Universitat Oberta de Catalunya
www.uoc.edu

PROJECTE "ACA"

MEMÒRIA

TFC Treball Fi de Carrera (Magatzem de dades)

Estudiant Enginyeria Tècnica Informàtica de Gestió
Ismael Santacruz Vera
Consultor
José Ángel Martín Carballo
Juny 2011

Agraïments infinits a la meva dona per tot l'ajut, comprensió i paciència que m'ha ofert en tot moment i per totes les hores que s'han perdut pel camí que a partir d'ara s'hauran de recuperar.

1. Resum del Treball Fi de Carrera

La incorporació de les noves tecnologies i la creixent necessitat d'oferir als clients el major servei, ha conduit a les empreses a realitzar importants millores en els seus sistemes informàtics, provocant l'increment del número de transaccions i en conseqüència, augmentant significativament el seu volum de dades.

Tot i la riquesa d'aquestes dades i la possibilitat d'anàlisi que ens podrien oferir, no resulta senzilla l'extracció i posterior estudi d'aquestes mitjançant els mètodes tradicionals.

Els magatzems de dades en forma d'eines BI amb tecnologies OLAP estan pensats per dissenyar i treure el màxim profit i rendiment de la seva base de dades.

Aquestes eines estan basades en interfícies gràfiques que ens possibiliten un anàlisi d'informació mitjançant implosió i explosió de les dades (drill-up/drill-down).

D'aquesta manera resulta senzill el fet de convertir informació transaccional fruit del treball operatiu de cada dia amb informació agregada i ben classificada significativa de negoci a fi i efecte de prendre decisions.

En aquest treball tractarem de posar en pràctica tota aquesta estratègia de transformació de dades. L'escenari que tractarem serà l'Acadèmia de Cinema Andorrà (ACA). L'ACA disposa d'un conjunt de fulls Excels amb els premis i nominacions obtingudes per diferents pel·lícules, actors, directors, certàmens, festivals, etc...

Davant la dificultat en la que es troben a l'hora d'analitzar la informació en aquest format d'Excel, intentarem crear un magatzem de dades que no només ens permeti emmagatzemar la informació de manera consistent i estandarditzada dels diferents premis, sinó també consolidar-la i emmarcar-la en un únic context homogeni.

Paraules clau

DataWarehouse, OLAP, BI, Business Intelligence, Magatzem de dades, Model Relacional, Model Multidimensional, Oracle DataBase Express Edition, Oracle Discoverer Administration, Oracle Discoverer Desktop, Dimensions, Mètriques, ETL, PL/SQL.

Àrea del TFC (Treball Fi de Carrera)

Magatzem de dades

Índex

1.Resum del Treball Fi de Carrera....3
2.Index de figures....6
3.Introducció....7
3.1. Justificació del TFC i context en el qual es desenvolupa. Punt de partida i aportació al TFC....7
3.2. Objectius del projecte....7
3.3. Enfocament i mètode seguit....7
3.4. Planificació del projecte....8
3.5. Productes obtinguts....13
3.6. Breu descripció dels altres capítols de la memòria....13
4.Anàlisi....13
4.1. Definició de documents d'entrada....13
4.2. Anàlisi de documents de sortida....14
4.3. Diagrama de casos d'ús....15
4.4. Diagrama del model conceptual....16
4.4.1. Dissenyar una estrella....16
4.4.2. Triar el fet....17
4.4.3. Trobar el grànul escaient....17
4.4.4. Escollir les dimensions i mesures....18
5.Disseny....19
5.1. Fase I Model temporal seqüencial....19
5.2. Fase II Model Relacional....19
5.3. Fase III Model Multidimensional....22
5.4. Depuració i Normalització de les dades....23
5.5. Evolució gràfica de la transformació de les dades....24
5.6. Disseny i descripció dels informes creats....25
5.6.1. Informe 1....25
5.6.2. Informe 2....28
5.6.3. Informe 3....29
5.6.4. Informe 4....31
5.6.5. Informe 5....33
5.6.6. Informe 6....35
5.6.7. Informe 7....37
5.6.8. Informe 8....39
6.Conclusions....41
7.Línies d'evolució futur....42
8.Glossari....42
9.Bibliografia....43

