TFC – Construcció i explotació d'un magatzem de dades

Marçal Isern Torrente
ETIG

Consultor: Cales Llorach Rius
6/01/2015

RESUM

Aquesta memòria recull les tasques realitzades durant l'elaboració del treball de fi de carrera dels estudis d'Enginyeria Tècnica en Informàtica de Sistemes basat en els magatzems de dades.

La memòria està estructurada en diferents parts. Començarem per una introducció on s'especifica la justificació del projecte, els seus objectius i s'adjunta una planificació de totes les tasques realitzades.

En la fase d'anàlisi mostrarem en detall els requeriments tant funcionals com no funcionals que ens determinaran tot el model conceptual del treball. També adjuntarem els diagrames necessaris per una millor comprensió de totes les tasques i processos.

Finalment en la part de desenvolupament parlarem de les eines del stack de Microsoft SQL Server que hem utilitzat.

Sobre el producte final, destacaria per un costat les eines de integració de dades que en permeten unificar informació de diferents fonts així com les possibilitats del model analític de Analysis Services que ens permet consultar a gran velocitat tota la informació del nostra magatzem de dades.

Índex

Índex d'il·lustracions

1. Introducció

Aquesta memòria detalla el treball realitzat de fi de Carrera (TFC) del semestre 2014-2015 basat en la construcció i explotació d'un magatzem de dades.

El document especifica tot el cicle de vida d'un projecte. Partint de les especificacions inicials, s'ha realitzat una planificació acurada de les tasques a realitzar.

Seguidament es detalla les fases de disseny i desenvolupament de la solució proposada i finalment suggerirem propostes de millora.

1.1 Justificació del Projecte

Actualment ens trobem en entorns cada vegada més heterogenis on la necessitat de consolidar la informació i disposar d'eines per explotar-la s'ha convertit en objectius de primera necessitat.

Els magatzems de dades ens permeten unificar tot tipus d'informació provinent de diferents orígens i d'aquesta manera ens permet obtenir una única visió de les dades d'una forma consolidada i homogènia.

Quan ens endinsem dins d'un projecte de magatzem de dades, no parlem únicament de construir un repositori per unificar les dades sinó que moltes vegades necessitem «normalitzar» i netejar les dades que ens venen dels diferents orígens per tal de mostrar una visió unificada, més pròxima als usuaris finals.

Necessitem crear una capa de metadades on el usuari de negoci es senti còmode amb la gestió de la informació. Ells no hi entenen ni de taules ni de columnes i molt menys de relacions. Per tant, és necessari crear una capa de metadades on l'usuari final vegi la informació d'una forma molt més senzilla i amb una nomenclatura i semàntica molt més clara.

1.2 Objectius

El treball de final de carrera té com a objectiu principal l'aplicació de coneixements adquirits al llarg dels estudis per tal de realitzar un producte.

1.2.1 Objectius Generals

L'objectiu principal del TFC es ensenyar a l'alumne el procés de la realització d'un producte per un client final.

Partint de la fase de planificació, anàlisi, implementació i defensa, l'alumne reproduirà el procés habitual de realització de projectes informàtics.

En el món laboral on a banda de la implementació, a la fase de prevenda, és molt habitual adjuntar una planificació així com la defensa del projecte davant del client una vegada finalitzat.

1.2.2 Objectius Específics

Pel que fa als objectius específics, durant el TFC es desenvoluparà un magatzem de dades per tal d'explotar unes dades provinents de twitter sobre un determinat esdeveniment col·lectiu.

Per la construcció d'aquest magatzem de dades necessitarem:

- Definir el model analític: Definició de mètriques, indicadors, dimensions i atributs del nostre model de dades.
- Construir el model relacional: Creació de la Base de dades que emmagatzemarà la informació del nostre magatzem de dades.
- Construcció de processos ETL per tal de manipular i incorporar dades al nostre datawarehouse.
- Creació del model analític OLAP: Creació taules de fets, dimensions i mètriques necessàries
- Definició i creació d'informes finals amb l'eina de reporting.

1.3 Enfocament i mètode seguit

Les tasques principals per desenvolupar aquest projecte han estat les següents:

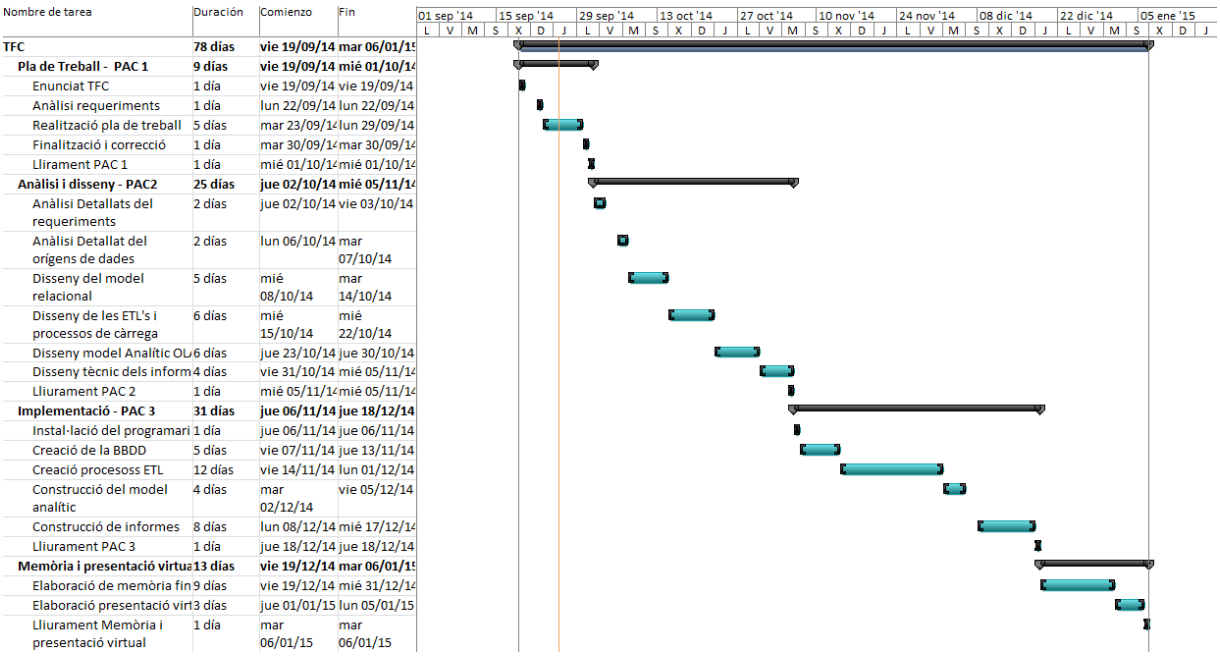
1. **Anàlisi dels requeriments:** A partir dels requeriments del client, es realitza un disseny preliminar del model multidimensional així com dels seus informes.
2. **Instal·lació del programari:** En la màquina virtual proporcionada es realitza la instal·lació de tot els software necessari.
3. **Disseny del model de dades:** A partir del model conceptual, crearem el model relacional final.
4. **Càrrega de dades:** Extraurem la informació del origen de dades i la incorporarem al nostre model relacional.
5. **Disseny multidimensional:** Crearem les diferents dimensions i mètriques necessaries per explotar el cub OLAP
6. **Realització dels informes:** En aquesta fase és crearan els diferent informes plantejats.
7. **Comprovació del producte final:** S'ha de verificar que el producte final assoleixi tots els requeriments fixats per el client.

1.4 Planificació del projecte

A continuació es detalla la planificació del projecte. Al ser un sol estudiant, totes les tasques es realitzaran de forma seqüencial.

Nombre de tarea	Duración	Comienzo	Fin
TFC	**78 días**	**vie 19/09/14**	**mar 06/01/15**
Pla de Treball - PAC 1	**9 días**	**vie 19/09/14**	**mié 01/10/14**
Enunciat TFC	1 día	vie 19/09/14	vie 19/09/14
Anàlisi requeriments	1 día	lun 22/09/14	lun 22/09/14
Realització pla de treball	5 días	mar 23/09/14	lun 29/09/14
Finalització i correcció	1 día	mar 30/09/14	mar 30/09/14
Llirament PAC 1	1 día	mié 01/10/14	mié 01/10/14
Anàlisi i disseny - PAC2	**25 días**	**jue 02/10/14**	**mié 05/11/14**
Anàlisi Detallats del requeriments	2 días	jue 02/10/14	vie 03/10/14
Anàlisi Detallat del orígens de dades	2 días	lun 06/10/14	mar 07/10/14
Disseny del model relacional	5 días	mié 08/10/14	mar 14/10/14
Disseny de les ETL's i processos de càrrega	6 días	mié 15/10/14	mié 22/10/14
Disseny model Analític OLAP	6 días	jue 23/10/14	jue 30/10/14
Disseny tècnic dels informes	4 días	vie 31/10/14	mié 05/11/14
Lliurament PAC 2	1 día	mié 05/11/14	mié 05/11/14
Implementació - PAC 3	**31 días**	**jue 06/11/14**	**jue 18/12/14**
Instal·lació del programari	1 día	jue 06/11/14	jue 06/11/14
Creació de la BBDD	5 días	vie 07/11/14	jue 13/11/14
Creació procesoss ETL	12 días	vie 14/11/14	lun 01/12/14
Construcció del model analític	4 días	mar 02/12/14	vie 05/12/14
Construcció de informes	8 días	lun 08/12/14	mié 17/12/14
Lliurament PAC 3	1 día	jue 18/12/14	jue 18/12/14
Memòria i presentació virtual	**13 días**	**vie 19/12/14**	**mar 06/01/15**
Elaboració de memòria final	9 días	vie 19/12/14	mié 31/12/14
Elaboració presentació virtua	3 días	jue 01/01/15	lun 05/01/15
Lliurament Memòria i presentació virtual	1 día	mar 06/01/15	mar 06/01/15

Il·lustració 1: LListat de Tasques



Nombre de tarea	Duración	Comienzo	Fin
TFC	**78 días**	**vie 19/09/14**	**mar 06/01/15**
Pla de Treball - PAC 1	**9 días**	**vie 19/09/14**	**mié 01/10/14**
Enunciat TFC	1 día	vie 19/09/14	vie 19/09/14
Anàlisi requeriments	1 día	lun 22/09/14	lun 22/09/14
Realització pla de treball	5 días	mar 23/09/14	lun 29/09/14
Finalització i correcció	1 día	mar 30/09/14	mar 30/09/14
Lliurament PAC 1	1 día	mié 01/10/14	mié 01/10/14
Anàlisi i disseny - PAC2	**25 días**	**jue 02/10/14**	**mié 05/11/14**
Anàlisi Detallats del requeriments	2 días	jue 02/10/14	vie 03/10/14
Anàlisi Detallat del orígens de dades	2 días	lun 06/10/14	mar 07/10/14
Disseny del model relacional	5 días	mié 08/10/14	mar 14/10/14
Disseny de les ETL's i processos de càrrega	6 días	mié 15/10/14	mié 22/10/14
Disseny model Analític OLA	6 días	jue 23/10/14	jue 30/10/14
Disseny tècnic dels inform	4 días	vie 31/10/14	mié 05/11/14
Lliurament PAC 2	1 día	mié 05/11/14	mié 05/11/14
Implementació - PAC 3	**31 días**	**jue 06/11/14**	**jue 18/12/14**
Instal·lació del programari	1 día	jue 06/11/14	jue 06/11/14
Creació de la BBDD	5 días	vie 07/11/14	jue 13/11/14
Creació procesoss ETL	12 días	vie 14/11/14	lun 01/12/14
Construcció del model analític	4 días	mar 02/12/14	vie 05/12/14
Construcció de informes	8 días	lun 08/12/14	mié 17/12/14
Lliurament PAC 3	1 día	jue 18/12/14	jue 18/12/14
Memòria i presentació virtua	**13 días**	**vie 19/12/14**	**mar 06/01/15**
Elaboració de memòria fin	9 días	vie 19/12/14	mié 31/12/14
Elaboració presentació virt	3 días	jue 01/01/15	lun 05/01/15
Lliurament Memòria i presentació virtual	1 día	mar 06/01/15	mar 06/01/15

Il·lustració 2: Diagrama de Grantt

1.5 Productes obtinguts

Al final del projecte s'entregarà un model analític que permetrà explotar tots els indicadors des de diferents perspectives .

També es lliuraran una sèrie d'informes on es mostrarà la fortalesa del model presentat detallant l'anàlisi dels diferents indicadors plantejats.

1.6 Breu explicació dels altres capítols

En els següents capítols es profunditza en cada una de les parts realitzades durant l'anàlisi, disseny i construcció de la solució.

En l'apartat d'anàlisi es detalla el llistat de requeriments que donarà lloc al model conceptual presentat així com els diagrames d'activitats necessaris per realitzar els processos d'integració.

En l'apartat de desenvolupament, es comenta la tecnologia utilitzada i s'explica tot el procés de la

construcció del projecte.

En els capítols finals s'adjunta una sèrie de captures de pantalles tant d'informes com d'anàlisis OLAP i finalitzarem amb les conclusions i línies de millora per fer evolucionar el producte.

2 Anàlisis

2.1 Requeriments Funcionals

A continuació es detalla la llista de requeriments funcionals :

ID	**Requeriment**
REQ-01	Analitzar el temes més parlats
REQ-02	Detectar usuaris més actius de l'esdeveniment.
REQ-03	Analitzar els recursos compartits en els tweets.
REQ-04	Descobrir els hashtags coincidents en els tweets.
REQ-05	Monitoritzar les activitats dels assistents.
REQ-06	Evolució del nombre de tweets al llarg del temps

2.2 Requeriments no Funcionals

A continuació detallem el compliment dels requeriments no funcionals :

ID	Requeriment
REQNF-1	L'accés al model analític i als seus informes ha d'estar restringit per els usuaris determinats..

REQNF-2	La solució necessita poder ser explotada tant a través d'informes predefinits com amb eines d'anàlisis.
REQNF-3	Els informes no poden tardar més de 5 segons
REQNF-4	Els usuaris han de saber quan es va realitzar l'última càrrega de les dades.

