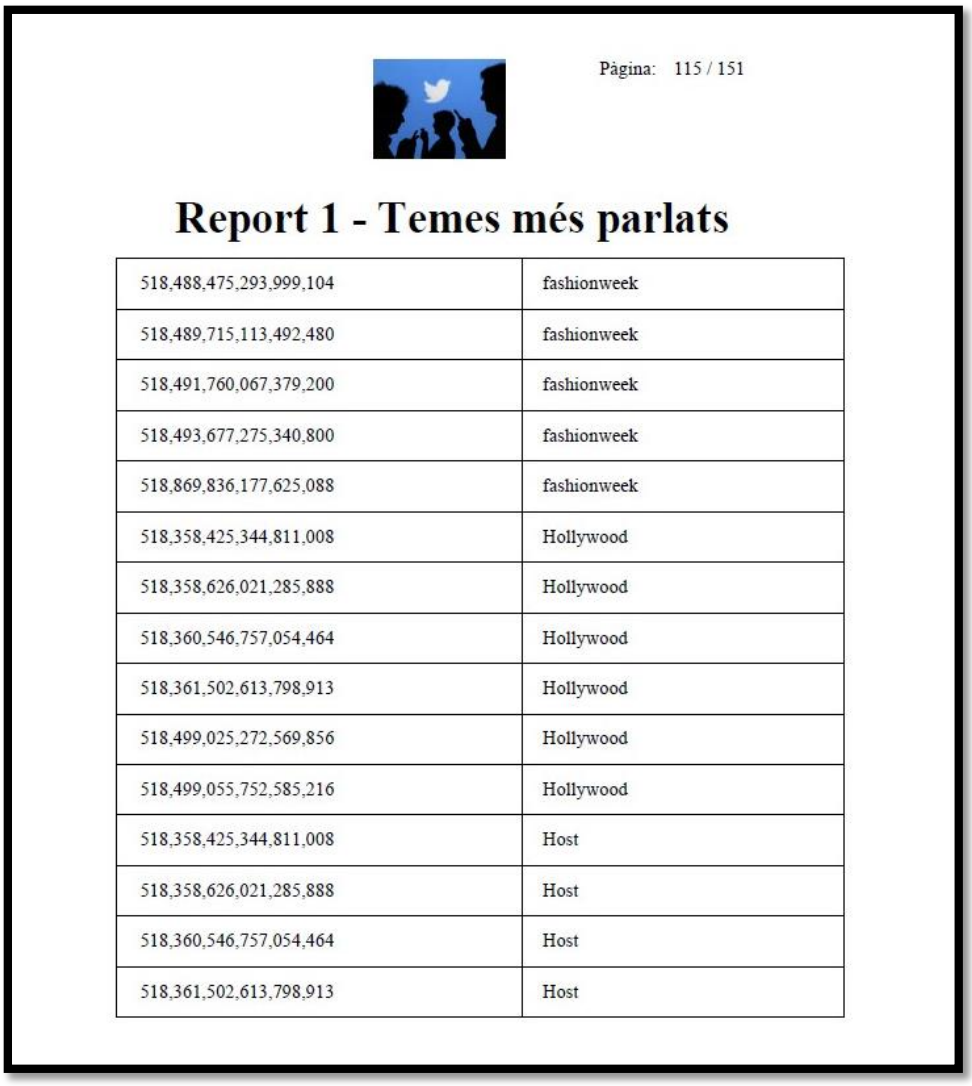
Destaquem en aquesta secció la tecnologia relacional per construir el magatzem de dades, la màquina virtual *Amazon* on s'hi instal·larà el programari necessari per dur a terme el projecte i les eines d'anàlisi i càrrega de dades (programari *Data Warehouse*) que indiqui el consultor de l'assignatura per extreure'n els informes. Cal tenir també present el llenguatge de programació PL/SQL utilitzat en el SGBD *PostgreSQL*. L'usuari i el *password* que faciliten l'accés a la màquina *Amazon* són ***Administrator*** i ***hUqHH)-(sH*** respectivament. Un cop introduïts l'usuari i el *password*, heu de tenir accés a un escriptori a partir d'una pantalla d'inici (*Start*) prement la icona anomenada "Desktop" (Escriptori). Noteu que a l'escriptori hi ha una icona d'accés directe a *PostgreSQL*. Vegem-ho:

7.4. Resultat obtingut (impressió de tots els reports amb una breu descripció)

El primer informe havia de respondre a la pregunta de quins eren els temes més parlats. Accedint a la taula etiqueta del meu magatzem de dades, font contenidora de tots els *hashtags*, he plantejat la següent consulta SQL per tal de trobar resposta a la qüestió plantejada:

```
SELECT e1.tweet_id,
e1.nometiqueta
FROM etiquetaEsmenada e1 LEFT JOIN (SELECT e2.nometiqueta,
                                    COUNT(*) as nombre
                                    FROM etiquetaEsmenada e2
                                    GROUP BY e2.nometiqueta) AS ndv
ON e1.nometiqueta=ndv.nometiqueta
ORDER BY ndv.nombre DESC, e1.nometiqueta ASC, e1.tweet_id ASC;
```

I la resposta es mostra en aquest primer informe, del qual en capturo només una imatge de pàgina aleatòria atès que n'ocupa 151 pàgines:



Pàgina: 115 / 151

Report 1 - Temes més parlats

518,488,475,293,999,104	fashionweek
518,489,715,113,492,480	fashionweek
518,491,760,067,379,200	fashionweek
518,493,677,275,340,800	fashionweek
518,869,836,177,625,088	fashionweek
518,358,425,344,811,008	Hollywood
518,358,626,021,285,888	Hollywood
518,360,546,757,054,464	Hollywood
518,361,502,613,798,913	Hollywood
518,499,025,272,569,856	Hollywood
518,499,055,752,585,216	Hollywood
518,358,425,344,811,008	Host
518,358,626,021,285,888	Host
518,360,546,757,054,464	Host
518,361,502,613,798,913	Host

En resposta a la pregunta plantejada que interessava als investigadors de la UOC, els temes més parlats es resolen mitjançant la següent consulta SQL i el posterior informe que en genera amb el Pentaho Report Designer. Vegem-ho:

```
SELECT e2.nometiqueta,
COUNT(*) as nombre
        FROM etiquetaEsmenada e2
        GROUP BY e2.nometiqueta
        order by nombre desc;
```

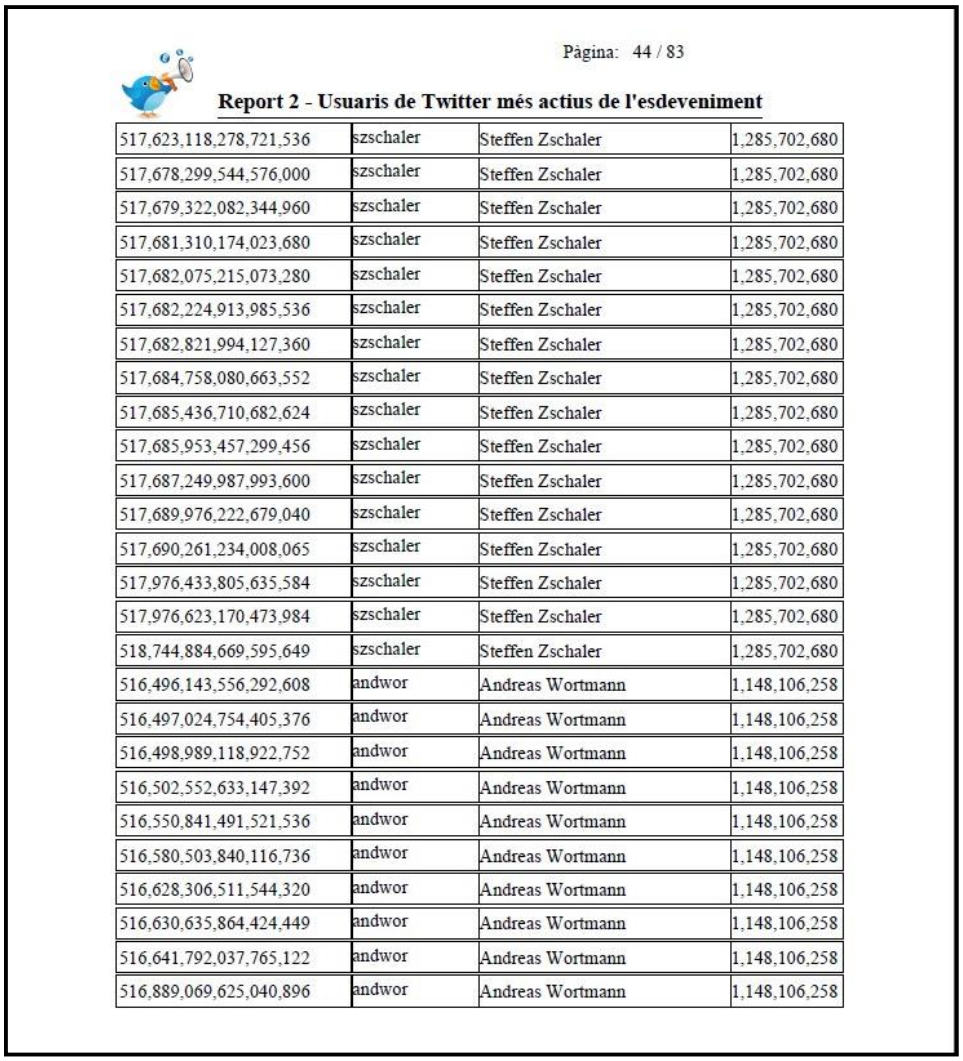
Temes mes parlats (nombre de vegades que hi ha tweets amb aquestes etiquetes)