10.Annexos 44
10.1. Annex 1. Credencials d'accés als diferents sistemes 44
10.2. Annex 2. Ubicació de fitxers 45
10.3. Annex 3. Procés ETL 45

2. Índex de figures

1. Figura 1. Planificació diagrama Gantt Overview projecte 10
2. Figura 2. Planificació diagrama Gantt Pac1 11
3. Figura 3. Planificació diagrama Gantt Pac2 11
4. Figura 4. Planificació diagrama Gantt Pac3 12
5. Figura 5. Planificació diagrama Gantt Final 12
6. Figura 6. Casos d'ús 1 15
7. Figura 7. Casos d'ús 2 16
8. Figura 8. BBDD Temporal 19
9. Figura 9. BBDD Relacional 22
10. Figura 10. BBDD Multidimensional 23
11. Figura 11. Gràfic de transformació de dades 24
12. Figura 12. Gràfic d'Arquitectura de Hardware i Software 25
13. Figura 13. Imatges d'Informe 1 26
14. Figura 14. Imatges d'Informe 2 29
15. Figura 15. Imatges d'Informe 3 30
16. Figura 16. Imatges d'Informe 4 33
17. Figura 17. Imatges d'informe 5 35
18. Figura 18. Imatges d'informe 6 36
19. Figura 19. Imatges d'informe 7 40
20. Figura 20. Imatges d'informe 8 41
21. Figura 21. Credencials d'Oracle Database Express Edition 45
22. Figura 22. Credencials d'Oracle Discoverer Administration 45
23. Figura 23. Credencials d'Oracle Discoverer Desktop 46
24. Figura 24. ETL 47
25. Figura 25. Scheduler 48

3. Introducció

3.1. Justificació del TFC i context en el qual es desenvolupa. Punt de partida i aportació al TFC

Durant més de 10 anys la meva vida laboral ha estat lligada al món de la programació i de la informàtica de gestió. No cal parlar de l'evolució que ha experimentat aquesta disciplina durant aquests temps; en els meus orígens realitzant programari per a terminals caràcter, fins avui realitzant projectes de programari web accessible des de qualsevol part del món i fins i tot amb dispositius mòbils que caben dins d'una butxaca.

Doncs bé, després de tota una carrera professional immers en el desenvolupament de software i eines per tal d'introduir i desar dades en una base de dades per tal de gestionar qualsevol tipus de negoci o activitat, he cregut convenient aprendre i endinsar-me en aquest món emergent del Business Intelligent i dels magatzems de dades.

Finalment, el fet que en la meva vida professional cada cop hi hagi més inclusions de projectes d'aquest tipus, la coincidència del meu Projecte de Fi de Carrera i la possibilitat de poder realitzar-lo en l'àrea dels magatzems de dades, ha estat un fet clau per voler aprendre el màxim d'aquesta disciplina i que sigui un punt de partida formatiu important per poder afrontar-ho professionalment amb el màxim de garanties.

3.2. Objectius del projecte

L'objectiu principal del projecte és adquirir experiència en el disseny, construcció i explotació d'un magatzem de dades a partir de la informació disponible en una base de dades transaccional.

A banda, també assolirem coneixements sobre:

- Com són i com funcionen els magatzems de dades.
- Conèixer programari de BI.
- Conèixer productes de Base de Dades Oracle.
- Aprendre a modelar dades originals i transformar-les al model final del magatzem mitjançant eines de transformació ETL.
- Aprofundir amb els llenguatges de programació de consulta SQL.

3.3. Enfocament i mètode seguit

Tot i que en cap moment és un projecte en el qual s'hagi de desenvolupar cap tipus de programari, el projecte el portarem a terme seguint una mena de cicle de vida clàssic, també anomenat cicle de vida en cascada.

En cadascuna de les etapes d'aquest mètode obtindrem uns documents que serviran de base per l'etapa següent. L'elecció dels tipus de document es farà en cadascuna de les etapes segons es cregui convenient amb la premissa d'utilitzar sempre la riquesa dels diagrames de l'estàndard UML.