2.3 Model Conceptual

A l'hora de definir el nostre model conceptual s'ha de tenir molt clar el «què volem analitzar» i el «com ho volem analitzar»

Per un costat definim com a fets fets o mesures totes aquelles coses que volem analitzar (comptar, sumar, etc).

Per l'altre costat tenim les dimensions amb els seus atributs que ens permetran segmentar els fets segons els eixos desitjats. Per exemple:

- Nombre de tweets durant un dia , o una hora.
- Nombre de tweets d'un usuari concret.
- Nombre de tweets amb una temàtica concreta .. etc ...

2.3.1 Dimensions

Com ja hem comentat anteriorment, les dimensions i atributs ens permetran segmentar la informació per diferents eixos.

Les dimensions utilitzades en el nostre model seran:

Dimensió Data:

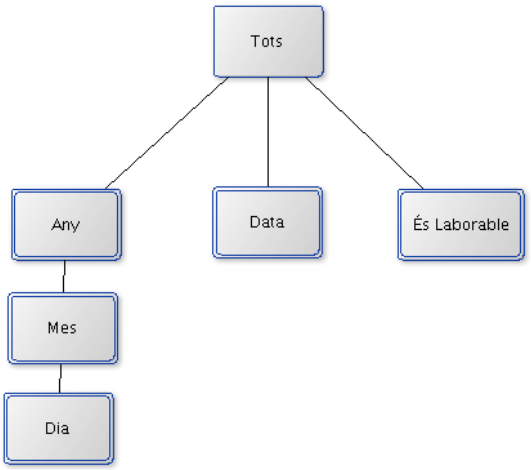
Aquesta dimensió està present en quasi tots els models analítics. Ens servirà per segmentar les dades en el temps.

Atributs:

- Any: Any de la data en format numèric.
- Mes: Mes de la data en format numèric.
- Dia: Dia del mes de la data en format numèric.
- És laborable: indica si la data en qüestió és un dia laborable o és cap de setmana.

Jerarquies:

- Any-Mes-Dia



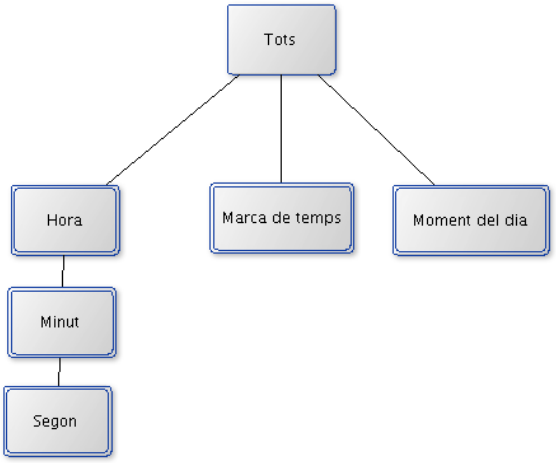
Il·lustració 3: Dimensió Data

Dimensió Temps:

Aquesta dimensió ens ajudarà a segmentar per hora, minut i segon. També crearem algun atribut addicional que ens serà d'utilitat.

Atributs:

- Hora: Indica la hora.
- Minut: Indica el minut.
- Segon: Indica el segón.
- Marca de Temps: Hora Completa HH:MM:SS
- Moment del dia: Indica en quin moment del dia s'ha realitzat el tweet



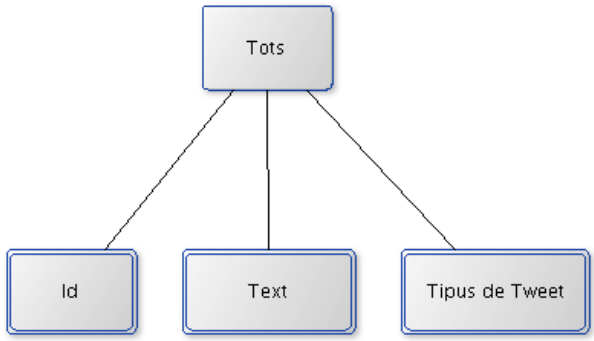
Il·lustració 4: Dimensió Temps

Dimensió Tweet:

La dimensió tweet conté la informació de les piulades capturades.
Aquesta dimensió és una mica peculiar ja que tindrà tants membres com registres tingui la taula de fets de tweets.
La raó per utilitzar aquesta dimensió és la de donar la possibilitat a l'usuari final de poder visualitzar els texts de les piulades o segmentar aquelles que siguin retweets o replays.

Atributs:

- Id: Identificador numèric dels tweets
- Text: Missatge del tweet.
- Tipus de tweet: Indica el tipus de tweet.



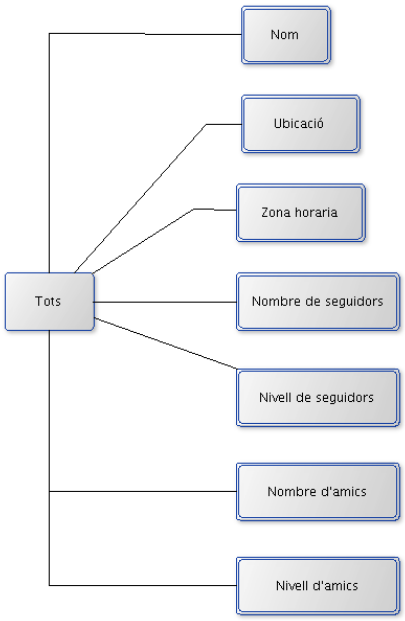
Il·lustració 5: Dimensió tweet

Dimensió Usuari

En la dimensió usuari hi trobarem tota la informació referent als usuaris de tweeter.

Atributs:

- Nom : Nom de l'usuari
- Ubicació: Ubicació (No Normalitzat).
- Zona Horària:
- Nombre de Seguidors
- Nivell de Seguidors
  - Valors possibles: (Baix/Mig/Alt)
- Nombre d'amics
- Nivell d'amics
  - Valors possibles: (Baix/Mig/Alt)



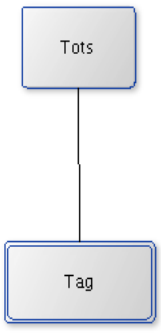
Il·lustració 6: Dimensió Usuari

Dimensió Hashtag

Aquesta dimensió és molt senzilla ja que només té un atribut.

Atributs:

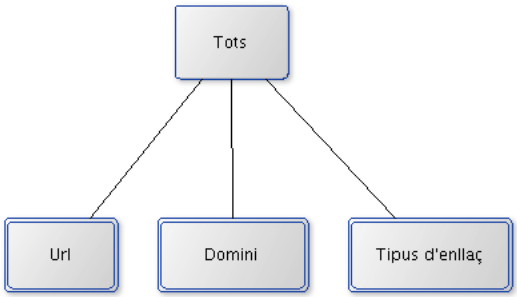
Tag: Paraula o tag utilitzat.



Il·lustració 7: DImensió Hashtag

Dimensió Recurs

Aquesta dimensió descriu els diferents recursos compartits en els tweets

Atributs: • Url: Enllaç del recurs. • Domini: Domini de l'enllaç del recursos • Tipus d'enllaç: Tipificació dels enllaços segona el seu domini.	 *Il·lustració 8: Dimensió Recurs*

2.3.2 Taules de Fets i Mesures

Tot l'anàlisi demanat està comprès dins l'àmbit del tweets. És a dir, principalment sempre estarem comptant tweets. Per tant tindrem una taula central de fets on s'emmagatzemaran tots els tweets recollits.

La granularitat de la taula estarà definida a escala de tweet. És a dir el nivell de detall necessari per realitzar l'explotació de dades serà el tweet. D'aquesta manera cada registre a la nostra taula principal, correspondrà a un tweet enviat per un usuari.

A banda del recompte de tweets, les especificacions també demanen tenir present el nombre de hashtags o temes, el nombre de recursos i el nombre de mencions.

Si mirem per exemple la relació entre tweets i hashtags veurem que no és una relació típica 1 a N, ja que un tweet pot tenir N tags i 1 tag pot estar a N tweets.

Aquesta casuística s'anomena «Many to many dimensions» i és present en molts models analítics com per exemple els llibres i els seus autors o els comptes bancaris i els seus titulars, etc ...

La forma d'implementar aquest tipus de dimensions ve condicionada per la tecnologia utilitzada.

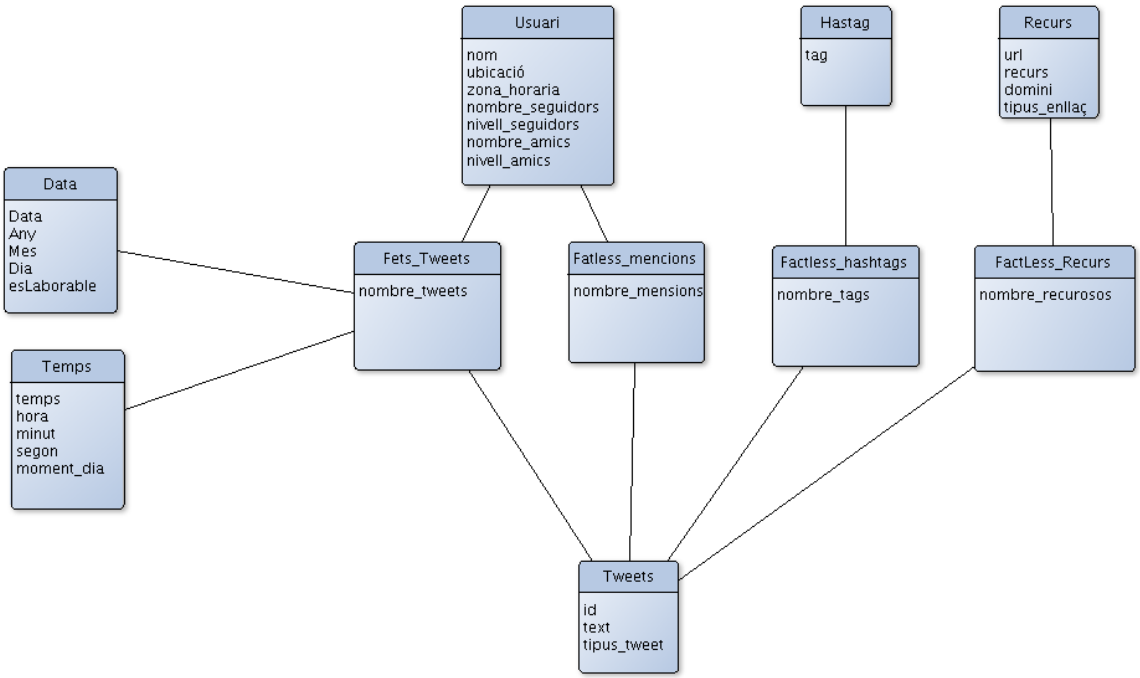
Pel que fa al nombre de hastags, recursos i mencions crearem diferents factables anomenades «factless table». És a dir, una taula de fets sense fets. Aquesta taula de fets només conté dues claus

foranes a les dimensions relacionades i serà posteriorment en la definició del model analític on indicarem aquest tipus de relació per tal de segmentar per la resta de dimensions.

Il·lustració 9: Exemple Dimensions Many 2 Many

Cal tenir present que altres tecnologies, com per exemple Mondrian, aquest tipus de relacions no són possibles i seria necessari crear diverses taules de fets (amb un cub virtual) relacionades amb el nombre màxim de dimensions possible.

Per tant el nostre model conceptual quedarà de la següent manera:



Il·lustració 10: Model Conceptual

2.3.3 Mesures

Com a mesures segons els requeriments funcionals tindrem:

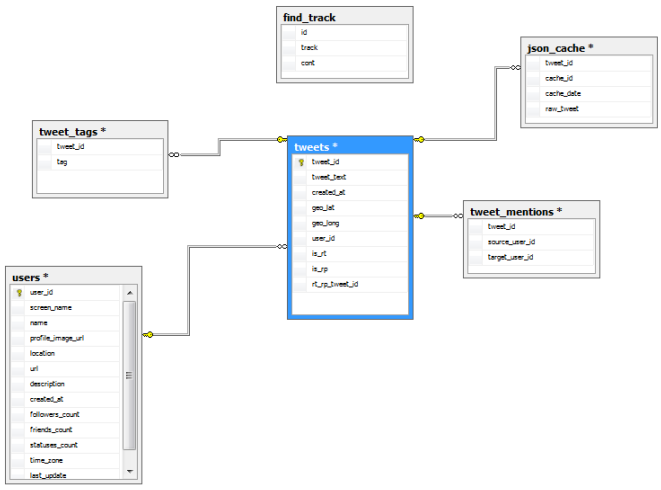
- Nombre de tweets → Recompte de files de la taula tweets
- Nombre de hashtags → Recompte de files de la taula de fets de hashtags
- Nombre de recursos → Recompte de files de la taula de fets dels recursos
- Nombre de mencions → Recompte de files de la taula de fets de mencions.

2.3.4 Origen de les dades

Identificació de la font de dades:

Per tal de realitzar el projecte amb èxit, primer s'ha d'estudiar el model de dades proporcionat i si conté tota la informació necessària que necessitem per explotar el nostre model analític.

Per la captura dels tweets s'ha utilitzat «140dev Streaming API Framework». Tota la informació capturada durant el congrés està disponible en una base de dades Mysql amb la següent estructura:



Il·lustració 11: Origen de Dades

Contrastant la informació demanada amb la disponible en la font de dades veiem que en l'anterior model de dades disposem de tota la informació:

- Temes dels tweets (tags)
- Usuaris més actius ---> Els que han realitzat més nombre de tweets
- Usuaris més seguits → Els usuaris amb major nombre de seguidors. (users → followers_cont)
- Els recursos més compartits → Taula tweet_urls
- Hastags coincidents → Taula tweet_tags
- Activitats dels assistents (enviament, reenviament de tweets) → Taula Tweet
- Si un usuari segueix a un altre → **No ho podem saber amb la font de dades proporcionades**.
- Evolució de tweets al llarg del temps → Taula tweets, camp created_at

Qualitat de les dades

Les dades recollides per l'API ja ens vénen normalitzades a les taules. De tota manera la informació referent a l'usuari és poc acurada, ja que per exemple, en el camp location no hi ha una estructura homogènia de les dades i es fa impossible poder segmentar per exemple per país o província.