Tema o hashtag	Tweets etiquetats amb aquest hashtag
models14	1,155
cloudmde	100
cmseba14	74
oss4mde	57
model	26
LOL	22
me14	19
gemoc	19
MDE	18
moda	17
mapreduce	16
xm14	15
fenyset	15
toomanyoptions	14
CloudMDE	12
shooting	12
fashion	12
EduSymp14	11
oldiebutgoldie	10
cloud	10

Pel que fa al segon informe, se n'ocupa de determinar els usuaris de Twitter més actius de l'esdeveniment que ha estat objecte d'estudi. Accedint a les taules piulada i usuari del meu magatzem, he plantejat la següent consulta en llenguatge SQL per tal de trobar una solució al problema plantejat:

```
SELECT p1.tweet_id,
       alies,
       nom,
       p1.usuari_id
  FROM piulada p1 LEFT JOIN (SELECT p2.usuari_id,
                                    COUNT(*) as nombre_de_vegades
                             FROM piulada p2
                             GROUP BY p2.usuari_id) AS ndv
  ON p1.usuari_id=ndv.usuari_id
  INNER JOIN usuari u ON p1.usuari_id=u.usuari_id
  ORDER BY ndv.nombre_de_vegades DESC, p1.tweet_id ASC;
```

La resposta obtinguda ha estat un informe de 83 pàgines, del qual només en mostro una captura de pantalla triada a l'atzar on hi figura l'identificador de tweet sobre el que ha actuat l'usuari, el nom d'usuari i el seu àlies així com el seu identificador d'usuari. He interpretat que un usuari és més o menys actiu en funció del volum d'autories del qual en sigui responsable en els diferents tweets de l'esdeveniment. Vegem-ne la captura:

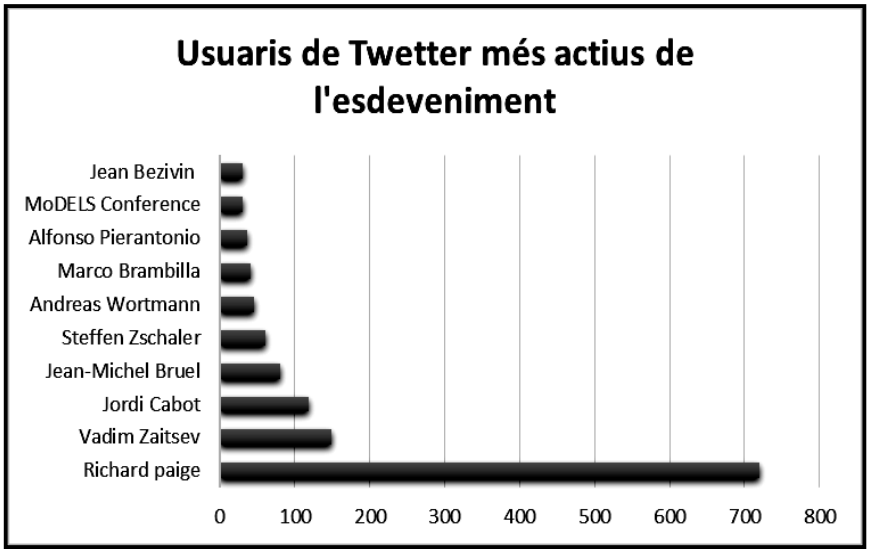


Pàgina: 44 / 83

Report 2 - Usuaris de Twitter més actius de l'esdeveniment

517,623,118,278,721,536	szschaler	Steffen Zschaler	1,285,702,680
517,678,299,544,576,000	szschaler	Steffen Zschaler	1,285,702,680
517,679,322,082,344,960	szschaler	Steffen Zschaler	1,285,702,680
517,681,310,174,023,680	szschaler	Steffen Zschaler	1,285,702,680
517,682,075,215,073,280	szschaler	Steffen Zschaler	1,285,702,680
517,682,224,913,985,536	szschaler	Steffen Zschaler	1,285,702,680
517,682,821,994,127,360	szschaler	Steffen Zschaler	1,285,702,680
517,684,758,080,663,552	szschaler	Steffen Zschaler	1,285,702,680
517,685,436,710,682,624	szschaler	Steffen Zschaler	1,285,702,680
517,685,953,457,299,456	szschaler	Steffen Zschaler	1,285,702,680
517,687,249,987,993,600	szschaler	Steffen Zschaler	1,285,702,680
517,689,976,222,679,040	szschaler	Steffen Zschaler	1,285,702,680
517,690,261,234,008,065	szschaler	Steffen Zschaler	1,285,702,680
517,976,433,805,635,584	szschaler	Steffen Zschaler	1,285,702,680
517,976,623,170,473,984	szschaler	Steffen Zschaler	1,285,702,680
518,744,884,669,595,649	szschaler	Steffen Zschaler	1,285,702,680
516,496,143,556,292,608	andwor	Andreas Wortmann	1,148,106,258
516,497,024,754,405,376	andwor	Andreas Wortmann	1,148,106,258
516,498,989,118,922,752	andwor	Andreas Wortmann	1,148,106,258
516,502,552,633,147,392	andwor	Andreas Wortmann	1,148,106,258
516,550,841,491,521,536	andwor	Andreas Wortmann	1,148,106,258
516,580,503,840,116,736	andwor	Andreas Wortmann	1,148,106,258
516,628,306,511,544,320	andwor	Andreas Wortmann	1,148,106,258
516,630,635,864,424,449	andwor	Andreas Wortmann	1,148,106,258
516,641,792,037,765,122	andwor	Andreas Wortmann	1,148,106,258
516,889,069,625,040,896	andwor	Andreas Wortmann	1,148,106,258

A partir de l'informe que s'ha generat, resolem que l'usuari més actiu de l'esdeveniment ha estat Richard Paige (àlies *richpaige*) amb identificador d'usuari 20773510 i 720 aparicions. El segon lloc del *rànking* és per a Vadim Zaitsev (àlies *grammarware*) amb identificador d'usuari 29290365 i 149 aparicions. La "medalla de bronze" és per a en Jordi Cabot (àlies *softmodeling*) amb identificador d'usuari 30841465 i 119 aparicions. El "diploma olímpic" o quarta posició és per a Jean-Michel Bruel (àlies *jmbruel*) amb identificador 30302607 i 80 tweets signats. La cinquena posició és per a Steffen Zschaler (àlies *szschaler*) amb l'identificador 1285702680 i 61 aparicions. La sisena, per a Andreas Wortmann (àlies *andwor*) amb l'identificador 1148106258 i 46 aparicions. La setena, per a Marco Brambilla (àlies *MarcoBrambi*) amb l'identificador 110420616 i 41 aparicions. La vuitena, per a Alfonso Pierantonio (àlies *APierantonio*) amb l'identificador 72976647 i 36 tweets assignats. La novena, per a l'entitat MoDELS Conference (àlies *modelsconf*) amb l'identificador 861145092 i 31 tweets assignats. I la desena posició del *rànking* és per a Jean Bezivin (àlies *JBezivin*) amb l'identificador 37587198 i 30 assignacions. Tot aquest nou coneixement que ens ha aportat l'informe de 83 pàgines queda prou ben resumit en aquest gràfic de barres:



El següent informe ens ha de permetre detectar els usuaris més seguits en temes (etiquetes) concrets. Se'ns demana que, un cop escollit un tema, de tots els que l'han "usat" (ja siguin tipus *tweet*, *retwet* o *reply*) es mostrin els usuaris que tenen més seguidors. A tal efecte s'ha plantejat la següent consulta en llenguatge SQL:

```
SELECT e1.nometiqueta,
       p.tweet_id,
       p.usuari_id,
       u.nom,
       u.recompte_seguidors
  FROM etiqueta e1 LEFT JOIN (SELECT e2.nometiqueta,
                                     COUNT(*) as nombre_de_vegades
                                FROM etiqueta e2
                                GROUP BY e2.nometiqueta) AS ndv
   ON e1.nometiqueta=ndv.nometiqueta
   INNER JOIN piulada p ON e1.tweet_id=p.tweet_id
   INNER JOIN usuari u ON p.usuari_id=u.usuari_id
   ORDER BY ndv.nombre_de_vegades DESC, u.recompte_seguidors DESC;
```

La resposta obtinguda és la mostrada en el següent informe, del qual n'afegeixo només una captura prou representativa d'imatge atès que n'ocupa 70 pàgines:

Pàgina: 54 / 70

Usuaris més seguits en temes concrets

516,911,018,283,646,977	xm14	37,587,198	Jean Bezivin	3,260
516,615,726,828,761,088	xm14	110,420,616	Marco Brambilla	2,101
516,526,847,308,595,200	xm14	20,773,510	Richard Paige	802
516,528,791,074,918,400	xm14	20,773,510	Richard Paige	802
516,522,152,942,895,104	xm14	20,773,510	Richard Paige	802
516,520,131,233,845,249	xm14	20,773,510	Richard Paige	802
516,523,439,214,313,472	xm14	20,773,510	Richard Paige	802
516,527,307,251,781,632	xm14	40,025,280	SÃ©bastien Mosser	355
516,515,225,320,062,976	xm14	72,976,647	Alfonso Pierantonio	345
516,487,755,510,325,249	xm14	72,976,647	Alfonso Pierantonio	345
516,519,484,077,932,545	xm14	72,976,647	Alfonso Pierantonio	345
516,520,007,082,442,752	xm14	72,976,647	Alfonso Pierantonio	345
516,533,421,393,600,512	xm14	66,634,735	Thanos Zolotas	173
516,916,711,304,753,154	xm14	1,967,099,448	MONDO Project	15
516,916,691,188,875,264	xm14	1,967,099,448	MONDO Project	15
518,859,738,235,559,936	toomanyoptions	30,841,465	Jordi Cabot	5,440
518,861,141,263,798,272	toomanyoptions	7,431,072	Pierre Lindenbaum	2,860
518,986,349,001,453,569	toomanyoptions	43,281,013	R0553R	2,295
518,865,910,699,151,360	toomanyoptions	255,718,856	Andrew Gonzales	1,098
518,860,347,629,789,184	toomanyoptions	470,402,184	John Murray	772
518,865,136,116,043,776	toomanyoptions	31,431,687	Pere Baleta	644
518,888,118,456,709,120	toomanyoptions	25,698,588	AndrÃ© Dietisheim	405

De l'informe anterior i amb les dades recollides fins al moment, se sap que en el tema "models 14" l'usuari més seguit ha estat ReleaseTEAM, amb identificador 15812482 i 11801 seguidors. Li segueixen Arboretum C & S i en Jordi Cabot en segona i tercera posició respectivament. Pel que fa al tema "cloudmde", l'usuari més seguit és en Jordi Cabot amb identificador 30841465 i un total de 5440 seguidors. Li segueixen Vadim Zaytsev i Marco Brambilla en segona i tercera posició i ja molt de lluny Hideki Kishida i Richard Paige. El tema "cmseba14" el lidera en Richard Paige, amb identificador 20773510 i un total de 802 seguidors. Li segueixen Marco LA1/4bbecke i Jean-Michel Bruel en segona i tercera posició. El tema "oss4mde" el lidera en Jordi cabot, amb identificador d'usuari 30841465 i un total de 5440 seguidors. Li segueixen Vadim Zaytsev i Marco Brambilla i en cinquena posició del rànking torna a aparèixer el nom de Richard Paige. El tema "LOL" torna a ser liderat per en Jordi Cabot, així com el tema "gemoc". Per aquest darrer, la segona i tercera posició són pels també coneguts Vadim Zaytsev i Richard Paige respectivament. El següent tema és "me14", que el lidera Vadim Zaytsev amb identificador d'usuari 29290365 i un total de 3395 seguidors. Li segueixen el ja conegut Richard Paige i Alfonso Pierantonio en segona i tercera posició i desmarcats del primer. El tema "xm14" és liderat per Jean Bezivin, amb identificador d'usuari 37587198 i 3260 seguidors. Darrera li van els ja coneguts Marco Brambilla i Richard Paige. Per acabar, el tema "toomanyoptions" torna a ser liderat pel ja citat i vell conegut Jordi Cabot.

A la vista dels resultats i comparant els usuaris citats en aquest informe amb els de l'informe anterior, que referenciava els usuaris més actius de l'esdeveniment, podem concloure que els usuaris amb més seguidors són els que assumeixen l'autoria de més tweets en qualsevol de les seves modalitats (respostes o retweets) i podem inferir, en resposta a una de les inquietuds del grup d'investigadors de la UOC, que són generadors d'opinions sobre un determinat tema. L'informe ha estat útil per a inferir aquest nou coneixement.

El següent informe que es demana ha de donar compte dels recursos compartits en tweets. A tal efecte, he dut a terme la següent consulta sobre la taula piulada del meu magatzem de dades. A grans trets, per cada tweet mostra si és retweet o si replica un altre i mostra també l'identificador de tweet del tweet del qual n'és rèplica o resposta. La consulta filtra totes aquelles piulades que no són ni *retweet* ni *reply* d'un altre tweet, atès que queden excloses de la categoria de recursos compartits en tweets. Vegem la consulta SQL:

```
select tweet_id,
    is_retweet,
    is_reply,
    retweet_reply_tweet_id from piulada
    where is_retweet <>0 or is_reply <>0 or (is_retweet=1 and is_reply=1)
    order by tweet_id ASC, retweet_reply_tweet_id ASC;
```

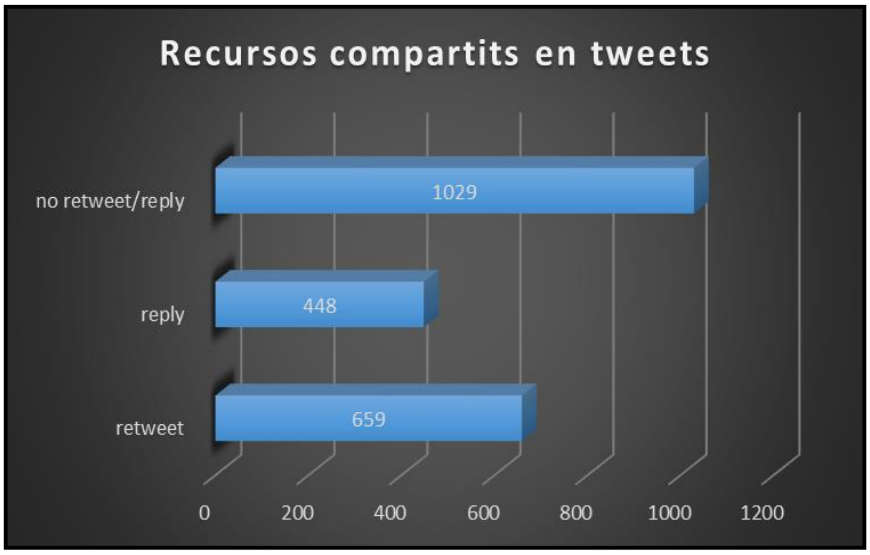
I vet aquí l'informe que s'ha generat amb l'eina *Pentaho Design Reports*, del qual només adjunto una captura de pantalla prou significativa atès que n'ocupa un total de 62 pàgines:

Pàgina: 1 / 62

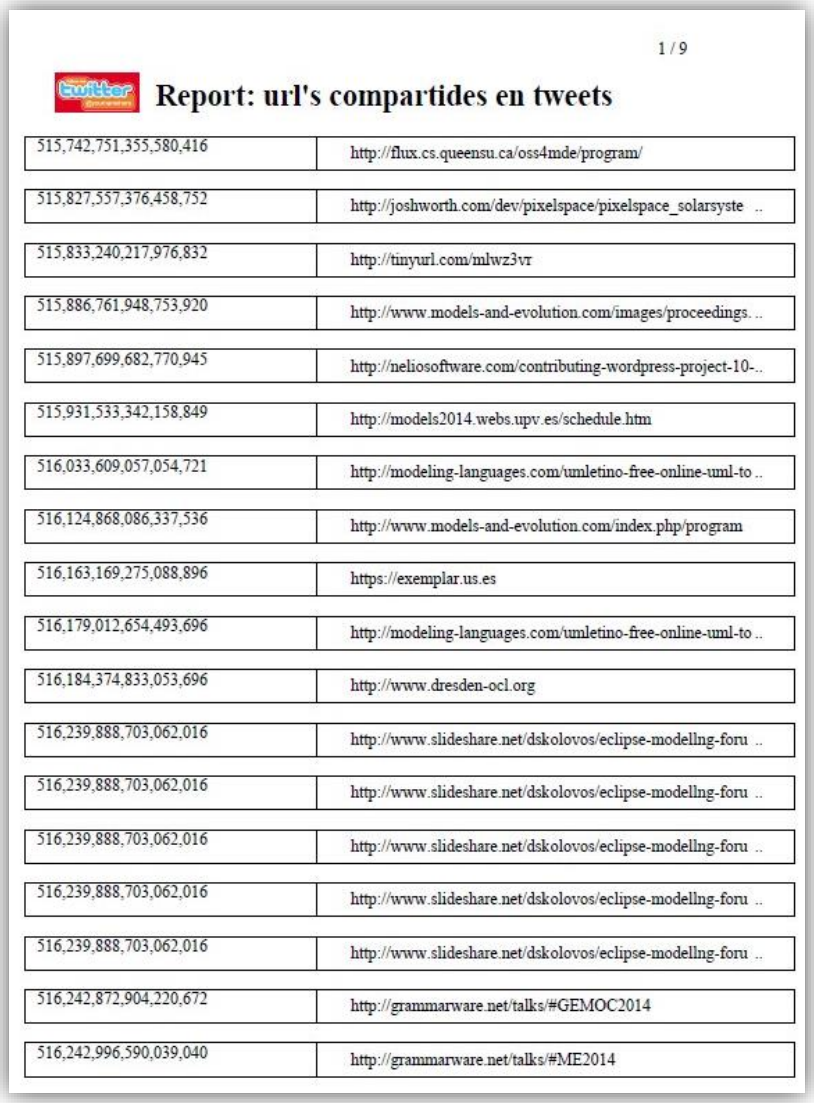
Report: Recursos compartits en tweets

Identificador de tweet	És retweet	És reply	Identificador de retweet o reply
515,251,916,637,761,536	1	0	515,060,543,481,446,400
515,309,015,719,772,160	0	1	515,168,239,874,498,560
515,360,717,378,433,025	0	1	515,309,015,719,772,160
515,380,154,097,745,920	1	0	515,059,023,524,425,728
515,418,818,949,419,008	1	0	515,418,634,039,345,153
515,426,191,688,626,176	0	1	515,419,239,722,004,480
515,426,229,579,964,417	1	0	510,414,185,931,763,712
515,426,405,929,480,192	0	1	515,396,777,747,677,184
515,426,818,338,222,082	0	1	515,426,191,688,626,176
515,428,666,839,281,664	1	0	514,749,966,192,222,208
515,444,287,946,117,120	0	1	513,677,253,432,336,384
515,481,459,889,827,840	1	0	514,701,941,331,726,337
515,491,071,699,468,289	1	0	514,749,966,192,222,208
515,491,213,806,665,728	1	0	515,487,168,286,846,977

De l'informe anterior obtenim que hi ha un total de 659 *retweets*, un total de 448 *replies* i un total de 1029 tweets que no són ni *retweet* ni *reply* d'un volum de 2136 tweets que hi ha al meu magatzem. La situació queda resumida en aquest gràfic de barres:



A efectes de tenir una visió més àmplia de recurs, els següents informes mostren un llistat amb les URLs i les imatges compartides en tweets:



1 / 9

Report: url's compartides en tweets

515,742,751,355,580,416	http://flux.cs.queensu.ca/oss4mde/program/
515,827,557,376,458,752	http://joshworth.com/dev/pixelspace/pixelspace_solarsyste ..
515,833,240,217,976,832	http://tinyurl.com/mlwz3vr
515,886,761,948,753,920	http://www.models-and-evolution.com/images/proceedings. ..
515,897,699,682,770,945	http://neliosoftware.com/contributing-wordpress-project-10-..
515,931,533,342,158,849	http://models2014.webs.upv.es/schedule.htm
516,033,609,057,054,721	http://modeling-languages.com/umletino-free-online-uml-to ..
516,124,868,086,337,536	http://www.models-and-evolution.com/index.php/program
516,163,169,275,088,896	https://exemplar.us.es
516,179,012,654,493,696	http://modeling-languages.com/umletino-free-online-uml-to ..
516,184,374,833,053,696	http://www.dresden-ocl.org
516,239,888,703,062,016	http://www.slideshare.net/dskolovos/eclipse-modellng-foru ..
516,239,888,703,062,016	http://www.slideshare.net/dskolovos/eclipse-modellng-foru ..
516,239,888,703,062,016	http://www.slideshare.net/dskolovos/eclipse-modellng-foru ..
516,239,888,703,062,016	http://www.slideshare.net/dskolovos/eclipse-modellng-foru ..
516,239,888,703,062,016	http://www.slideshare.net/dskolovos/eclipse-modellng-foru ..
516,242,872,904,220,672	http://grammarware.net/talks/#GEMOC2014
516,242,996,590,039,040	http://grammarware.net/talks/#ME2014