El projecte que ens ocupa comporta una particularitat especial i resideix en el fet que els objectius finals són realitzar uns informes amb una eina que no coneixem. Per aquest motiu el plantejament serà combinar la fase d'anàlisi i la fase de disseny amb

l'aprenentatge d'aquesta eina, per tal de poder arribar a la fase de l'elaboració d'informes amb els coneixements suficients per tal de dur-la a terme. A banda el fet de conèixer l'eina durant tot el procés ens ajudarà a dissenyar i analitzar certs aspectes.

3.4. Planificació del projecte

La planificació de cadascuna de les fases ha estat condicionada a les dates dels diferents lliuraments assenyalades en el Pla Docent de l'assignatura. D'aquesta manera cadascuna de les fites del projecte es correspon amb una data de lliurament:

Titol	Enunciat	Lliurament
PAC1	05/03/2011	16/03/2011
PAC2	17/03/2011	20/04/2011
PAC3	21/04/2011	25/05/2011
Lliurament Final	26/05/2011	13/06/2011

16/03/2011 – Lliurament PAC1

Confecció i lliurament del Pla de Treball amb la descripció del projecte, objectius i planificació.
Aquest lliurament disposarà d'una primera fase d'anàlisi molt incipient i preliminar dels indicadors i dimensions que utilitzarem. També disposarà d'alguns esborranys d'informes que mostraran la informació.

20/04/2011 – Lliurament PAC2

Aquest lliurament constarà de la documentació amb l'anàlisi detallat de requeriments basat en l'anàlisi preliminar realitzat.
També hi constarà tot el model dimensional de dades del magatzem que haurà de donar suport a tot l'assortit d'informes.
Hi farem constar també tota la documentació orgànica referent a la transformació de dades des d'Excel fins al Magatzem de dades final mitjançant l'eina ETL.

25/05/2011 – Lliurament PAC3

Aquest lliurament està catalogat com a control de seguiment i es situa en la part avançada de la fase d'implementació.

13/06/2011 – Lliurament final

Lliurament de la Memòria del projecte.
Lliurament del programari i manual d'utilització.
Lliurament de la presentació del projecte.

Tasca	Dates	Hores
Abans de començar el TFC	05/03/2011>16/03/2011	
Documentar-se sobre les diverses àrees de TFC		0.5
Apuntar-se a la sol·licitud i acceptació de l'opció triada		0.5
Llegir-se el Pla Docent de l'assignatura		0.5

Un cop començat TFC (previ a la PAC1)	05/03/2011>16/03/2011	
Descarregar documentació inicial de l'assignatura		0.5
Descarregar software de l'assignatura (VirtualBox)		4
Preparar altre software administratiu (Visio, Word, Project, etc...)		1
Familiaritzar-se amb software de Base de Dades Oracle		1
Familiaritzar-se amb software BI Discoverer Oracle		1
Total		
Lliurament PAC1	05/03/2011>16/03/2011	
Recerca d'informació sobre Business Intelligence i magatzems de dades.		1
Elaboració del pla de treball (objectius, tasques, material, plans de contingència imprevistos) per tal d'arribar a totes les fites de manera satisfactòria.		10
Encaixar en un planning / calendari totes les tasques planificades.		5
Total		25
Lliurament PAC2	16/03/2011>20/04/2011	
Anàlisi exhaustiu de tots els requeriments de cadascun dels informes sol·licitats.		10
Anàlisi exhaustiu de tots els documents originals d'entrada (Excels facilitats per l'ACA).		10
Anàlisi i disseny del model de dades d'un magatzem de dades capaç de donar suport als requeriments sol·licitats.		35
Recerca i familiarització amb un producte de transformació i càrrega de dades ETL.		10
Anàlisi i disseny de processos ETL per tal de transformar la informació original d'Excel al magatzem de dades final.		10
Total		**75**
Lliurament PAC3	21/04/2011>25/05/2011	
Construcció i instal·lació del magatzem de dades.		10
Instal·lació de l'eina BI d'explotació de les dades.		10
Programació dels informes prèviament analitzats.		40
Validació i tests finals.		6
Total		**66**
Lliurament FINAL	26/05/2011>13/06/2011	
Recull de tota la documentació realitzada i confecció de la memòria.		30
Revisió de la memòria (ortografia, redacció, continguts, etc...).		2
Síntesi de la memòria i extracció d'idees fonamentals.		2
Realització d'una presentació PowerPoint d'unes 20 diapositives.		6
Total		**40**

A continuació es mostren els diagrames de Gantt corresponents al TFC.