Un altre dada important que ens ajudaria a netejar les dades no vàlides seria la de geolocalització. D'aquesta manera podríem detectar els tweets generats més enllà d'una àrea geogràfica concreta i revisar els tweet coincidents per tal d'esbrinar si el tag utilitzat al congrés (@model2014) pertany a un tweet del congrés o no...

Anomalies detectades:

S'ha detectat alguns tweets que fan referencia a uns usuaris inexistents a la taula users. Per tal de resoldre aquesta problemàtica, es crearà un membre nou dins de la dimensió Usuari anomenat «desconegut» per tal d'agrupar tots aquells usuaris no identificats.

2.4 Model Relacional

A continuació detallarem el disseny del nostre magatzem de dades a partir del model conceptual descrit anteriorment.

Totes les claus primàries del nostre magatzem de dades seran el que s'anomena **claus subrogades.**

Les claus subrogades són de tipus numèric i seqüencial i no tenen cap significat especial. La utilització d'aquest tipus de claus en un model de datawarehouse ens ofereix els següents avantatges:

- Ocupen menys espai que les claus natural i això repercuteix directament en l'eficiència del nostre magatzem de dades sobretot si el volum de dades és gran.
- Faciliten la construcció i manteniment dels índexs.
- El nostre DW no dependrà de la codificació interna de la font de dades (OLTP).
- Permet la correcta aplicació de tècniques com SCD (Slow Changing Dimensions).

Per tant a l'hora de definir les dimensions , hem de mirar quin volum de dades tenim per tal de triar el tipus de dada adequat per la nostra clau subrogada.

2.4.1 Dimensions

Dimensió Data

Per la dimensió temps, utilitzarem un script TSQL per generar els registres necessaris. Si observem la font de dades, podem veure que els tweets més antics daten de setembre del 2014. Així doncs podríem fer una taula de temps que anés des de el 2014 fins el 2016.

La nostra clau subrogada serà de tipus int que només ocupa 4 bytes i dóna escalabilitat al model, ja que podem emmagatzemar fins a 2.147.483.647 registres.

Taula Dim_data

Nom	Tipus de dada	Clau	Descripció

id	int	PK,SK	Clau primària
Data	Date	-	Format dd/mm/aaaaa
Any	int	-	Format de 4 dígits
Mes	int		Número de mes
Mes_cat	varchar(30)		Nom del més (Català)
dia	int	.	Dia del mes
Eslaborable	Int		Indica si es laborable o no (0/1)

Dimensió temps

La dimensió temps tindrà una mida fixa, ja que l'interval va des de les 00:00:00 a les 23:59:00.

Per tant si multipliquem 24h * 60 minuts * 60 segons, tindrem un total de 86400 registres.

Taula Dim_temps

Nom	**Tipus de dada**	**Clau**	**Descripció**
TimeSK	int	PK,SK	Clau primària
Time	char(8)	-	Format hh:mm:ss
Hour	char(2)	-	Format de 2 dígits de 0 a 24h
Minute	char(2)		Format de 2 dígits
segon	char(2)		Format de 2 dígits
Range	Varchar(30)		Moment del dia (matí/tarda/nit)

Dimensió Tweet

El volum de la dimensió tweet dependrà de les dades carregades. En el nostre escenari inicial 2230 tweets però cara a donar escalabilitat al nostre model, la nostra clau subrogada serà de tipus int.

Taula Dim_tweet

Nom	Tipus de dada	Clau	Descripció
tweet_id	int	PK,SK	Clau primària
source_pk	bigint		Clau primària a OLTP
text	varchar(140)	-	Texte del tweet
Tipus de tweet	char(1)		Número de mes

Dimensió Usuari

El volum de la dimensió usuari també serà dinàmic ja que depèn del nombre de tweets que tinguem carregats en el nostre datawarehouse.

Si fem un recompte a la font de dades trobem 447 usuaris diferents. Tot i això al igual que els tweets, definirem com a clau subrogada un tipus de dada tipus int.

Taula Dim_users

Nom	Tipus de dada	Clau	Descripció
user_id	int	PK,SK	Clau primària
source_pk	bigint	-	Clau primària a OLTP
name	varchar(50)	-	Texte del tweet
location	varchar(50)		Número de mes
timezone	varchar(50)		
n_followers	int		
n_friends	int		

Dimensió Hashtag

Al igual que la dimensió usuari, la dimensió hastags és dinàmica i també depèn del nombre de tweets del nostre model.

A la font de dades tenim un total de 355 hashtags diferents tot i això definirem la nostra clau subrogada com a int per tal de donar escalabilitat al model.

Taula Dim_hastag

Nom	Tipus de dada	Clau	Descripció
hashtag_id	int	PK,SK	Clau primària
description	varchar(100)		Texte del tag.

Dimensió Recurs

A la font de dades trobem 162 recursos diferents. Al igual que les dimensions anteriors definirem la nostra clau subrogada com a enter.

Taula Dim_resources

Nom	Tipus de dada	Clau	Descripció
resource_id	int	PK,SK	Clau primària
description	varchar(140)		url_completa
domain	varchar(80)	-	Domini URL
content_type	tinyint		Tipus enllaç

2.4.2 Taules de fets

Fact_tweets

Taula principal de fets on es guardaran tots els tweets registrats.

Taula Fact_tweets

Nom	Tipus de dada	Clau	Descripció
id	int	PK	Clau primària
tweet_fk	int	FK	Fk a la dimensió tweet

user_fk	int	FK	FK a la dimensió usuari
data_fk	int	FK	FK a la dimensió data
temps_fk	int	FK	FK a la dimensió temps

Fact_hastags

Factless table que relaciona la dimensió tweet amb la dimensió hashtag.

Taula Fact_hastags

Nom	**Tipus de dada**	**Clau**	**Descripció**
tweet_fk	int	FK	FK a la dimensió tweet
hashtag_fk	int	FK	FK a la dimensió hashtag

Fact_resources

Factless table que relaciona la dimensió tweet amb la dimensió resources.

Taula Fact_resources

Nom	**Tipus de dada**	**Clau**	**Descripció**
tweet_fk	int	FK	FK a la dimensió tweet
resource_fk	int	FK	FK a la dimensió recurs

Fact_mentions

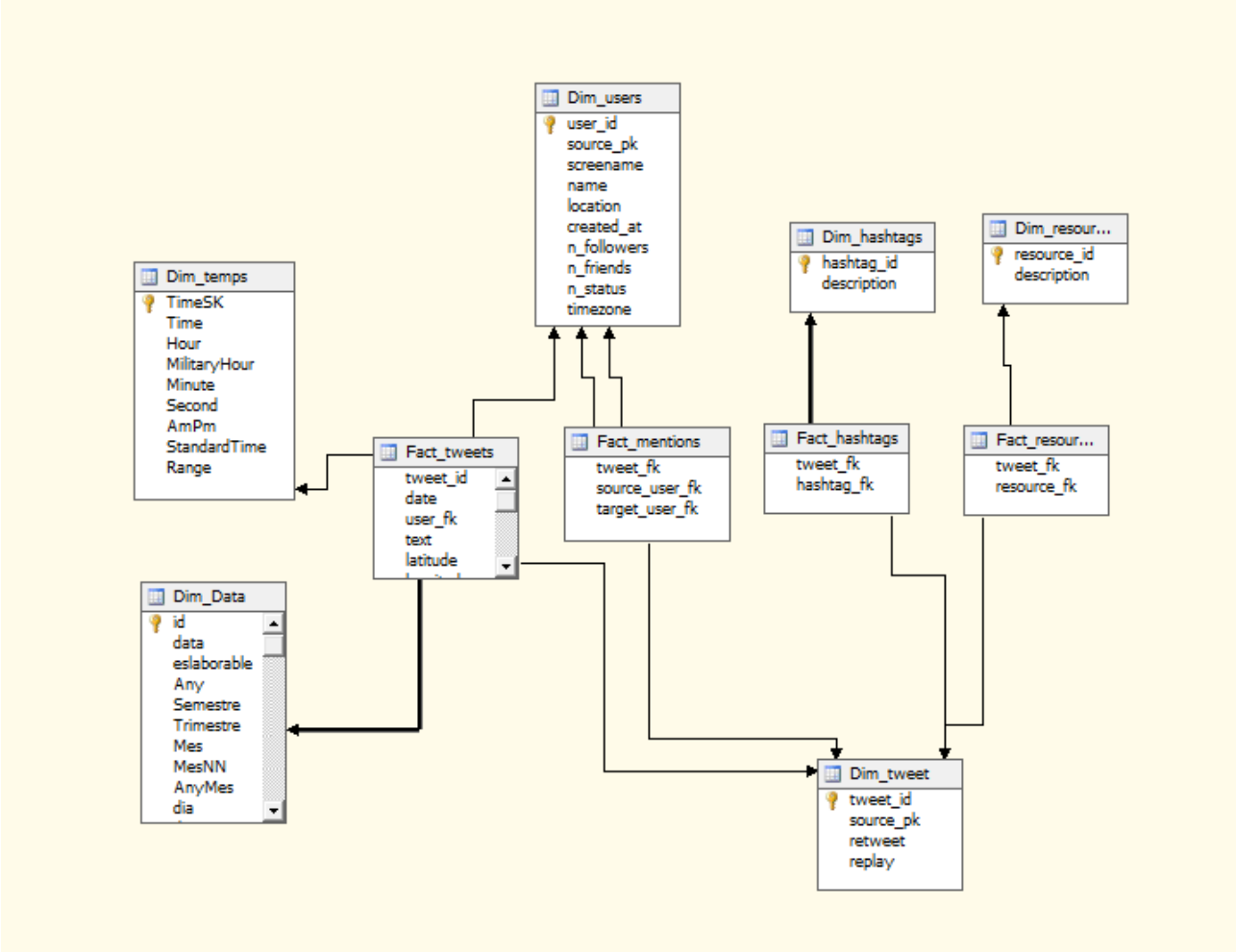
Factless table que relaciona la dimensió tweet amb la dimensió usuari.

Taula Fact_mentions

Nom	**Tipus de dada**	**Clau**	**Descripció**
tweet_fk	int	FK	FK a la dimensió tweet

source_user_fk	int	FK	FK a la dimensió usuari
target_user_fk	Int	FK	FK a la dimensió usuari

2.4.3 Diagrama E-R



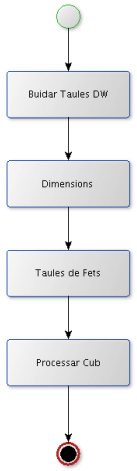
Il·lustració 12: Diagrama E-R

2.5 Procés ETL

Definirem un procés ETL de càrrega completa. És a dir, cada vegada que tinguem dades noves, eliminarem les dades del nostre datawarehouse i carregarem totes les dades disponibles de la font de dades.

El procés general de la càrrega serà:

1. Buidar taules DW
2. Càrrega de dimensions
3. Càrrega de taules de fets
4. Processament del cub.



Il·lustració 13: Diagrama Activitats - Procés ETL

2.5.1 Dimensions

Dimensió Data

Per la creació dels valors de la taula dim_data, utilitzarem un script TSQL que ens generi tota la informació necessària entre dues dates.

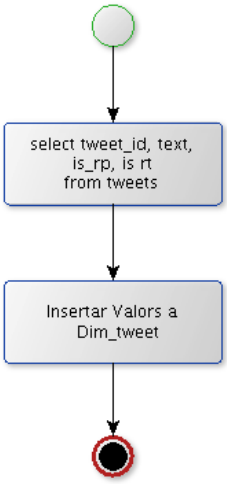
Dimensió Temps

Per la creació de la dimensió temps, també utilitzarem un script per generar tots els valors necessaris entre les 00:00:00 i les 23:59:00

Dimensió Tweet

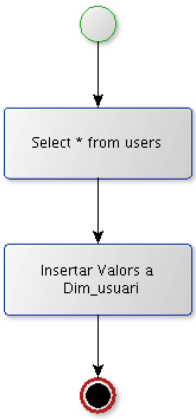
A origen tenim la taula tweets. Per tal de construir la nostra taula de dimensió només necessitem els següents camps :

- tweet_id
- text
- is_rp
- is_rt

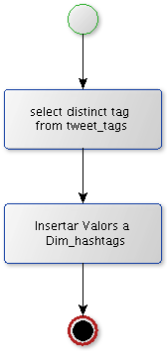


Il·lustració 14: Diagrama Activitat: Dim tweet

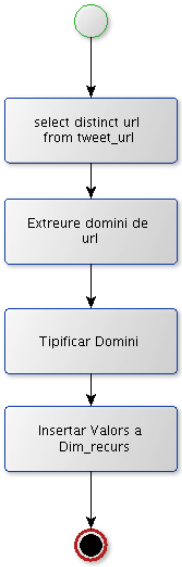
Dimensió Usuari

Per crear la dimensió usuari utilitzarem tots els camps de la taula mysql users. Per tant el ETL serà molt senzilla.	 *Il·lustració 15: Diagrama Activitat: Dim Usuari*

Dimensió Hashtag

De la taula origen seleccionem tots els tweets diferents.	 *Il·lustració 16: Diagrama Activitat: Dim Hashtag*

Dimensió Recurs

De cada recurs a partir de la seva url, extraurem el domini i mirarem si el domini el tenim tipificat per tal de categoritzar-lo.	 *Il·lustració 17: Diagrama Activitat: Dim Recurs*

2.5.2 Fets

El procés de les diferents taules de fets serà molt similar. Al treballar amb claus subrogades haurem d'anar a buscar a la taula de dimensió la seva clau primaria per tal d'introduir-la a la taula de fets

Per tal d'agilitzar el procés és molt important indicar a la ETL que carregui tota la taula de dimensió en memòria per cercar les claus primàries. De no fer-ho d'aquesta manera, el procés ETL hauria de realitzar una *select* per cada registre per tal d'esbrinar la seva clau primària i el temps de càrrega seria molt superior.