1 / 70

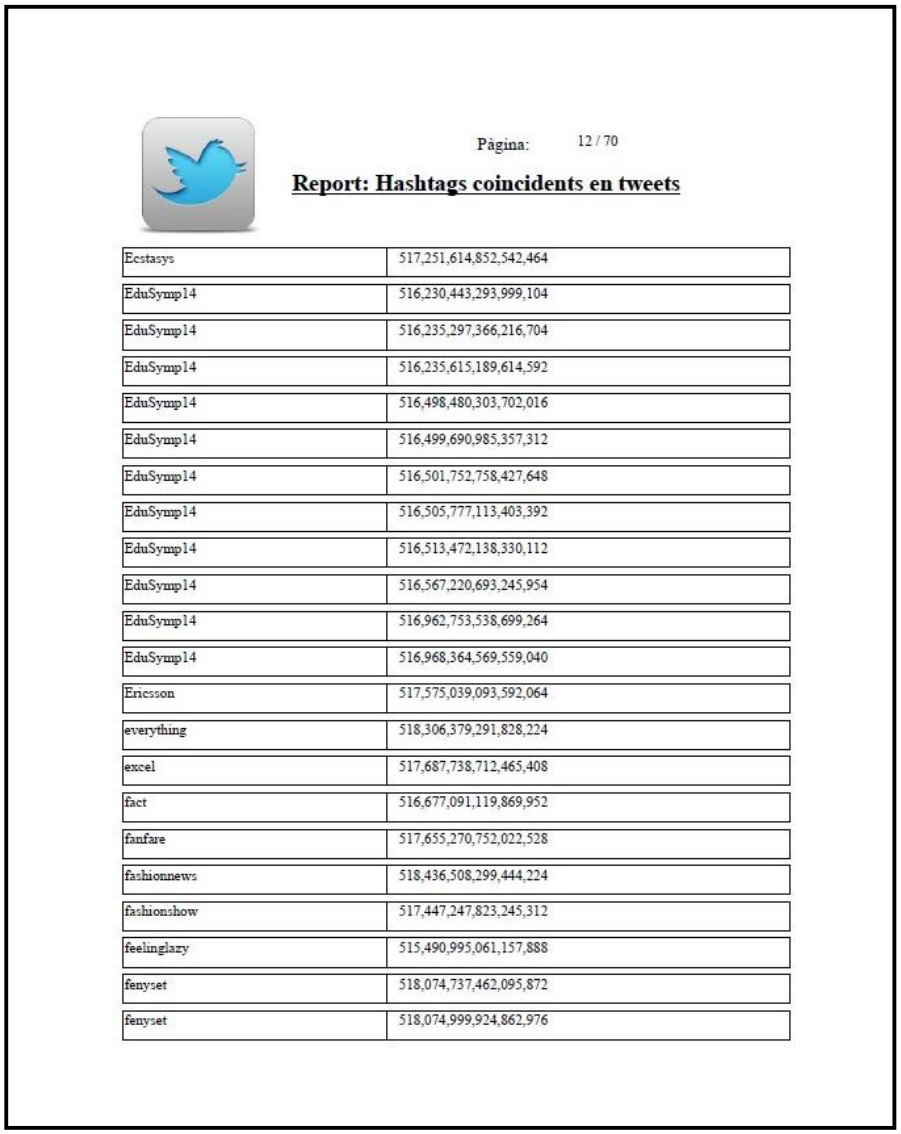
Report: imatges compartides en tweets

515,309,015,719,772,1 ..	Tzo4OiJzdGRDbGFzcyI6MjU6e3M6MTA6ImNyZWF0ZWRfY ..
515,396,777,747,677,1 ..	Tzo4OiJzdGRDbGFzcyI6MjU6e3M6MTA6ImNyZWF0ZWRfY ..
515,418,634,039,345,1 ..	Tzo4OiJzdGRDbGFzcyI6MjU6e3M6MTA6ImNyZWF0ZWRfY ..
515,418,634,039,345,1 ..	Tzo4OiJzdGRDbGFzcyI6MjU6e3M6MTA6ImNyZWF0ZWRfY ..
515,418,634,039,345,1 ..	Tzo4OiJzdGRDbGFzcyI6MjU6e3M6MTA6ImNyZWF0ZWRfY ..
515,418,634,039,345,1 ..	Tzo4OiJzdGRDbGFzcyI6MjU6e3M6MTA6ImNyZWF0ZWRfY ..
515,418,634,039,345,1 ..	Tzo4OiJzdGRDbGFzcyI6MjU6e3M6MTA6ImNyZWF0ZWRfY ..
515,419,239,722,004,4 ..	Tzo4OiJzdGRDbGFzcyI6MjU6e3M6MTA6ImNyZWF0ZWRfY ..
515,426,191,688,626,1 ..	Tzo4OiJzdGRDbGFzcyI6MjU6e3M6MTA6ImNyZWF0ZWRfY ..
515,511,834,464,763,9 ..	Tzo4OiJzdGRDbGFzcyI6MjU6e3M6MTA6ImNyZWF0ZWRfY ..
515,512,370,945,609,7 ..	Tzo4OiJzdGRDbGFzcyI6MjU6e3M6MTA6ImNyZWF0ZWRfY ..
515,512,486,087,241,7 ..	Tzo4OiJzdGRDbGFzcyI6MjU6e3M6MTA6ImNyZWF0ZWRfY ..
515,512,566,538,194,9 ..	Tzo4OiJzdGRDbGFzcyI6MjU6e3M6MTA6ImNyZWF0ZWRfY ..

El següent informe se n'ocupa de donar compte de les etiquetes o *hashtags* coincidents en tweets. Es tracta doncs d'aplegar tots els tweets que es poden associar a una mateixa etiqueta. Per donar resposta a aquesta qüestió, s'ha elaborat una consulta SQL d'estructura i plantejament bastant semblants a la plantejada per resoldre el primer informe. Vegem-la tot seguit:

```
SELECT e1.nometiqueta,
e1.tweet_id
FROM etiqueta e1 LEFT JOIN (SELECT e2.tweet_id,
COUNT(*) as nombre
FROM etiqueta e2
GROUP BY e2.tweet_id) AS ndv
ON e1.tweet_id=ndv.tweet_id
ORDER BY ndv.nombre DESC, e1.nometiqueta ASC,
e1.tweet_id ASC;
```

I el resultat que ens retorna és aquest informe de 70 planes, el qual ens dóna un llistat dels identificadors de tweet associats a cadascuna de les etiquetes mostrades en ordre creixent alfabètic. Se n'adjunta una captura de pantalla per il·lustrar l'informe:



Pàgina: 12 / 70

Report: Hashtags coincidents en tweets

Ecstasys	517,251,614,852,542,464
EduSymp14	516,230,443,293,999,104
EduSymp14	516,235,297,366,216,704
EduSymp14	516,235,615,189,614,592
EduSymp14	516,498,480,303,702,016
EduSymp14	516,499,690,985,357,312
EduSymp14	516,501,752,758,427,648
EduSymp14	516,505,777,113,403,392
EduSymp14	516,513,472,138,330,112
EduSymp14	516,567,220,693,245,954
EduSymp14	516,962,753,538,699,264
EduSymp14	516,968,364,569,559,040
Ericsson	517,575,039,093,592,064
everything	518,306,379,291,828,224
excel	517,687,738,712,465,408
fact	516,677,091,119,869,952
fanfare	517,655,270,752,022,528
fashionnews	518,436,508,299,444,224
fashionshow	517,447,247,823,245,312
feelinglazy	515,490,995,061,157,888
fenyset	518,074,737,462,095,872
fenyset	518,074,999,924,862,976

De l'anterior captura de pantalla podem inferir, per exemple, que l'etiqueta *fenyset* apila com a mínim a dos tweets donats pels identificadors 518074737462095872 i 518074999924862976 respectivament.

El següent informe demanat ha de donar compte de les activitats dels assistents, tant si envien / reenvien tweets com si segueixen a d'altres. A tal efecte he creat una consulta SQL que utilitza les taules piulada i usuari del meu magatzem de dades per extreure'n la informació requerida. Vegem-la:

```
select u.usuari_id,
nom,
tweet_id,
is_retweet,
is_reply,
retweet_reply_tweet_id
from usuari u
INNER JOIN piulada p
ON u.usuari_id = p.usuari_id
order by usuari_id ASC, nom ASC;
```

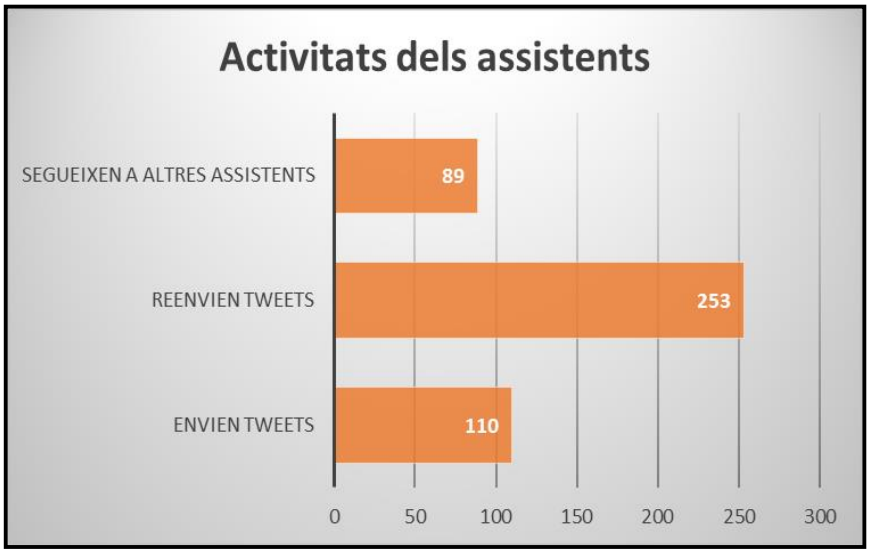
La consulta retorna l'identificador i el nom de l'usuari de la taula usuari del meu magatzem, així com informació del tweet sobre el qual l'usuari hi té una interacció de la taula piulada i ordena les dades per identificador d'usuari en ordre ascendent i per nom d'usuari en ordre ascendent en cas d'empat. Vegem l'informe que genera l'eina *Pentaho Report Designer* a partir de només una captura prou representativa, atès que l'informe ocupa 134 planes.

Report: activitat dels assistents

Pàgina: 1 / 134

Identificador d'usuari	Nom	Identificador de tweet	Retweet	Reply	id tweet as retweet or reply
610,323	Serge Stinckwich	517,314,840,319,516,672	1	0	517,314,620,240,191,489
610,323	Serge Stinckwich	517,934,278,689,427,456	1	0	517,934,020,290,940,929
763,224	James Robertson	515,531,878,074,691,585	1	0	515,517,617,747,279,872
822,087	Keith Mantell	516,514,343,937,384,448	1	0	516,507,422,681,497,600
6,310,412	Anders Aspns	515,934,871,349,653,504	1	0	515,768,560,757,792,768
6,484,132	Richie Rump	517,289,091,185,254,401	0	1	517,288,343,059,828,737
6,484,132	Richie Rump	517,288,938,277,699,584	0	1	517,288,343,059,828,737
6,484,132	Richie Rump	517,288,218,157,260,801	0	0	0
7,431,072	Pierre Lindenbaum	516,857,195,825,479,680	1	0	516,808,726,305,443,840
7,431,072	Pierre Lindenbaum	518,861,141,263,798,272	1	0	518,859,738,235,559,936
7,601,492	Meredith L Patterson	516,278,254,911,905,792	0	1	516,277,581,621,243,905
7,959,122	Pavan Yara	519,061,933,824,032,770	0	1	518,859,738,235,559,936
8,252,602	webhat/redhat	518,698,323,721,875,456	0	1	518,630,462,944,980,992