Diagrama de Gantt General. Overview general de tot el projecte. Des de 05/03/2011 fins al 13/06/2011.

Nombre de tarea	Comienzo	Fin
Abans de començar el TFC	**vie 04/03/11**	**dom 06/03/11**
Documentar-se sobre les diverses àrees de TFC	vie 04/03/11	dom 06/03/11
Apuntar-se a la sol•licitud i acceptació de la opció tri	vie 04/03/11	dom 06/03/11
Llegir-se el Pla Docent de la assignatura	vie 04/03/11	dom 06/03/11
Un cop començat TFC (previ a la PAC1)	**vie 04/03/11**	**lun 07/03/11**
Descarregar documentació inicial de l'assignatura	vie 04/03/11	lun 07/03/11
Descarregar software de l'assignatura (VirtualBox)	vie 04/03/11	lun 07/03/11
Preparar altre software administratiu (Visio, Word, P	vie 04/03/11	lun 07/03/11
Familiaritzar-se amb software de Base de Dades Or	vie 04/03/11	lun 07/03/11
Familiaritzar-se amb software BI Discoverer Oracle	vie 04/03/11	lun 07/03/11
Total	vie 04/03/11	lun 07/03/11
Lliurament PAC1	**vie 04/03/11**	**mié 16/03/11**
Recerca d'informació sobre Business Intelligence i n	vie 04/03/11	jue 10/03/11
Elaboració del pla de treball (objectius, tasques, mat	vie 04/03/11	jue 10/03/11
Encaixar en un planning / calendari totes les tasques	jue 10/03/11	mié 16/03/11
Total	mié 16/03/11	mié 16/03/11
Lliurament PAC2	**mié 16/03/11**	**mié 20/04/11**
Anàlisi exhaustiu de tots els requeriments de cadasc	mié 16/03/11	mié 23/03/11
Anàlisi exhaustiu de tots els documents originals d'e	mié 23/03/11	mar 29/03/11
Anàlisi i disseny del model de dades d'un magatzen	mar 29/03/11	lun 11/04/11
Recerca i familiarització amb un producte de transfo	mar 12/04/11	vie 15/04/11
Anàlisi i disseny de processos ETL per tal de transfo	lun 18/04/11	mié 20/04/11
Total	mié 20/04/11	mié 20/04/11
Lliurament PAC3	**jue 21/04/11**	**jue 26/05/11**
Construcció i instal•lació del magatzem de dades.	jue 21/04/11	mié 27/04/11
Instal•lació de l'eina BI d'explotació de les dades.	jue 28/04/11	mié 04/05/11
Programació dels informes prèviament analitzats.	jue 05/05/11	lun 23/05/11
Validació i tests finals.	mar 24/05/11	mié 25/05/11
Total	jue 26/05/11	jue 26/05/11
Lliurament FINAL	**jue 26/05/11**	**lun 13/06/11**
Recull de tota la documentació realitzada i confecció	jue 26/05/11	vie 03/06/11
Revisió de la memòria (ortografia, redacció, contingu	lun 06/06/11	lun 06/06/11
Síntesi de la memòria i extracció d'idees fundamenta	mar 07/06/11	mar 07/06/11
Realització d'una presentació PowerPoint d'unes 20	mié 08/06/11	lun 13/06/11
Total	lun 13/06/11	lun 13/06/11