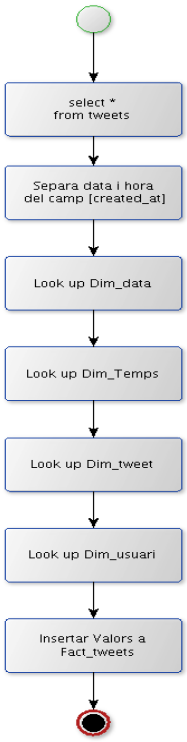
A continuació detallem els diagrames d'activitats de les càrregues de les taules de fets:

Fact_tweets

Primerament haurem de separar el camp *creted_at* amb dos camps: Un per la data i l'altre per l'hora.

Després haurem de buscar les claus primàries a les taules de dimensions corresponents (Operació lookup)

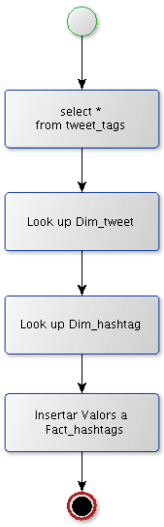
Finalment introduirem les dades a la taula de fets amb les seves claus primàries.



Il·lustració 18: Diagrama Activitat: Fact_tweet

Fact_hastags:

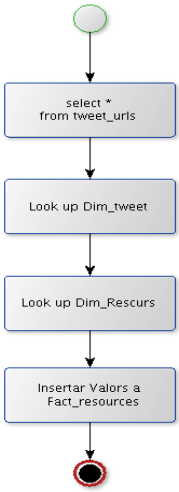
El procés d'aquesta càrrega és molt similar a l'anterior. Únicament hem de buscar les claus primàries de la dimensió tweet i la dimensió Hash_tag



Il·lustració 19: Diagrama Activitat: Fact_hashtag

Fact_resources:

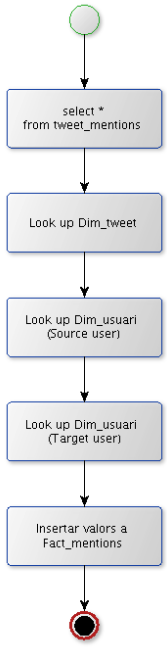
En el cas de la taula de fets de recursos, necessitem les claus primàries de la dimensió tweet i de la dimensió recurs.



Il·lustració 20: Diagrama Activitat: Fact_resources

Fact_mentions

A la taula Fact mentions al tenir dues claus foranes contra la mateixa dimensió haurem de fer dos lookups. Un per el camp source_user i l'altre per el target_user.



Il·lustració 21: Diagrama Activitat: Fact_mentions

2.5.3 Processar Cub

L'última part del nostre procés ETL serà l'encarregat de processar les dimensions i el cub amb les dades carregades anteriorment. Integrations Services disposa d'un component especial per realitzar aquesta tasca.

2.5.4 Execució ETL i errors de càrrega

Per tal d'executar el procés ETL, crearem un JOB a l'agent del SQL Server per tal que executi tot el procés de càrrega.

També configurarem aquest Job perquè envii un correu electrònic en cas de que la càrrega no s'executi correctament.

Per tal d'executar aquest treball, crearem un informe dins el Reporting Services que ens indiqui la data i l'hora de l'última càrrega així com un botó que ens permeti executar la ETL de forma manual.

Com que la nostra font de dades prové d'una altre base de dades, no trobarem errors de tipus de dades incorrectes ja que venen correctament tipades des de origen.

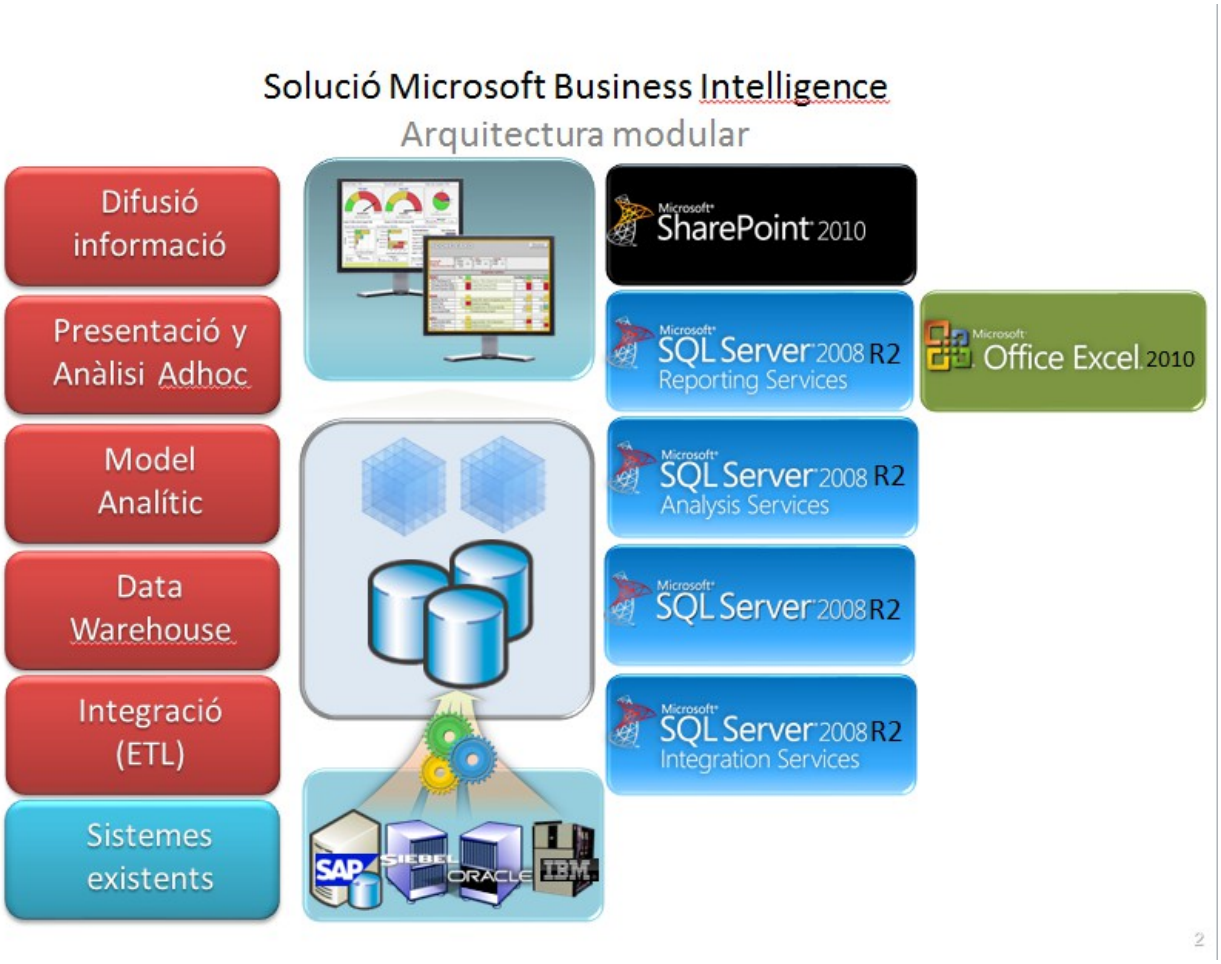
El que si ens podem trobar a l'hora de processar el nostre cub son registres a les taules de fets que fan referència a membres inexistents d'alguna dimensió.

En el nostre cas concret, hem trobat tweets que feien referència a usuaris que no estaven a la taula users. Per tal de resoldre aquest problema, hem creat un membre nou dins la dimensió usuari anomenat «Desconegut» que recull tots aquells tweets de usuaris no identificats.

3. Desenvolupament

3.1 Programari Utilitzat

A continuació presentem l'arquitectura de Microsoft per la realització del projecte.



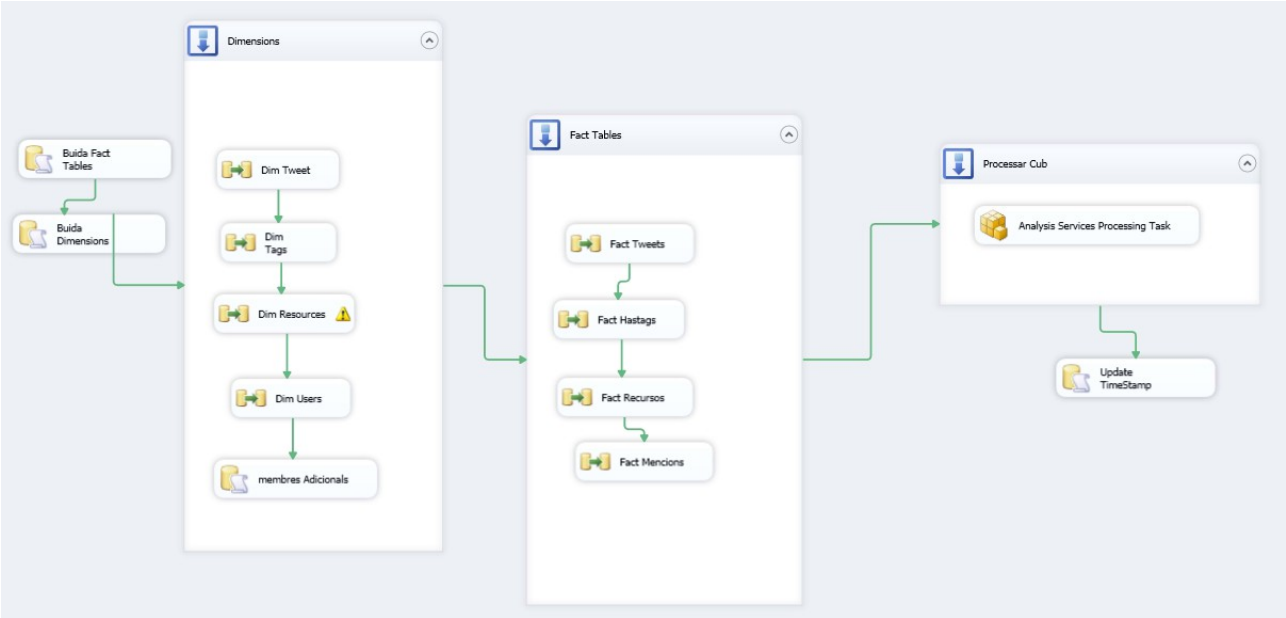
Il·lustració 22: Stack Microsoft SQL Server

3.2 Procés ETL

Per la realització de la càrrega de dades utilitzarem Integration Services (SSIS). Aquesta eina proporciona diferents components que faciliten la càrrega i transformació de les dades de diferents orígens.

Tots els components de l'eina, s'organitzen en paquets (fitxers) que permeten executar els processos d'una forma seqüencial o concurrent.

A continuació es mostra l'estructura del paquet utilitzat que es correspon en bona mesura al diagrama d'activitats descrit en l'apartat de disseny de la solució:



Il·lustració 23: SSIS: Paquet principal

Com es pot veure, el paquet SSIS està estructurat en diferents parts i les fletxes indiquen el flux d'execució del procés:

1. Eliminem dades prèvies en les taules de fets i dimensions
2. Realitzem la càrrega de les dimensions
3. Realitzem les càrrega de les taules de fets
4. Processem el cub OLAP
5. Actualitzem la taula de control

3.2.1 Càrrega de Dimensions

En aquest apartat es realitza la càrrega de les dades de totes les dimensions a excepció de les

taules dim_Data i dim_temps que ja s'han creat en la part de construcció del model relacional.

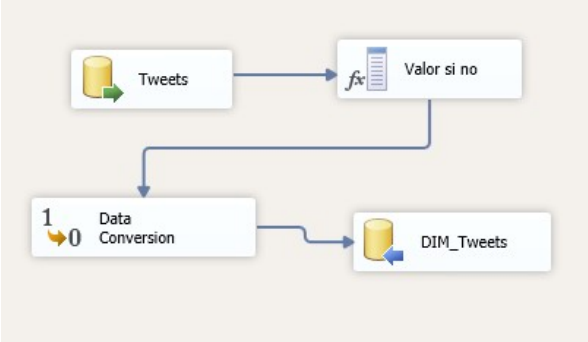
Per cada dimensió utilitzarem un Data Flow Task que ens permetrà realitzar tot el procés d'extracció, transformació i càrrega de dades en les taules del nostre datawarehouse.

Dimensió tweet:

En el primer pas executem la següent consulta sobre la taula tweets:

« **select** tweet_id, is_rp,is_rt,tweet_text **from** tweets;»

Amb el resultat apliquem un transformació a les columnes is_rp, is_st i seguidament introduïm les dades a la taula Dim_tweet del nostre magatzem de dades.

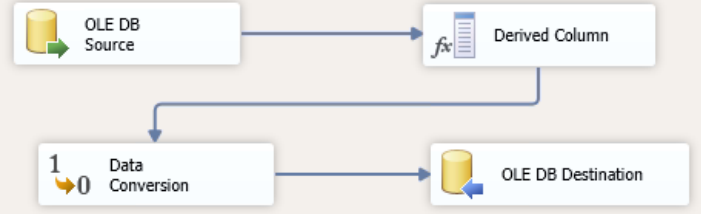


Il·lustració 24: SSIS: Dim tweet

Dimensió hashtag

A partir de la taula tweet_tags seleccionem els tags únics a partir de la següent sentència SQL:

« select distinct tag **from** tweet_tags »



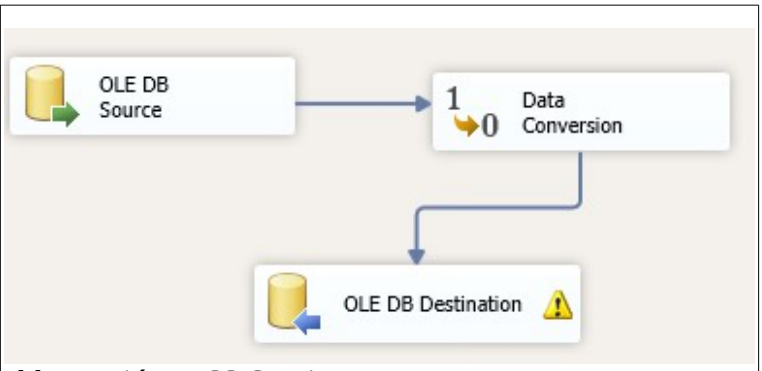
Il·lustració 25: SSIS: Dim Hashtag

Dimensió recurs

Per carregar els recursos utilitzarem la següent SQL, juntament amb la funció parseURL per tal d'extreure el domini de la url.