De l'anterior informe s'acaba obtenint el següent nou coneixement: del total de 446 usuaris registrats diferents que conté la taula usuari, 110 s'han dedicat exclusivament a enviar tweets i prou, 253 han reenviat tweets i 89 han fet un seguiment a d'altres assistents. La situació es pot resumir en el següent gràfic de barres, que il·lustra l'activitat dels assistents:



Per acabar aquesta secció, el darrer informe ens ha de mostrar l'evolució del nombre de tweets al llarg del temps. A tal efecte s'ha elaborat la següent consulta simple SQL, que ens retorna l'identificador de tweet i la data de creació de tots els tweets inserits a la taula piulada del meu magatzem, ordenant les dades pel camp *data_creacio* en ordre ascendent:

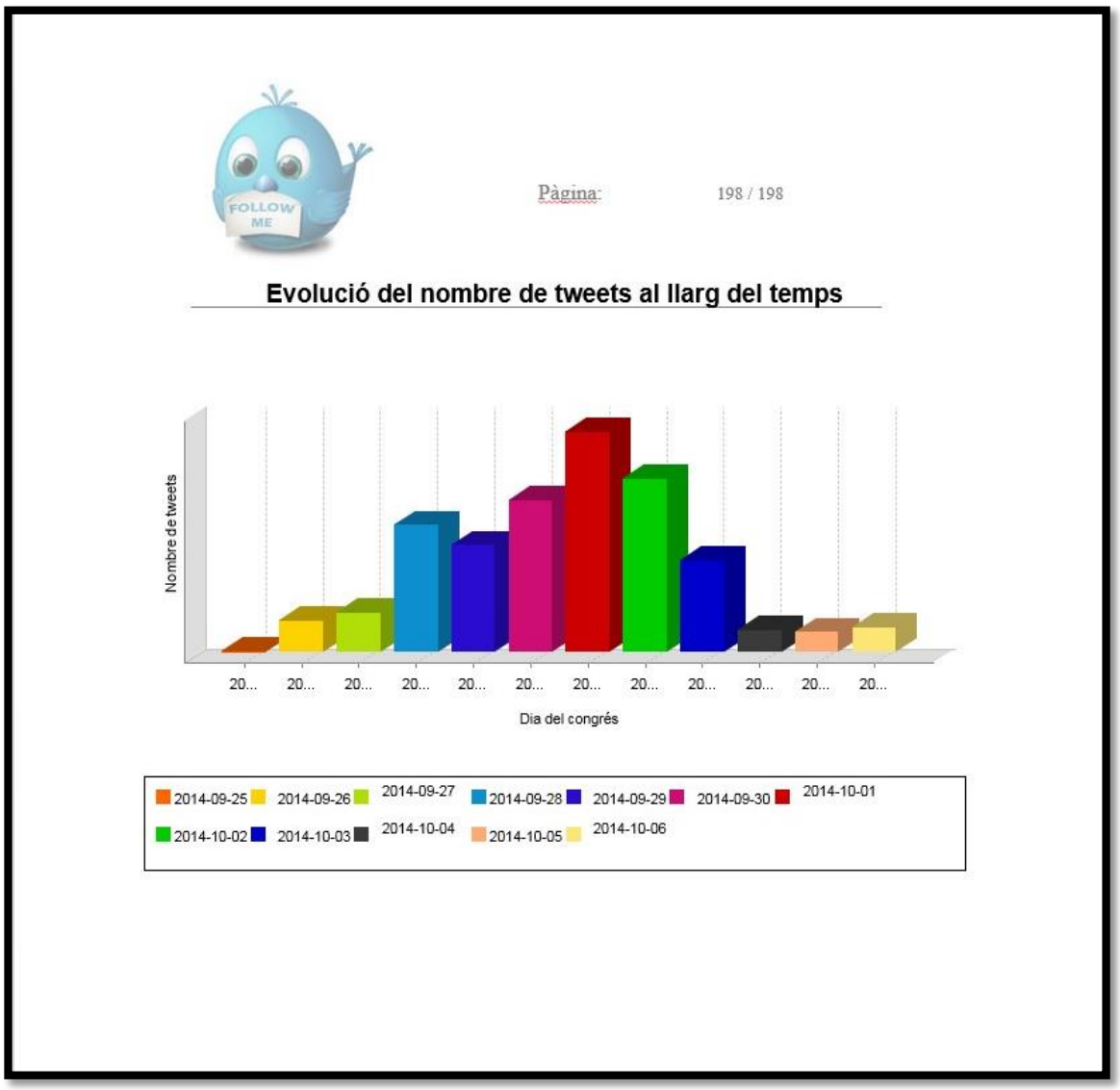
```
select tweet_id, data_creacio
from piulada
order by data_creacio ASC;
```

I l'informe que se'n deriva de la consulta SQL amb l'eina Pentaho ocupa 107 planes i té l'aspecte que es mostra tot seguit, adjuntant-ne una captura de pantalla corresponent a la primera pàgina de l'informe, que enregistra els tweet del dijous 25 de setembre i el divendres 26 de setembre de 2014 respectivament, i de l'última, que inclou un gràfic resum de Pentaho:

Pàgina: 1 / 107

Report: Evolució del nombre de tweets al llarg del temps

Identificador de tweet	Data de creació
515,251,916,637,761,536	Thu Sep 25 00:00:00 UTC 2014
515,309,015,719,772,160	Thu Sep 25 00:00:00 UTC 2014
515,389,443,651,006,465	Fri Sep 26 00:00:00 UTC 2014
515,590,291,794,173,952	Fri Sep 26 00:00:00 UTC 2014
515,591,761,796,681,728	Fri Sep 26 00:00:00 UTC 2014
515,605,341,103,718,400	Fri Sep 26 00:00:00 UTC 2014
515,606,040,826,900,480	Fri Sep 26 00:00:00 UTC 2014
515,607,129,177,456,640	Fri Sep 26 00:00:00 UTC 2014
515,610,521,203,384,320	Fri Sep 26 00:00:00 UTC 2014
515,612,275,928,211,456	Fri Sep 26 00:00:00 UTC 2014
515,612,570,087,333,888	Fri Sep 26 00:00:00 UTC 2014
515,641,389,837,586,432	Fri Sep 26 00:00:00 UTC 2014
515,668,713,500,979,201	Fri Sep 26 00:00:00 UTC 2014
515,670,853,552,717,824	Fri Sep 26 00:00:00 UTC 2014
515,380,154,097,745,920	Fri Sep 26 00:00:00 UTC 2014
515,396,777,747,677,184	Fri Sep 26 00:00:00 UTC 2014



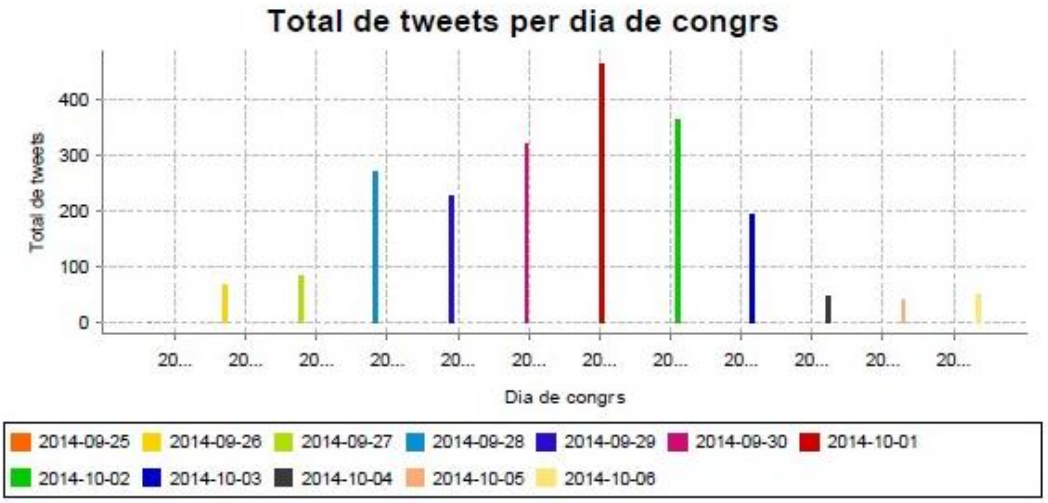
Pel que fa a l'evolució pròpiament dita del nombre de tweets al llarg del temps sense entrar en detalls de quins són els tweets, que és el tema que ara ens interessa, el gràfic anterior ja ens mostra prou bé quina ha estat la tendència. Si volem esbrinar en xifres quants n'hi ha hagut cada dia, resulta útil la consulta:

```
SELECT data_creacio, COUNT(*) AS total FROM piulada GROUP BY data_creacio ORDER BY data_creacio;
```