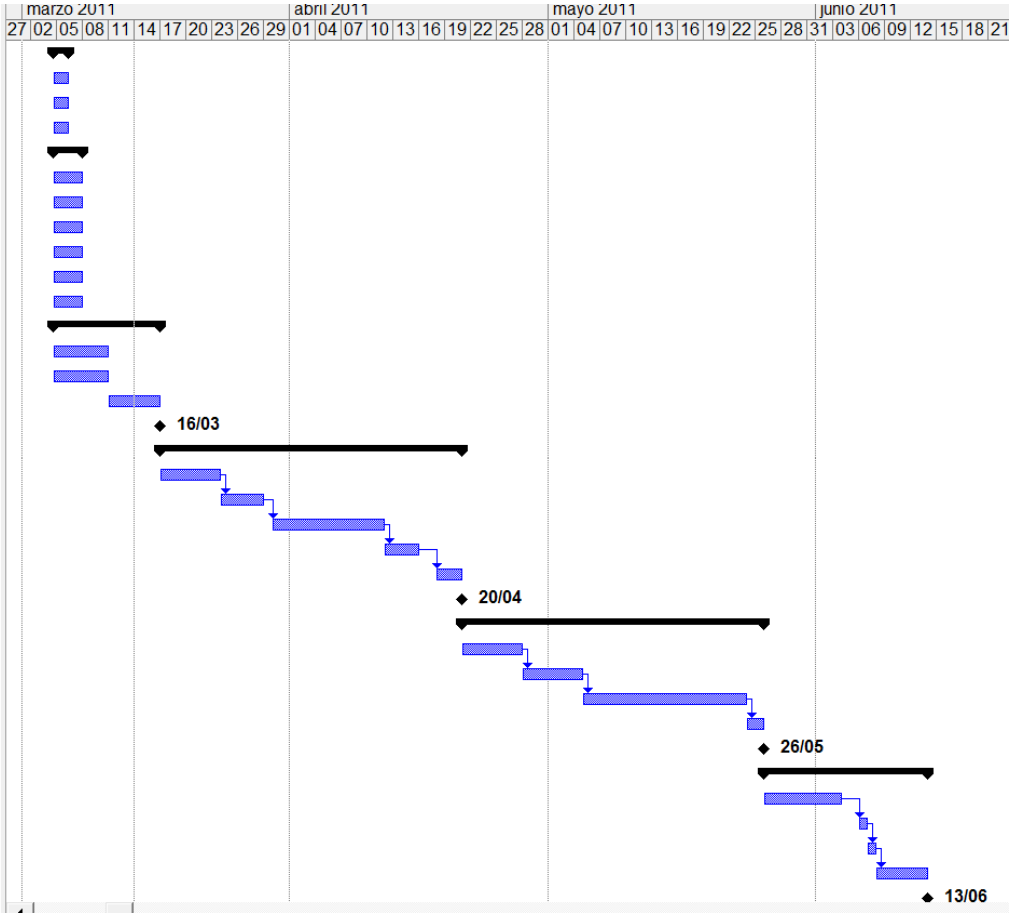


Diagrama de Gantt del previ TFC fins la PAC1

Abans de començar el TFC	**vie 04/03/11**	**dom 06/03/11**
Documentar-se sobre les diverses àrees de TFC	vie 04/03/11	dom 06/03/11
Apuntar-se a la sol•licitud i acceptació de la opció tri	vie 04/03/11	dom 06/03/11
Llegir-se el Pla Docent de la assignatura	vie 04/03/11	dom 06/03/11
Un cop començat TFC (previ a la PAC1)	**vie 04/03/11**	**lun 07/03/11**
Descarregar documentació inicial de l'assignatura	vie 04/03/11	lun 07/03/11
Descarregar software de l'assignatura (VirtualBox)	vie 04/03/11	lun 07/03/11
Preparar altre software administratiu (Visio, Word, P	vie 04/03/11	lun 07/03/11
Familiaritzar-se amb software de Base de Dades Or	vie 04/03/11	lun 07/03/11
Familiaritzar-se amb software BI Discoverer Oracle	vie 04/03/11	lun 07/03/11
Total	vie 04/03/11	lun 07/03/11
Lliurament PAC1	**vie 04/03/11**	**mié 16/03/11**
Recerca d'informació sobre Business Intelligence i n	vie 04/03/11	jue 10/03/11
Elaboració del pla de treball (objectius, tasques, mat	vie 04/03/11	jue 10/03/11
Encaixar en un planning / calendari totes les tasques	jue 10/03/11	mié 16/03/11
Total	mié 16/03/11	mié 16/03/11