« **select distinct** url, [dbo].parseURL(url) **as domain from** tweet_urls »

El resultat l'emmagatzemarem a la taula Dim_resources.



Il·lustració 26: SSIS: Dim Recurs

Dimensió usuari

La taula usuaris la traspassem amb tots els camps de l'origen de dades (taula users) al nostre datawarehouse a la taula Dim_users.

Il·lustració 27: SSIS: Dim Usuari

En l'anàlisi de la font de dades, vam descobrir que hi ha havia alguns tweets que feien referència a usuaris que no eren presents a la taula users.

Per tal de tenir en compte aquests tweets, afegim a la dimensió usuari un membre anomenat «Desconegut» per tipificar totes aquestes piulades.

```
INSERT INTO [dbo].[Dim_users]
        (user_id, [source_pk] ,[screename] ,[name]  ,[location] ,[created_at],[n_followers],[n_friends],[n_status],
[timezone])
VALUES
        (999999,999999,'Desconegut ' ,'Desconegut ','Desconegut ' ,null  ,0 ,0 ,0,'Desconegut ');
```

3.2.2. Càrrega de taules de Fets

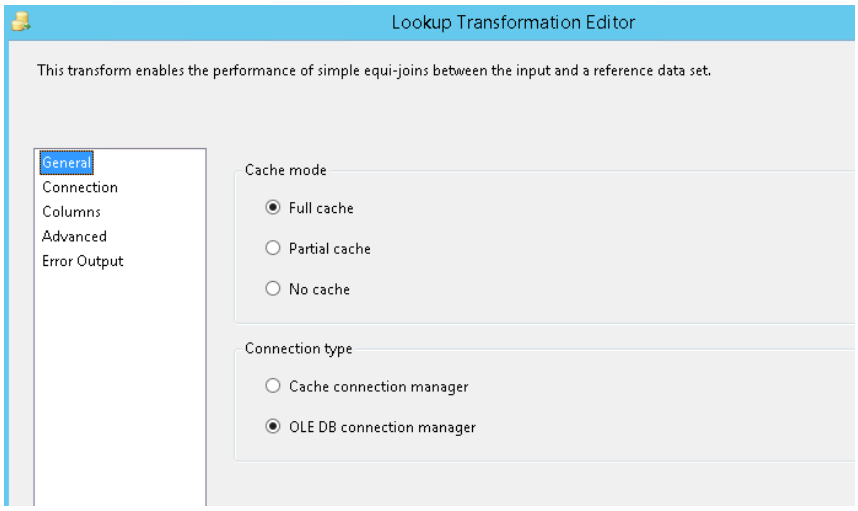
El procés de la càrrega de dades en les taules de fets serà molt semblant per totes elles.

Al treballar amb claus subrogades, per cada clau forana que tinguem haurem d'anar a buscar a la taula de dimensió la clau primària real.

SSIS disposa d'un component anomenat *lookup* que permet buscar un valor del flux de dades en una altre taula.

S'ha d'anar molt en compte a l'hora d'utilitzar aquest component ja que si no el tenim ben configurat ens pot penalitzar molt el temps del procés de càrrega ja que per cada registres de la fact table hauríem de fer tantes consultes addicionals com dimensions disposi el model.

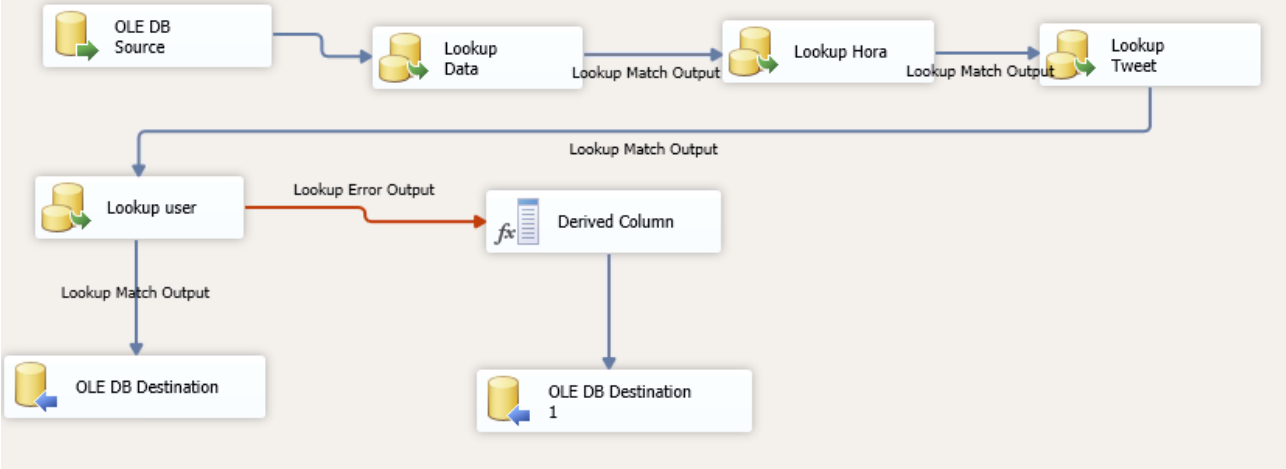
Per tal de configurar correctament aquest component, s'ha d'indicar que realitzi *full cache* de les dades ja que d'aquesta manera carrega en memòria tots els registres de la taula de la dimensió i fa el lookup en memòria millorant substancialment el temps de procés.



Il·lustració 28: SSIS: Component Lookup

Una altre bona alternativa per a realitzar aquesta funcionalitat sense utilitzar el component de lookup és la de realitzar-ho en el SGBD mitjançant «joins». Si per exemple utilitzem *lef join* sabrem quins valors son coincidents i quins no. De vegades és mes idoni crear una àrea de Stagging on dipositar les taules origen i treballar-les mitjançant l*eft joins*.

Taula Fact_tweets



Il·lustració 29: SSIS: Taula de fets tweets

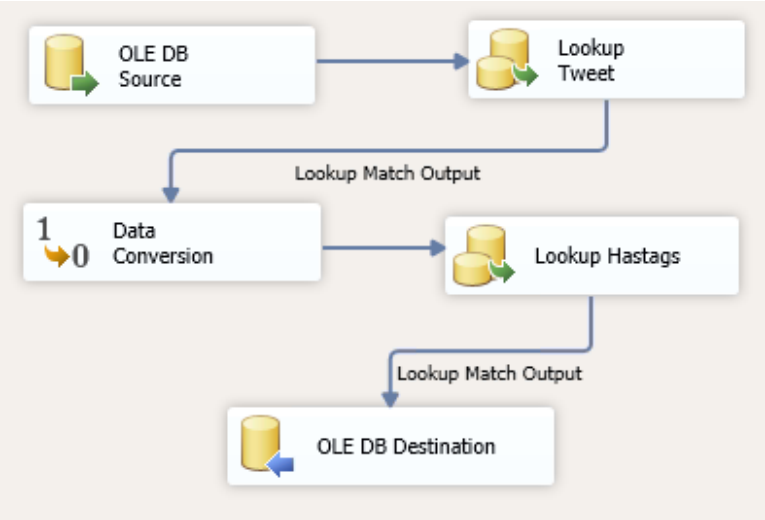
La sentència SQL utilitzada per obtenir les dades és la següent:

```
select *, Convert(varchar,[created_at],103) as data_sola, Left( convert(varchar,created_at, 114),8) as
hora_minut
from tweets;
```

Hem de separar la data i l'hora del component datetime que tenim en origen per tal de poder treballar amb la dimensió data i la dimensió temps.

També cal destacar el lookup al component usuari ja que s'ha definit que si no es troba l'usuari a la taula de la dimensió, posi el valor del usuari «Desconegut» definit anteriorment.

Taula Fact_hashtags



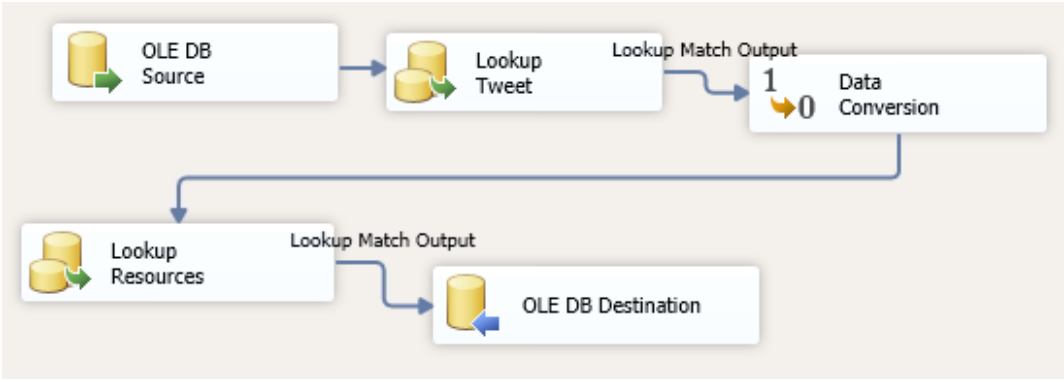
Il·lustració 30: SSIS: Taula de fets hashtags

El procediment per la taula de fets de hashtags és molt similar a l'anterior.

La sentència SQL utilitzada per recuperar les dades és la següent:

```
select *,upper(tag) as tag_upper from tweet_tags
```

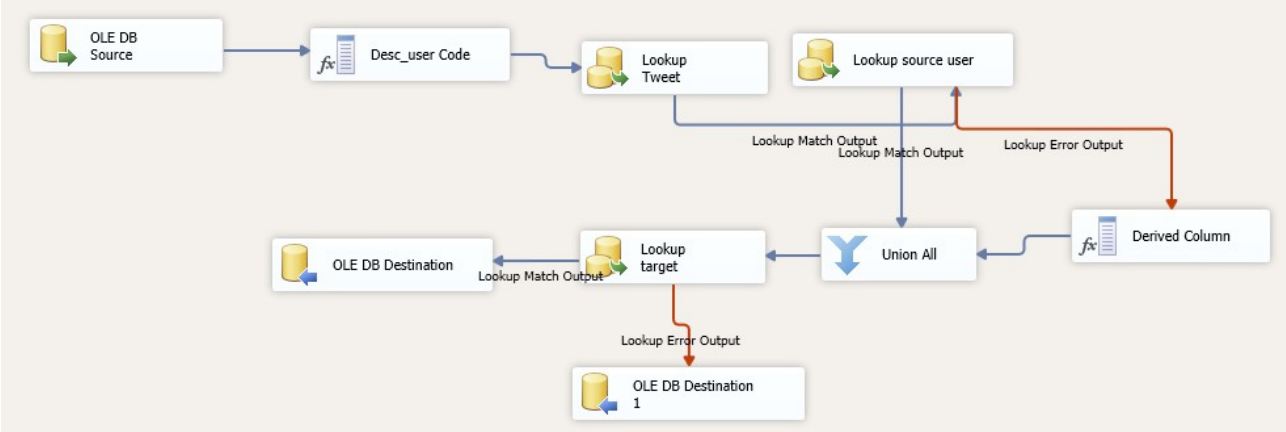
Taula Fact_resources:



Il·lustració 31: SSIS: Tauls de fets de recursos

Per tractar els recursos, recuperem totes les dades de la taula *tweet_urls* de orígens i una vegada realitzats els lookups corresponents emmagatzemem el resultat a taula *Fact_resources*.

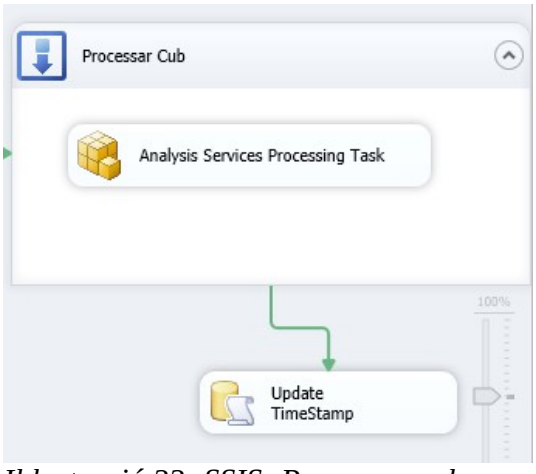
Taula Fact_Mentions



Il·lustració 32: SSIS: Taula de fets de mencions

Al igual que en els recursos, utilitzarem totes les dades de la taula *tweet_mentions* per tal de obtenir les claus necessàries per guardar les dades a la taula *Fact_Mentions*

Processament del cub i actualització de taula de control



Il·lustració 33: SSIS: Processar cub

Al treballar amb un sistema MOLAP, és necessari una vegada tenim les dades carregades en el nostre datawarehouse processar el cub per tal de tenir el model analític actualitzat.

SSIS disposa de uns components propis per processar el cub amb les seves dimensions.

Una vegada finalitzat el tot el procés, actualitzem la taula de control per registrar la marca de temps de la càrrega realitzada.

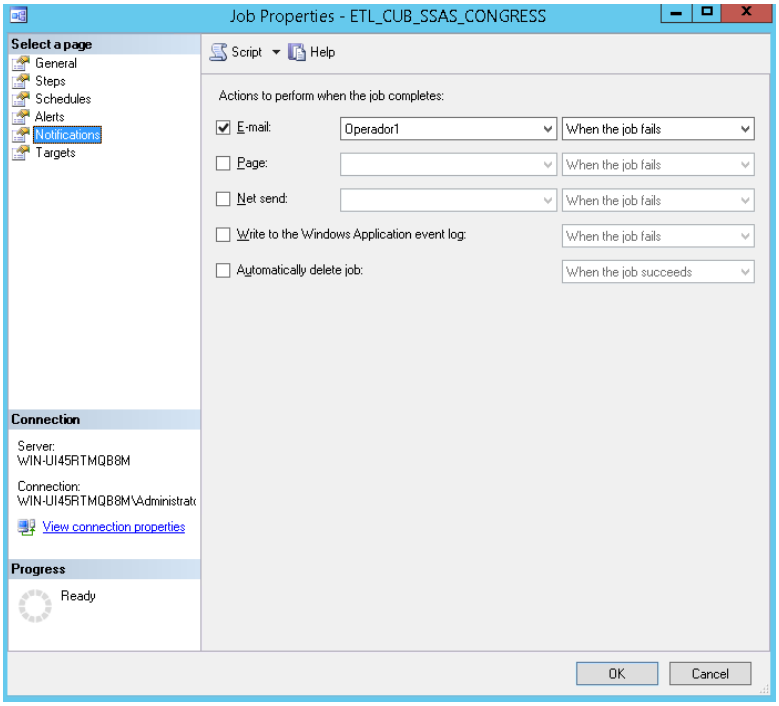
3.2.3 Automatització de càrrega i control d'errors

El servei del SQL AGENT permet crear treballs per realitzar tot tipus de tasca.