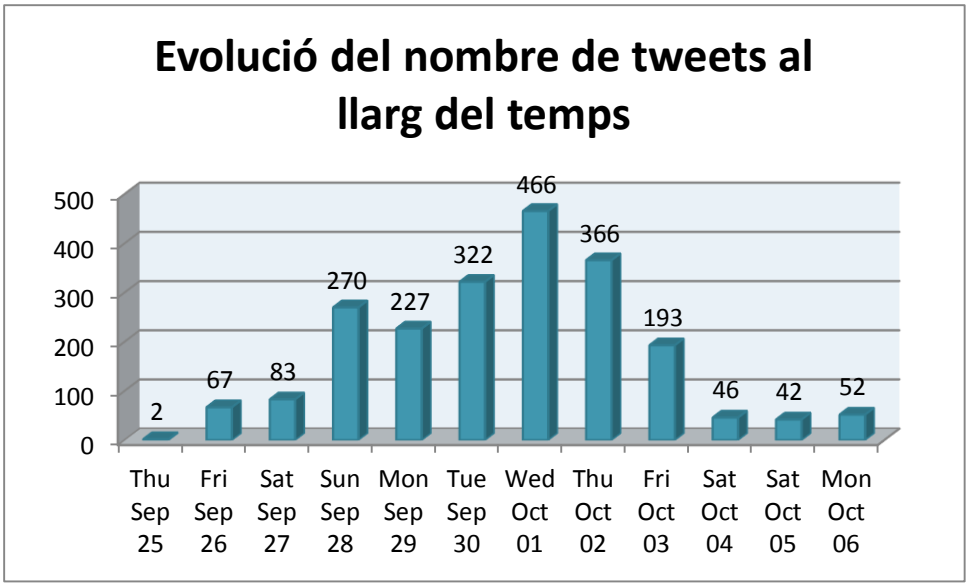
Aquesta consulta permet el següent informe complementari:

Total de tweets per dia de congrés

Data de creació	Total de tweets
Thu Sep 25 00:00:00 UTC 2014	2
Fri Sep 26 00:00:00 UTC 2014	67
Sat Sep 27 00:00:00 UTC 2014	83
Sun Sep 28 00:00:00 UTC 2014	270
Mon Sep 29 00:00:00 UTC 2014	227
Tue Sep 30 00:00:00 UTC 2014	322
Wed Oct 01 00:00:00 UTC 2014	466
Thu Oct 02 00:00:00 UTC 2014	366
Fri Oct 03 00:00:00 UTC 2014	193
Sat Oct 04 00:00:00 UTC 2014	46
Sun Oct 05 00:00:00 UTC 2014	42
Mon Oct 06 00:00:00 UTC 2014	52



I amb aquest informe i el següent gràfic de barres se'ns deixa ben clar que els dies que han enregistrat més tweets són els dies de la franja del dimarts 30 de setembre al dijous 2 d'octubre de 2014, que són els dies centrals en la durada de la conferència. Com a la majoria d'esdeveniments d'una certa durada, són els dies centrals els que representen més activitats. El màxim de tweets enregistrats en una jornada correspon al dimecres 1 d'octubre de 2014, amb un total de 466 tweets enregistrats, i al voltant d'aquest dia la resta de tweets es reparteixen de forma gairebé simètrica. Vegem-ho:



8. Treball futur

Tot i haver construït un magatzem de dades prou solvent i de qualitat, és ara el moment de recollir un recull de propostes que resten pendents de dissenyar i de desenvolupar, de tal manera que poden resultar molt útils en el futur al nostre client per a la gestió del magatzem de dades ja construït. Vegem quines són aquestes propostes d'evolutiu:

• Millores del model E/R. Bàsicament, es podria considerar afegir dades de control a les taules del model de negoci, de manera que en produir-se una inserció o modificació de registre en una taula el sistema registrés quin usuari ha estat responsable i quan ho ha fet. Amb aquesta mesura es podrien endevinar problemes de manipulació "involuntària" de dades i s'aconseguiria un control més acurat sobre qui treballa la informació. La implementació d'aquesta millora es realitzaria de forma senzilla afegint-hi uns pocs atributs a cada taula i condicionant els procediments d'alta i modificació per tal que desin la informació d'usuari i data. Atès que no s'ha demanat necessàriament considerar diversos perfils d'usuari i que no es desitja omplir el codi de funcionalitats restant-ne llegibilitat, s'ha optat per no implementar l'esmentada millora.

• Creació de metadades *Pentaho* per a *reporting ad-hoc*. La suite *Pentaho BI* ofereix l'eina *Pentaho Metadata Editor* que permet definir models de negoci a partir de magatzems de dades i fonts diferents, per tal que posteriorment els usuaris menys especialitzats el facin servir en l'elaboració dels seus propis informes i quadres de comandament. L'objectiu d'aquesta eina és fer un mapatge de l'estructura física de la base de dades a un model lògic de negoci conegut per l'usuari, de tal manera que l'usuari es desvinculi de la implementació física que hi ha darrera del model i sigui capaç d'explotar el magatzem de dades sense coneixement de llenguatges de manipulació de dades com ara les sentències en llenguatge SQL.

• Definició de paràmetres de seguretat i diferents nivells d'accés a les dades. Tal i com està construït el magatzem, tots els usuaris disposen del mateix compte d'accés a la

plataforma i accés a tots els informes que dins d'ella estan publicats a la carpeta *reports*. Ens podríem plantejar crear diferents perfils d'accés amb diferents competències pel que fa a la informació visible a la plataforma. Seria convenient fer la següent distinció pel que fa a visibilitat:

- Que afecti a objectes i funcionalitats. És un nivell de seguretat que afecta als informes i als components de les estructures de carpetes del servidor a les quals u cert perfil d'usuari hi té accés. Podria ser el cas que hi haguessin informes només disponibles per a un cert perfil d'usuari, o bé un seguit de funcionalitats no accessibles per als usuaris, creació d'informes propis i d'altres.

- Que afecti a la visibilitat de la dada. Aquest és un tipus de restricció referit al cas en què, tot i que dos usuaris tinguin accés al mateix informe, la seva execució no retorna la mateixa informació per tots dos usuaris. Una utilitat seria que només segons quins usuaris tinguessin accés a un cert camp de la taula de piulades (camp és *retweet* o *reply*), essent tota la resta de camps visible als usuaris de tots els perfils.

• Creació de quadres de comandament i informes dinàmics. El lliurament inicial contempla a la seva secció d'anàlisi un conjunt d'informes ja definits, que tot i que abasten un gran ventall de tipus d'informes i usos de components, a més d'incloure un munt de filtres per a consultes, són informes estàtics amb capacitat d'anàlisi online limitada. Seria interessant la construcció de quadres de comandament empresarials adreçats a l'àrea de direcció, que permetin a partir d'una simple ullada veure l'estat dels seus indicadors o claus de negoci, així com funcionalitats per a profunditzar i sintetitzar (*Drill down/up*) que permetin explorar-los o analitzar part del document depenent de la secció realitzada.

• Introducció de tasques relacionades amb la millora en la qualitat de les dades.

• Contemplar l'enviament de documents i informes a través de l'aplicació tant de forma manual com automàtica per a múltiples comptes de correu electrònic i de forma dinàmica.

• Control i comunicació de denegació de permisos d'accés.

• Augmentar el conjunt de dades d'origen, incloent-hi tweets sense cap mena de filtre i durant tots els dies que ha durat l'esdeveniment científic, en comptes de limitar-se a una petita mostra.

• Establir en el sistema la capacitat del multi-idioma, de manera que en accedir al sistema l'usuari esculli l'idioma en què apareixeran els diferents literals. A tal efecte, com que per a la creació i execució dels informes calen determinades eines, cal assegurar-s'hi que aquestes eines permeten complir amb aquest requeriment.

• Augmentar les capacitats del procés ETL de manera que tingui en consideració dades corregides.

• Fer servir tècniques de mineria de dades que permetin inferir possibles comportaments. Això seria especialment útil en l'estudi de l'evolució del nombre de tweets amb el temps.

• Afegir noves dades al sistema, per exemple permetent reconstruir converses en la seva totalitat, sabent per cada piulada qui és el seu pare i també les possibles piulades fills.

9. Conclusions

Un cop finalitzat el projecte i havent acabat la redacció del document de memòria, considero oportú fer-ne les oportunes conclusions i valoracions. Aquest és doncs el propòsit d'aquesta secció.