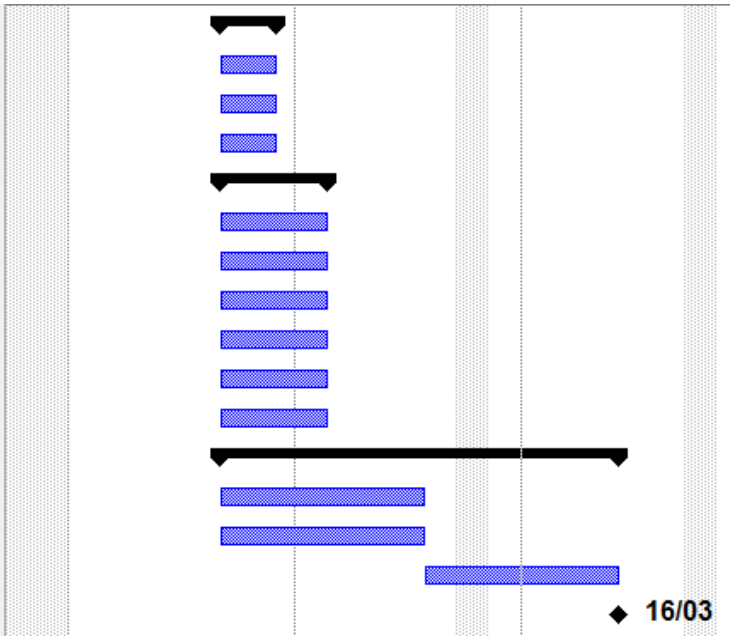


Diagrama de Gantt de la PAC2

Lliurament PAC2	**mié 16/03/11**	**mié 20/04/11**
Anàlisi exhaustiu de tots els requeriments de cadasc	mié 16/03/11	mié 23/03/11
Anàlisi exhaustiu de tots els documents originals d'e	mié 23/03/11	mar 29/03/11
Anàlisi i disseny del model de dades d'un magatzen	mar 29/03/11	lun 11/04/11
Recerca i familiarització amb un producte de transfo	mar 12/04/11	vie 15/04/11
Anàlisi i disseny de processos ETL per tal de transfo	lun 18/04/11	mié 20/04/11
Total	mié 20/04/11	mié 20/04/11

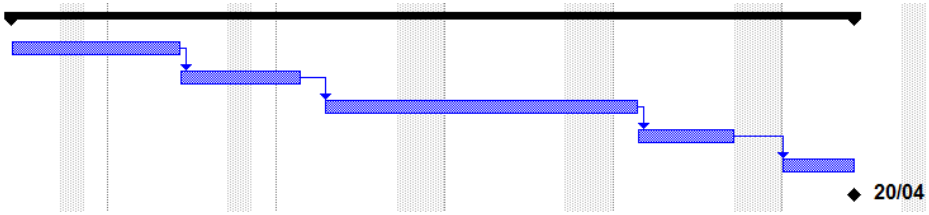


Diagrama de Gantt de la PAC3

	Lliurament PAC3	jue 21/04/11	jue 26/05/11
	Construcció i instal•lació del magatzem de dades.	jue 21/04/11	mié 27/04/11
	Instal•lació de l'eina BI d'explotació de les dades.	jue 28/04/11	mié 04/05/11
	Programació dels informes prèviament analitzats.	jue 05/05/11	lun 23/05/11
	Validació i tests finals.	mar 24/05/11	mié 25/05/11
	Total	jue 26/05/11	jue 26/05/11