En el nostre cas, crearem un treball perquè executi el nostre paquet de Integration Services i en cas de error envii un correu electrònic a l'operador definit

L'agent de SQL té un registre dels treballs executats amb el seu log associat de tal manera que si per alguna raó fallés la càrrega de dades, tindríem la informació de l'error perfectament classificada.

Cal tenir en compte que el cub sempre el tindrem disponible amb una versió estable. És a dir, en cas de fallada de la ETL o del processament del cub, sempre tindrem disponible per l'usuari la versió que hi havia abans d'executar el job.



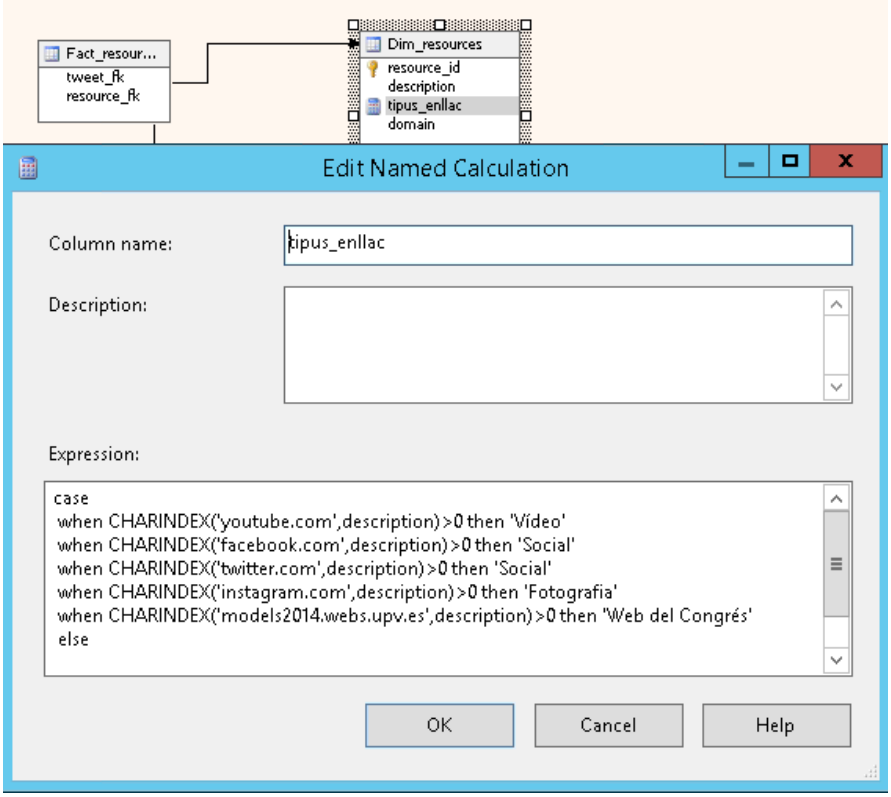
Il·lustració 34: SQL Agent: Configuració de l'operador

3.3 Fitxer DSV

A l'hora de definir el model analític, Analysis Services crea una capa lògica del model relacional. És a dir, una vista de totes les metadades del model relacional on els principals avantatges son :

1. Incloure taules o vistes de diferents bases de dades: Per exemple si tenim en una base dades la taula de la dimensió temps, no cal copiar la taula de temps a tots els magatzems de dades. En el dsv, definim la connexió on s'ubica la taula de temps i la importem en el nostre projectes
2. Creació de camps calculats o vistes de taules: És molt comú anar adaptant el model analític amb nous atributs que s'obtenen a partir de la informació existent. Mitjançant el dsv podem definir aquestes nous camps calculats i d'aquesta manera no s'ha de modificar tot el

procés ETL.



Il·lustració 35: DSV: Creació de camps calculats

A continuació detallem els camps calculats definits en el fitxer dsv:

Nom Taula	Nom Camp Calculat	SQL
Dim_resources	tipus_enllac	case when CHARINDEX('youtube.com',description)>0 then 'Vídeo' when CHARINDEX('facebook.com',description)>0 then 'Social' when CHARINDEX('twitter.com',description)>0 then 'Social' when CHARINDEX('instagram.com',description)>0 then 'Fotografia' when CHARINDEX('models2014.webs.upv.es',description)>0 then 'Web del Congrés' else

Nom Taula	Nom Camp Calculat	SQL
		'Sene tipificar' **end**
Dim_users	nivell_seguidors	**case** **when** n_followers<50 **then** 'Baix' **when** n_followers<1000 **then** 'Moderat' **else** 'Alt' **end**
Dim_users	nivell_amics	**case** **when** n_friends<50 **then** 'Baix' **when** n_friends<100 **then** 'Moderat' **else** 'Alt' **end**

3.4 Model Multidimensional

3.4.1 Dimensions

La creació de les dimensions mitjançant SQL server tools és un procés molt senzill. A través d'un assistent es selecciona la taula de la dimensió que es vol crear i s'especifica quines columnes de la taula volem tenir com a atributs.
Una vegada generada la dimensió podem crear jerarquies arrossegant els atributs a la pestanya «Hierarchies»

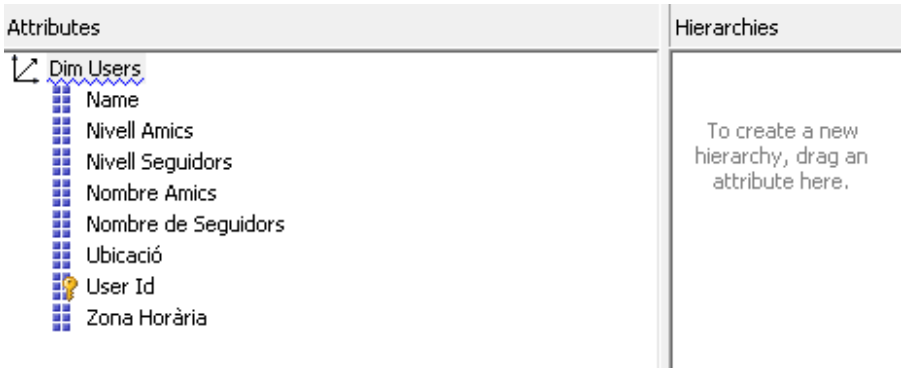
Dim_data:

Il·lustració 36: SSAS: Dimensió Data

Dim Tweet:

Il·lustració 37: SSAS: Dimensió Tweet

Dim Users:



Il·lustració 38: SSAS: Dimensió Usuari

Dim Temps:

Il·lustració 39: SSAS: Dimensió Temps

Dim Hashtag:

Il·lustració 40: SSAS: Dimensió Hashtag

Dim Recurs:

Il·lustració 41: SSAS: Dimensió Recurs

3.4.2 Grups de mesures / Taules de fets

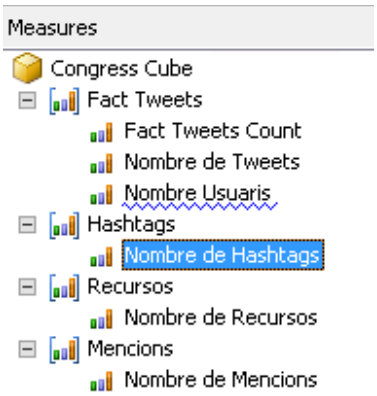
En el model analític, els grups de mesures equivalen habitualment a les taules de fets del nostre magatzem de dades.

A partir de la definició del nostre model relacional detallat en el fitxer dev, indicarem les taules de fets amb les mètriques que utilitzarem en el model, juntament amb el seu tipus d'agregació.

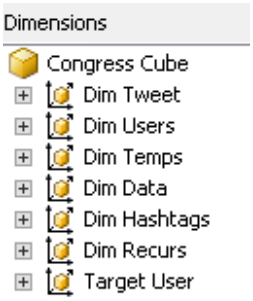
A la següent taula resumim totes les mètriques creades a en el projecte

Nom Mètrica	Grup de Mesures	Taula de fets	Tipus Agregació
Nombre de Tweets	Tweets	Fact_tweets	Row count
Nombre de Usuaris	Tweets	Fact_tweets	Distinct Count (user_fk)
Nombre de Hashtags	Hashtags	Fact_hashtags	Row count
Nombre de Recursos	Recursos	Fact_resources	Row Count
Nombre de Mencions	Mencions	Fact_Mencions	Row Count

Finalment, en les següents imatges mostrem les mètriques associades a cada grup de mesura així com les dimensions associades al cub.



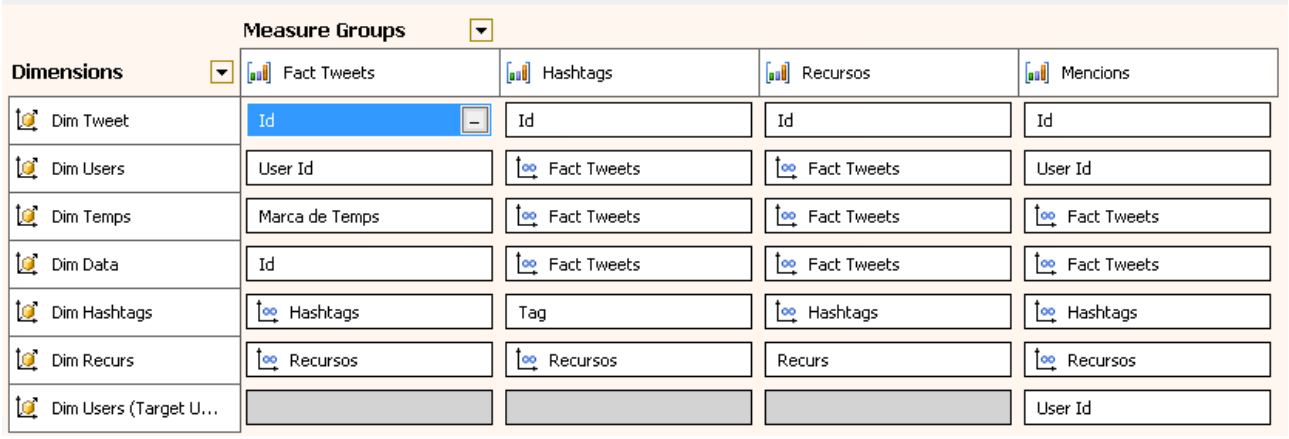
Il·lustració 43: SSAS: Grups de mesures i mètriques



Il·lustració 42: SSAS: Dimensions associades

3.4.3 Relacions entre grups de mesures

L'ultim aspecte i no menys important és la relació que tenen les dimensions amb els diferents grups de mesures definits:



Il·lustració 44: SSAS: Relacions Grups de mesures

Aquí aprofitem la possibilitat de definir les relacions «many 2 many» comentades en la fase de disseny.

És molt important tenir clar que la possibilitat que ens dóna Analysis Services de definir relacions many 2 many ja que ens determina tant la construcció del model analític com la construcció del nostre model relacional.

3.4.4 Càlculs MDX

Per la realització dels informes, hem definit uns quants càlculs MDX que ens seran de molta utilitat:

Dynamic sets: Els conjunts dinàmics permet definir quins membres s'han de mostrar en la consulta MDX, tenint en compte el segmentador utilitzat a la consulta.

En l'apartat dels informes que veurem posteriorment hem definit diferents datasets que utilitzen aquests conjunts. Per exemple, tenim un dataset que ens mostra els 10 tags més utilitzats. Així doncs si apliquem aquest conjunt dinàmic en una consulta segmentada per usuari, ens mostrarà els 10 tags més utilitzats per l'usuari seleccionat.

La utilització de conjunts dinàmics ens millora l'eficiència dels informes ja que el cub ens mostrarà només la informació que necessitem i no serà necessària filtrar la informació en la capa de reporting, millorant així la quantitat de dades que envia el cub cap a l'informe.

Càlculs MDX – Conjunts Dinàmics

```
CALCULATE;

CREATE SESSION DYNAMIC SET [Congress Cube].[Top 10 Hashtags]
 AS TOPCOUNT([Dim Hashtags].[Tag].[All].Children, 10,[Measures].[Nombre de Hashtags]);


CREATE SESSION DYNAMIC SET [Congress Cube].[Top 30 Dominis]
 AS TOPCOUNT([Dim Recurs].[Domini].[All].Children, 30,[Measures].[Nombre de Recursos]);


CREATE SESSION DYNAMIC SET [Congress Cube].[Usuaris més populars]
 AS TOPCOUNT([Target User].[User Id].[All].Children, 10,[Measures].[Nombre de Mencions]);
```

3.5 Informes

La construcció dels informes s'ha realitzat amb el Reporting Services de Microsoft.