• El primer que puc concloure és que a la llarg de l'elaboració del projecte s'han assolit amb èxit els objectius que s'havien fixat inicialment, de tal manera que s'ha procurat seguir de forma el més acurada possible la planificació que s'havia detallat la Pla de Treball (PAC 1). Ha estat a les acaballes del projecte, durant les fites d'implementació i de presentació final, que he hagut de dedicar una mica més de temps del compte i que m'he desviat una mica respecte el nombre de dies de planificació original. Atès que la qualitat del producte final obtingut no és excel·lent però sí prou solvent i adequada, puc afirmar també que el repartiment de l'esforç dedicat a cadascuna de les fites ha estat també l'adequat.

• Per a dur a terme aquest projecte, que correspon a l'assignatura TFC, he hagut de posar en pràctica les destreses de planificació, anàlisi, disseny i implementació apreses durant la realització dels estudis que ara conclouen. En especial, el projecte ha estat útil per refermar i practicar els coneixements del llenguatge SQL apresos en l'assignatura *Bases de dades I* en un entorn de SGBD tipus *PostgreSQL*. El projecte s'ha centrat en l'àrea de coneixement del magatzem de dades o *Data Warehousing*, per la qual cosa ha estat necessari adquirir nous coneixements en aquesta àrea de treball a partir dels materials de l'assignatura disponibles a l'aula virtual en els diferents formats. L'experiència ha estat prou enriquidora atès que l'àrea temàtica del magatzem de dades és força habitual en el món informàtic empresarial i està a l'ordre del dia. Pel que fa als objectius propis i específics del projecte, el sistema implementa amb èxit un magatzem de dades i permet obtenir el conjunt d'informes necessaris com per a generar el coneixement d'interès per al grup d'investigadors de la UOC.

• Voldria assenyalar també que no puc assegurar a ciència certa si el pla de contingència (anàlisi de riscos) ha estat l'adequat, atès que no s'ha produït cap d'important llevat el fet d'haver-me matriculat en un màster de SAP de Sistemes i Business Intelligence a partir del novembre, fet que ha minvat la meva dedicació al projecte en les seves acaballes. Tot i així, la creació d'un pla de riscos detallat és de gran importància a l'inici de la realització d'un projecte, perquè amb ell es cobreixen tots (o gairebé tots) els

possibles inconvenients que poden aparèixer i, en cas que se'n produeixi cap, indica com s'ha d'actuar de manera que el projecte no experimenti demores.

• Per acabar aquesta secció, també volia comentar-vos que la meva valoració global respecte aquest projecte ha estat del tot satisfactòria perquè se n'aprèn molt, tot i que exigeixi un ritme de treball constant d'inici a fi. Els informes del producte final m'hagués agradat que fossin més vistosos i elaborats i que permetessin un accés i gestió més sofisticats, però en qualsevol cas responen a les qüestions plantejades i acaben generant el nou coneixement requerit. També valoro moltíssim els comentaris del consultor al final de cada fita o PAC en publicar les qualificacions, ja que en general han estat sempre d'ànim constructiu i molt encoratjadors per continuar amb la feina feta.

10. Bibliografia

Materials de suport:

Pradel Miquel, Jordi , Raya Martos, Jose (2014). "Mòdul didàctic 3: Requisits". Pradel Miquel J., Raya Martos, J. *Enginyeria del programari* (2a edició, setembre 2014). Barcelona: *Oberta UOC Publishing, SL*.

Rius Gavídia, Àngels , Serra Vizern, Montse, Abelló Gamazo, Alberto, Samos Jiménez, José, Vidal Portolés, José, Curto Díaz, Josep (2014). "*Data warehouse*. Magatzem de dades i models multidimensionals. Mòdul 1: Introducció a l'emmagatzematge de dades – PID_00189731". Barcelona: Fundació per a la Universitat Oberta de Catalunya.

Sáenz Higueras, Nita , Rut Vidal, Oltra (2014). "Redacció de textos cientificotècnics". Barcelona: FUOC • P08/19018/00445.

Beneito Montagut, Roser (2014). "Presentació de documents i elaboració de presentacions". Barcelona: FUOC • P08/19018/00446.

Abelló Gamazo, Alberto (2014). "*Data Warehouse. Magatzems de dades i models multidimensionals. Mòdul 4. Disseny multidimensional* – PID_00189734". Barcelona: Fundació per a la Universitat Oberta de Catalunya.

Turón Manzanares, Fernando (2013). "*Construcción y Explotación de un almacén de datos para el anàlisis de información sobre alojamientos turísticos. Análisis de Requerimientos. Diseño Conceptual y Técnico – Trabajo Fin de carrera PEC 2*".

Pereiras Magariños, Alexandre (2013). "*Construcción y explotación de un almacén de datos para el anàlisis del sistema de ventas de una distribuidora farmacéutica. Proyecto Final de Carrera*".

Mora Pérez, Fernando (2014). "*Construcción y explotación de un almacén de datos para el anàlisis de información sobre tránsito de vehículos. Memoria del proyecto. Grado de Ingeniería Informática. Trabajo Fin de Grado*".

Moreno Sánchez, José Manuel (2013). "*Construcció i explotació d'un magatzem de dades per a l'anàlisi d'informació sobre allotjaments turístics. Memòria. TFC Magatzem de dades. Enginyeria Tècnica d'Informàtica de Gestió. Universitat Oberta de Catalunya".*

Hervás Bolaños, Pedro (2012). "*Construcción y explotación de un almacén de datos para la empresa inmobiliaria "un techo para todos". Memoria Trabajo Fin de carrera Ingeniería Técnica en Informática de Sistemas".*

Martín Escofet, Carme. "El llenguatge SQL". Barcelona: Mòdul didàctic de l'assignatura "Bases de dades I". Universitat Oberta de Catalunya • P05/05002/00529.

"Pentaho Report Designer User Guide". 2011 Pentaho Corporation.

"Getting Started with Pentaho Report Designer". 2014 Pentaho Corporation.

Webs de suport:

http://www.ieru.org/projects/mavsel/index.html

http://ca.wikipedia.org/wiki/Magatzem_de_dades

http://ca.wikipedia.org/wiki/Planificaci%C3%B3_de_Recursos_Empresarials

http://es.wikipedia.org/wiki/Extract,_transform_and_load

http://es.wikipedia.org/wiki/OLAP

http://es.wikipedia.org/wiki/Diagrama_de_Gantt

Manual de Microsoft Project 2007 - 2XMIL Soluciones
*(*www.2xmil.es/pdf/PROJECT_2007.pdf*)*

http://nairobi.uoc.es/congress/

http://nairobi.uoc.es/phpmyadmin

http://es.wikipedia.org/wiki/Hashtag

http://www.infor.uva.es/~mlaguna/is1/apuntes/2-requisitos.pdf

http://churriwifi.wordpress.com/2010/04/22/15-3-analisis-dimensiones-hechos/

http://es.wikipedia.org/wiki/Modelo_entidad-relaci%C3%B3n

http://www.alegsa.com.ar/Dic/dise%C3%B1o%20fisico%20de%20bases%20de%20datos.php

http://es.slideshare.net/errroman/diseo-lgico-y-diseo-fsico

http://advenis.wordpress.com/2010/04/21/tipos-de-datos-en-mysql/

http://www.desarrolloweb.com/articulos/1054.php

http://www.dataprix.com/blogs/respinosamilla/herramientas-etl-que-son-para-que-valen-productos-mas-conocidos-etl-s-open-sour

https://www.fing.edu.uy/inco/cursos/caldatos/Transparencias/4-TratamientoSI.pdf

http://es.wikipedia.org/wiki/Limpieza_de_datos

http://www.dataprix.com/blogs/respinosamilla/herramientas-etl-que-son-para-que-valen-productos-mas-conocidos-etl-s-open-sour?utm_campaign=procesos-etl&utm_source=hs_automation&utm_medium=email&utm_content=9063552&_hsenc=p2ANqtz-9hXdnZo04IUnFIgu4Ks-qu55TWSuGdE-XvHAwy8ppvYEipV4BAGNgCsBeZg_UqEkXYIRkOBQr7kO6pcM6Yzj6jVg7A5Q&_hsmi=9063552

http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/arquitectura-del-data-warehouse/34-datawarehouse-manager#at_pco=smlwn-1.0&at_si=5447eb75c396ca08&at_ab=per-1&at_pos=0&at_tot=1

http://es.wikipedia.org/wiki/Extract,_transform_and_load

http://es.wikipedia.org/wiki/Trigger_%28base_de_datos%29

https://www.youtube.com/watch?v=mkXWFa41VRw

http://www.xarxanet.org/projectes/noticies/com-hem-de-definir-els-objectius-dun-projecte

http://www.gestio.suport.org/index.php?option=com_content&view=article&id=105:que-es-el-pla-estrategic&catid=34:pmf-activitats&Itemid=44