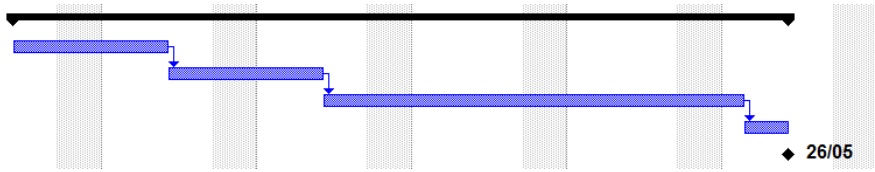
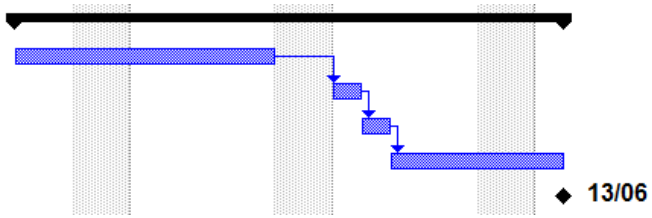


Diagrama de Gantt del lliurament FINAL

	Lliurament FINAL	jue 26/05/11	lun 13/06/11
	Recull de tota la documentació realitzada i confecció	jue 26/05/11	vie 03/06/11
	Revisió de la memòria (ortografia, redacció, contingu	lun 06/06/11	lun 06/06/11
	Síntesi de la memòria i extracció d'idees fundamenta	mar 07/06/11	mar 07/06/11
	Realització d'una presentació PowerPoint d'unes 20	mié 08/06/11	lun 13/06/11
	Total	lun 13/06/11	lun 13/06/11



3.5. Productes obtinguts

A continuació es detalla tot un seguit material, programari i documentació ha estat tot el que s'ha creat a partir de la realització d'aquest projecte.

- Document explicatiu de la planificació del projecte (PAC1).
- Document explicatiu de l'anàlisi i el disseny del projecte (PAC2).
- Document explicatiu de la implementació del projecte (PAC3).
- Un disc virtual VirtualBox que conté:
    - Les bases de dades en Oracle DataBase Express Edition.
    - Oracle Discoverer Administration.
    - Oracle Discoverer Desktop.
- Fitxers originals amb format Excel. Ubicats en la màquina virtual.
- Fitxers transformats amb format csv. Ubicats en la màquina virtual.
- Memòria final del projecte (TFC).
- Presentació virtual en format de Vídeo.

3.6. Breu descripció dels altres capítols de la memòria

A continuació en aquesta memòria aprofundirem en altres aspectes i fases del projecte.
Un dedicat a l'anàlisi previ dels requeriments i l'altre al disseny final que hem adoptat per tal de portar a terme la sol·lució final.

Posteriorment veurem la implementació final i com hem resolt mitjançant informes tots els requeriments sol·licitats.

Per acabar, tindrem un capítol de conclusió i de possibles millores de futur que podríem implementar a la sol·lució actual.

4. Anàlisi

En aquesta fase d'anàlisi, tractarem d'analitzar tots els documents d'entrada (Excels) per tal de veure quines opcions i problemes ens ofereixen.

Finalment, per tal de poder complir els requeriments finals amb les dades dels documents originals, s'haurà de confeccionar tota l'anàlisi dimensional de la base de dades definitiva, així com definir tot el procés i estratègia de transformació de les dades per obtenir al final els documents de sortida definitius.

4.1. Definició de documents d'entrada

Com a documents d'entrada, l'ACA ens ha proporcionat 5 llibres Excel. Un Excel per cada festival on ens apareixen totes les dades transaccionals de tots els festivals amb totes les pel·lícules que han entrat a concurs.

A grans trets les dades que hi surten representades són les següents:

Nom del Camp	Descripció i comentari
Original title	Títol Original de la pel·lícula.

Localized title	Títol local de la pel·lícula.
Sort title	Títol d'ordenació de la pel·lícula.
Year	Any de celebració del certamen.
Country	País origen de la pel·lícula.
Award	Nom del certamen o festival.
Category	Categoria del premi.
Winner	Guanyador.
Nominee	Nom dels nominats.
Edition	Edició del certamen.

Després d'analitzar aquests documents Excels i després d'estudiar els informes finals que volem obtenir, ens adonem que les diferents fonts de dades (Excels) ens poden donar deficiències generals en les dades.