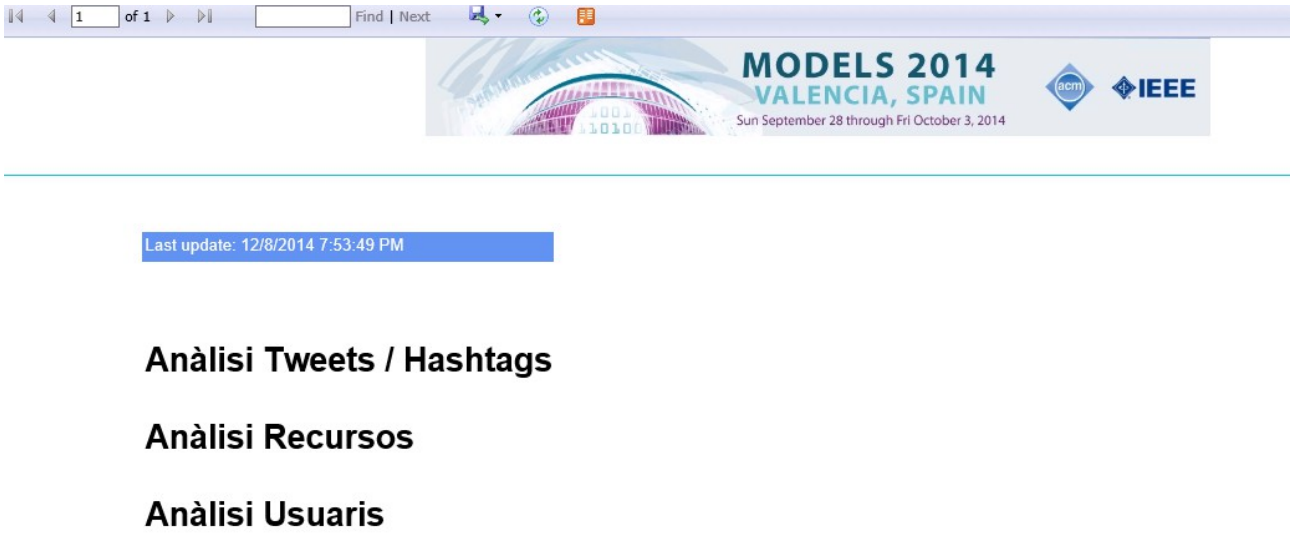
Dins de la solució TFC_2014 s'ha creat un projecte SSRS_Congress on en poden trobar tots els informes:

A continuació es detallarà cada un dels informes construïts:

index.rdl

Aquest informes és molt simple. És només una pàgina inicial que ens servirà de menú on s'indica els diferents informes creats i mostra la data de l'ultima actualització del cub.

Tots els informes tenen la funcionalitat de tornar a la pàgina del menú mitjançant el logo definit en la part superior dels informes.



Il·lustració 45: SSRS: Informe portada

tweets.url

En aquest informe es mostra tots els indicador més significatius per l'anàlisi dels tweets i els seus hashtags.

A l'apartat de filtre, tenim la dimensió temps que ens permet filtrar les dades de l'informe per qualsevol atribut de la dimensió (any,mes o dia).

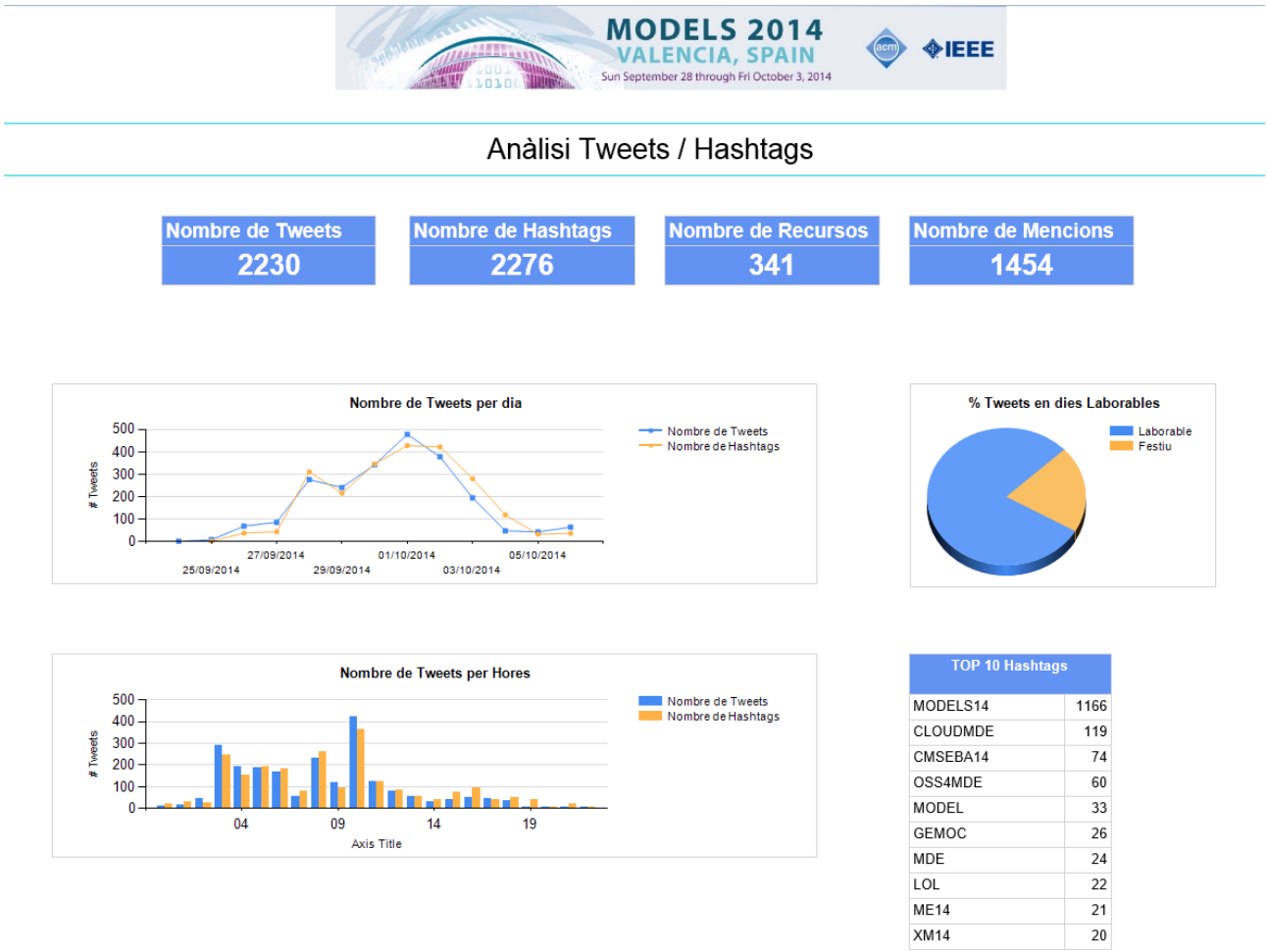
L'informe està format per quatre datasets :

Data_set_tweets_mensuals: Consulta MDX que retorna els tweets per mes.

DataSet1_hores: Consulta MDX que retorna el nombre de tweets per hora.

DataSet1_resumTotal: Consulta MDX que retorna els valors totals de cada una de les mètriques.

DataSet1_topHashTags: Consulta MDX que mostra els 10 tags més utilitzats.



Il·lustració 46: SSRS: Anàlisi de tweets

recursos.rdl

En aquest informe es mostra tota la informació relacionada amb els recursos. El nombre de recursos compartits, elt tipus de recursos I els dominis compartits mé populars.

A l'apartat de filtre, tornem a tenir la dimensió temps que ens permet filtrar les dades de l'informe per qualsevol atribut de la dimensió (any,mes o dia).

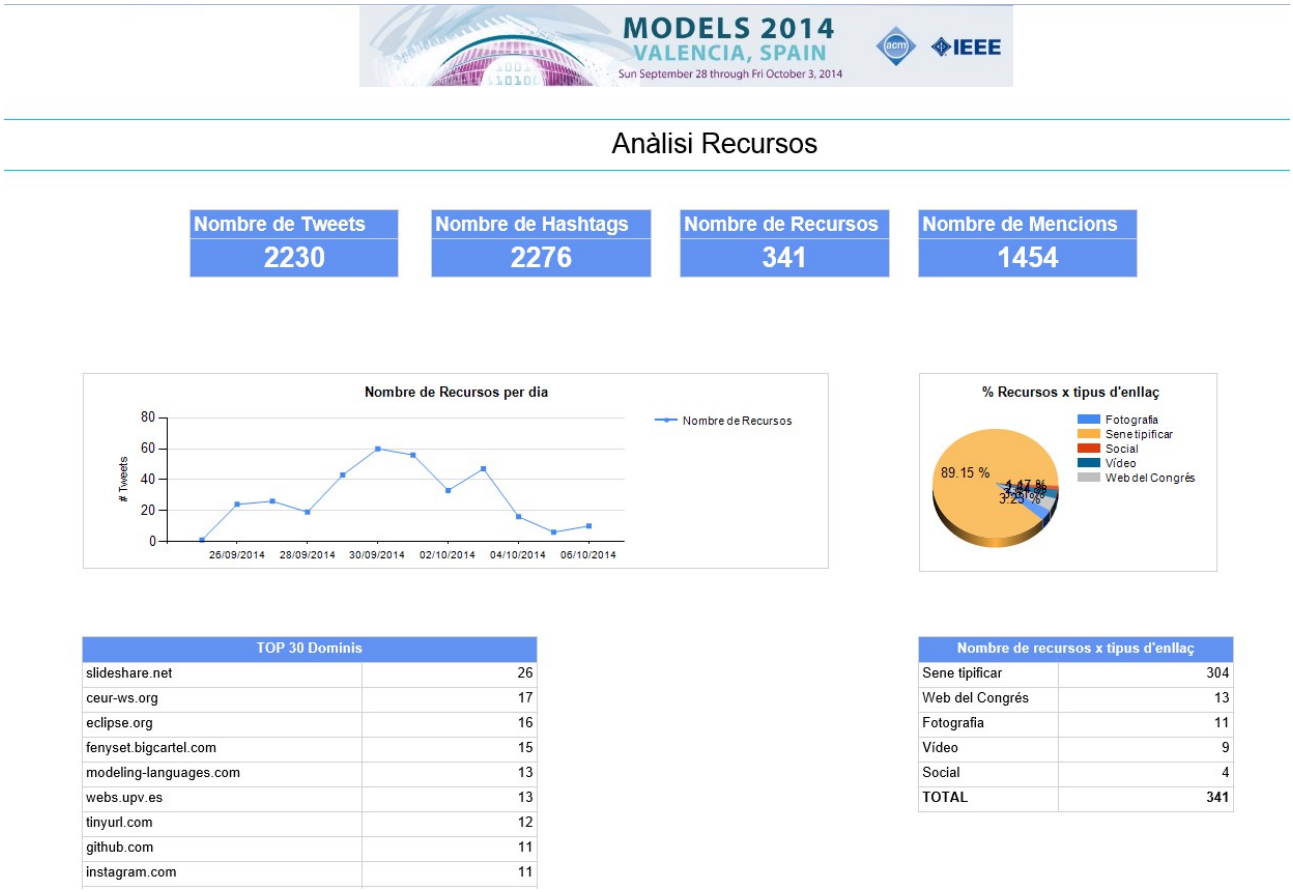
L'informe està format per quatre datasets :

DataSet_recursos_diaris: Consulta MDX que retorna els recursos compartits per dia.

DataSet_tipus_recurs: Consulta MDX que retorna el nombre recursos segmentat per tipus..

DataSet1_resumTotal: Consulta MDX que retorna els valors totals de cada una de les mètriques.

DataSet1_topDominis: Consulta MDX que mostra els 30 dominis més compartits.



Il·lustració 47: SSRS: Anàlisi de recursos

usuaris.rdl

L'informe d'usuaris, ens mostra el nombre de participants en l'esdeveniment així com la seva activitat durant el congrés.

A l'apartat de filtre, tornem a tenir la dimensió temps que ens permet filtrar les dades de l'informe per qualsevol atribut de la dimensió (any,mes o dia).

Datasets:

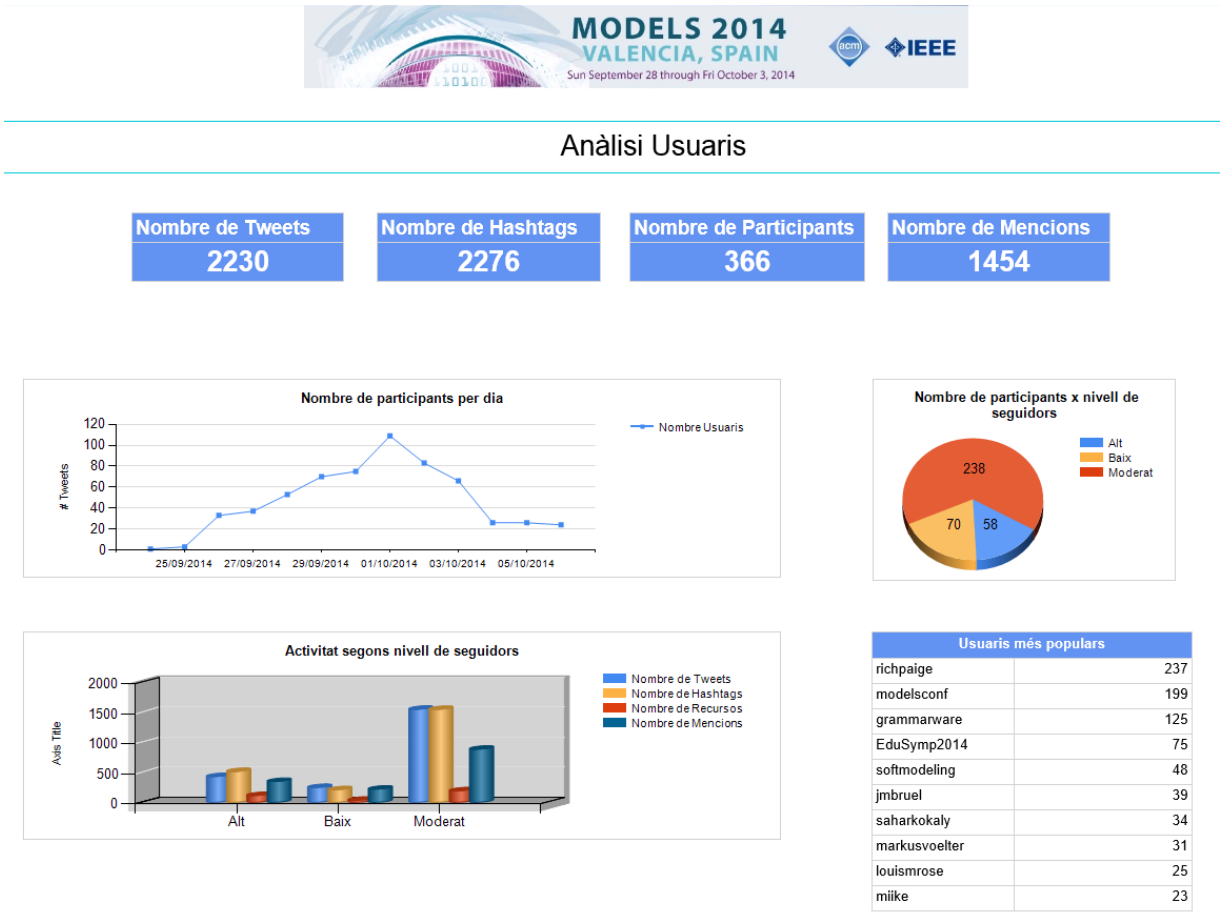
L'informe està format per cinc datasets :

DataSet_usuaris_diaris: Consulta MDX que retorna el nombre de usuaris diferents per dia.

DataSet1_usuaris_populars: Consulta MDX que retorna els usuaris que han rebut més mencions durant el congrés.

DataSet1_resumTotal: Consulta MDX que retorna els valors totals de cada una de les mètriques.

DataSet_tipus_usuari: Consulta MDX que mostra el el nombre d'usuaris segmentat per tipus.DataSet_metriques_tipus_usuari: Consulta MDX que mostra l'activitat dels usuaris segmentat per el nivell de seguidors.



Il·lustració 48: SSRS: Anàlisi de usuaris

3.6 Justificació del compliment del requisits

A continuació detallem el compliment dels requeriments no funcionals :

ID	**Requeriment**	**Justificació**
REQ-01	Analitzar el temes més parlats	En l'informe tweets.rdl es mostra els temes més parlats durant el congrés. El model analític construït permet explotar els temes per qualsevol eix del model.
REQ-02	Detectar usuaris més actius de l'esdeveniment.	Mitjançant el nostre model, podem explotar el nombre de tweets, recursos i temes per usuari.
REQ-03	Analitzar els recursos compartits en els tweets.	La dimensió recurs, ens permet analitzar tots els recursos compartits en l'esdeveniment a nivell més atòmic o segmentat per tipus de recurs.
REQ-04	Descobrir els hashtags coincidents en els tweets.	A través de la dimensió Hashtag podem esbrinar en quants tweets hi apareix.
REQ-05	Monitoritzar les activitats dels assistents.	En el informe usuaris.rdl, tenim un exemple de l'activitat dels assistents segmentat per tipus d'usuari.
REQ-06	Evolució del nombre de tweets al llarg del temps	En el informe tweets.rdl mostrem una gràfica on es veu l'evolució diària de tweets

A continuació detallem el compliment dels requeriments no funcionals :

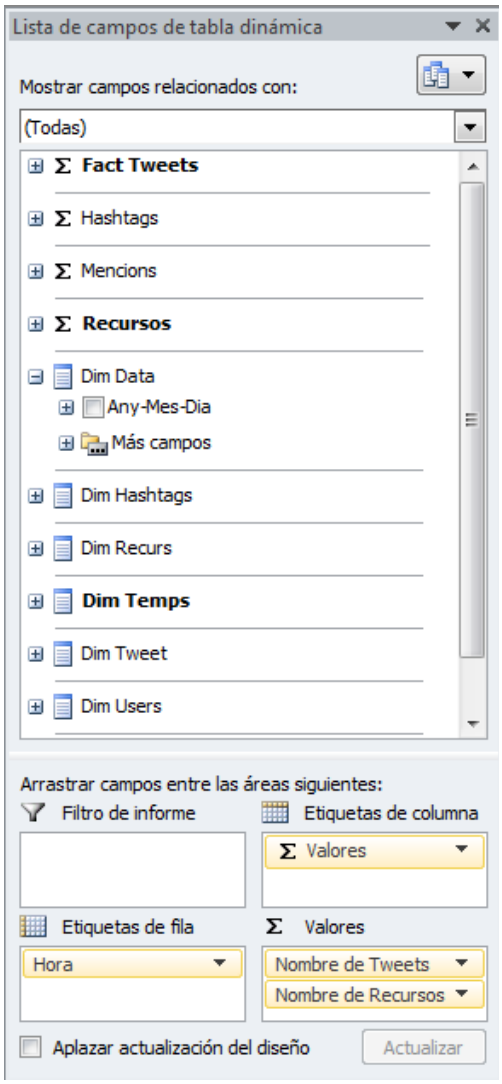
ID	Requeriment	Resolució
REQNF-1	L'accés al model analític i als seus informes ha	Tant l'accés al cub Congress

	d'estar restringit per els usuaris determinats..	com als informes de Rerporting Services es gestionen mitjançant autentificació Windows.
REQNF-2	La solució necessita poder ser explotada tant a través d'informes predefinits com amb eines d'anàlisis.	La visualització dels informes predefinits es realitza mitjançant un navegador i accedint al servidor d'informes. D'altra banda per l'anàlisi del cub es pot utilitzar Microsoft Excel per crear anàlisis a mida.
REQNF-3	Els informes no poden tardar més de 5 segons	Al treballar amb un sistema MOLAP on es té pre-calculats tots els agregats juntament amb que el volum de dades no és molt gran fa que el temps de resposta dels informes sigui molt òptim.
REQNF-4	Els usuaris han de saber quan es va realitzar l'última càrrega de les dades.	En l'informe del menú, es mostra la data de l'última actualització de les dades.

4 Captures de Pantalla

En l'apartat anterior ja hem vist algunes captures dels informes realitzats amb Reporting Services. Ara veurem unes altres captures de Microsoft Excel on es mostra les possiblitats del model analític construït.

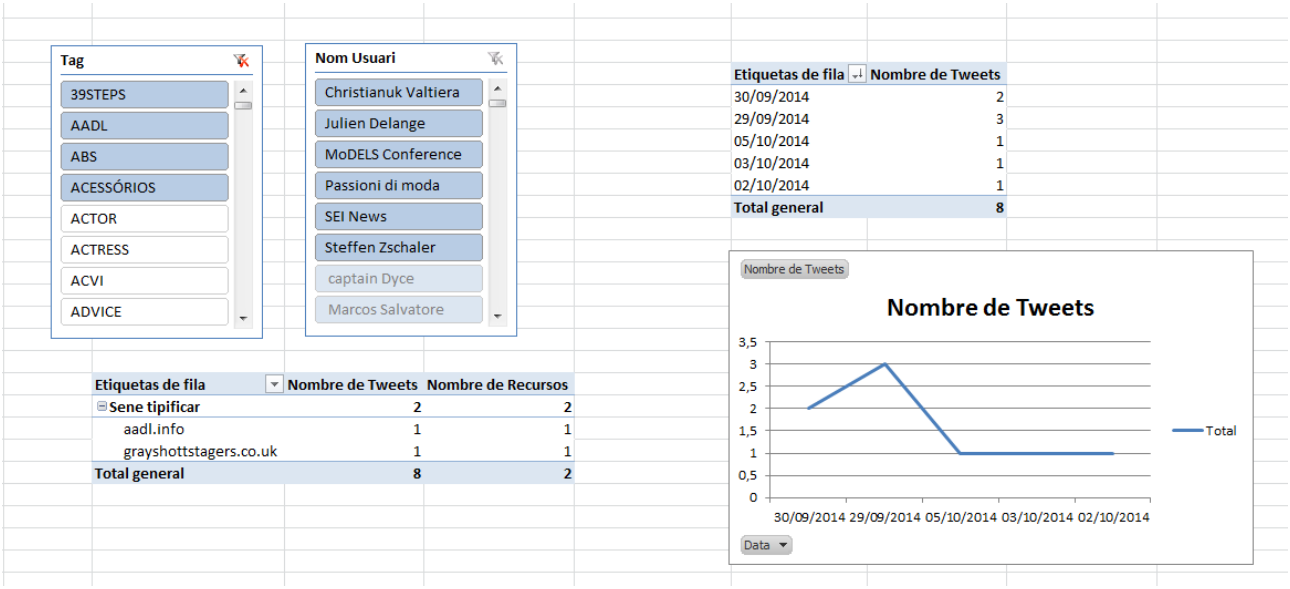
L'explotació de dades amb MS Excel es realitza a través de les taules dinàmiques. Es crea una connexió al cub i seguidament apareixen totes les mètriques i dimensions per ser utilitzades en la nostra taula dinàmica.



Il·lustració 49: MS Excel: Model Analític

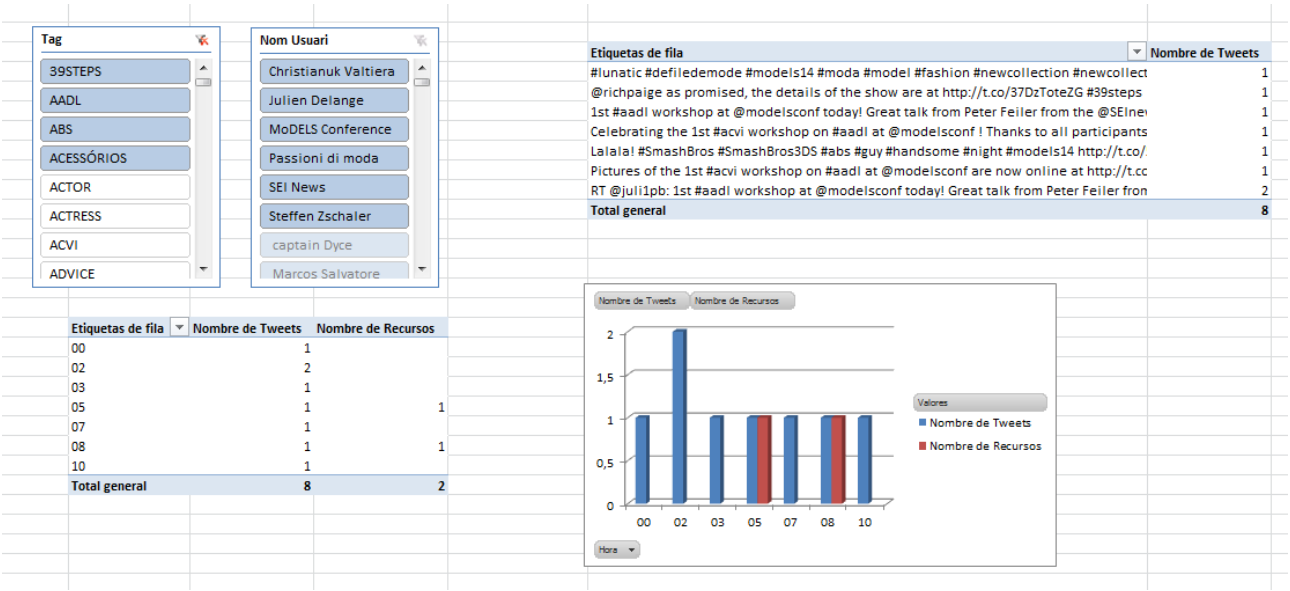
En la següent captura es mostra l'ús de segmentadors per filtrar la informació a diverses taules dinàmiques i gràfics dinàmics.

Com es pot observar a partir d'uns hashtags seleccionats, tenim la informació referent al nombre de tweets generats amb aquests tags i als seus recursos.



Il·lustració 50: MS Excel: Anàlisi de tweets I

En la següent captura podem veure un detall dels tweets generats a partir d'uns hashtags seleccionats.



Il·lustració 51: MS Excel: Anàlisi de tweets II

5 Conclusions

Podem concloure que els objectius del projecte han estat assolits. Hem vist el procés complet de creació d'un magatzem de dades i diferents formes d'explotar la seva informació.

A partir d'uns requeriments inicials, s'ha dissenyat una solució que ens ha permès desenvolupar uns informes determinats mostrant els indicadors més importants del nostre model.

S'ha realitzat una documentació exhaustiva del producte final, acompanyada amb una presentació virtual on s'ha pogut veure l'explicació i una demostració del treball realitzat.

Personalment ha estat una experiència molt enriquidora ja que no s'ha tractat únicament el desenvolupament d'un producte informàtic sinó que s'ha treballat aspectes com la planificació, el disseny i la presentació del producte final. En el món professional a banda de realitzar un bon producte a nivell tècnic és imprescindible la realització d'aquestes tasques per tenir èxit en la realització de projectes.

Com a aspiració personal, m'agradaria aprofundir amb temes de mineria de dades per descobrir patrons de comportament en la informació i així establir relacions que puguin ser útils i puguin generar noves oportunitats de negoci.

6 Línies d'evolució futures

Un dels aspectes més fluixos del nostre model és tota la relació referent als usuaris de tweeter. La informació no està normalitzada i ens aporta poc valor afegit.

Com a proposta es podria incorporar una base de dades dels assistents i ponents del esdeveniment, . Aquestes dades les podríem vincular amb el usuari de tweeter i d'aquesta manera poder segmentar per atributs més interessants com el perfil professional, l'empresa, etc....

Un altre apartat que podria ser interessant seria aplicar Social intelligence al nostre model. Mitjançant algoritmes podríem detectar si s'estan fent comentaris positius o negatius sobre l'esdeveniment o sobre els seus ponents, descobrir tendències o preveure el comportament del usuaris per identificar noves oportunitats de negoci.

Finalment seria interessant mostrar l'explotació de les dades mitjançant un client OLAP com

Microsoft Excel per tal de mostrar la potència del nostre model analític. Per motius tècnics de falta d'espai a la màquina virtual no hem pogut instal·lar Microsoft Excel per mostrar els anàlisis.

8 Bibliografia

Enllaços de interès:

- http://blog.oaktonsoftware.com/2011/04/bridge-tables-and-many-to-many.html
- http://datawarehouse4u.info/Data-warehouse-schema-architecture-star-schema.html

Tutorials de Microsoft SQL Server:

- http://msdn.microsoft.com/es-es/library/jj590844.aspx (SQL Server)
- http://msdn.microsoft.com/es-es/library/jj720568.aspx (Integration Services)
- http://msdn.microsoft.com/es-es/library/hh231701.aspx (Analysis Services)
- http://msdn.microsoft.com/es-es/library/bb522859.aspx (Reporting Services)