Algunes d'aquestes deficiències són:

- Entitats no relacionades
- Entitats sense restriccions d'integritat
- Valors inconsistents
- Redundància de dades
- Incoherència de dades
- Acumulació de dades sense indexar
- Alguns camps sense informar
- Estructura tabular/seqüencial davant de la Multidimensional de un DW
- Estructura no orientada a l'explotació de dades
- Estructura orientada bàsicament a l'acumulació transaccional de dades

A banda d'aquestes deficiències estructurals, ens trobem amb algunes característiques de les dades que haurem de tenir en compte a l'hora d'homogeneïtzar-les i estandarditzar-les mitjançant la transformació.

Nom del Camp	Característica de les dades
Year	Els anys els podem trobar amb format yyyy o yyyy1-yyyy2. En el segon cas s'agafarà com a any yyyy2.
Country	El país si no està informat o s'informa amb un 0 es tractarà d'una pel·lícula dels Estats Units.
Winner	En el camp Winner, si hi apareix una X és que s'ha guanyat el premi de la categoria. En altre cas, no.
Nominee	Nom dels nominats (només té sentit per a persones).

4.2. Anàlisi de documents de sortida

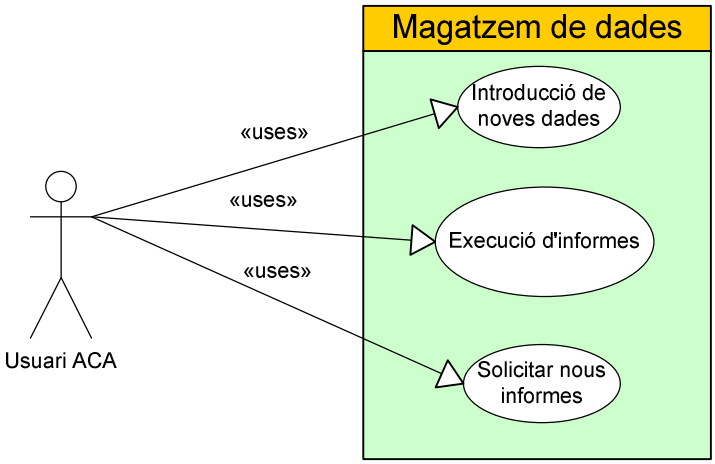
Els informes que el nostre banc de dades haurà de fer front a partir dels documents Excels són els següents:

Informe	Dimensions
Nombre de premis aconseguits per un actor/actriu per any i festival. L'informe ha d'estar totalitzat per saber el nombre total de premis.	Entitat, certamen, any i resultat
Nombre de premis aconseguits per un actor/actriu per any, festival i categoria.	Entitat,certamen, any, categoria i

	resultat
Nombre de premis aconseguits per un director per any, festival i pel·lícula L'informe ha d'estar totalitzat per saber el nombre total de premis.	Entitat,certamen, any, pel·lícula i resultat
Nombre de premis aconseguits per any, festival, categoria (s'ha de poder triar si es vol veure l'estàndard o la pròpia) i pel·lícula.	Certamen, categoria, any, pel·lícula i resultat
Donada una data a seleccionar per l'ACA, per cada tipus de festival mostrarà el nombre d'edicions realitzades del festival, nombre total d'actors que han estat nominats, nombre total d'actrius que han estat nominades i nombre total de pel·lícules nominades.	Certamen, entitats, pel·lícules
Per a cada any, nombre de premis guanyats per cada pel·lícula i festival (i subtotal per any/pel·lícula).	Any, categoria, pel·lícula, certamen i resultat
Addicionalment hem de proporcionar una definició de "pel·lícula exitosa" i "pel·lícula fracassada". A partir d'aquesta definició, i per a cada any, hem de llistar la pel·lícula més exitosa i més fracassada.	Any, pel·lícula i resultat
Finalment, també ens demanen que expliquem si hi ha una relació entre guanyar un premi a un festival i guanyar-ne el mateix a un altre. Detallar també quins criteris s'han seguit per arribar al resultat donat.	Certamen, categoria i resultats

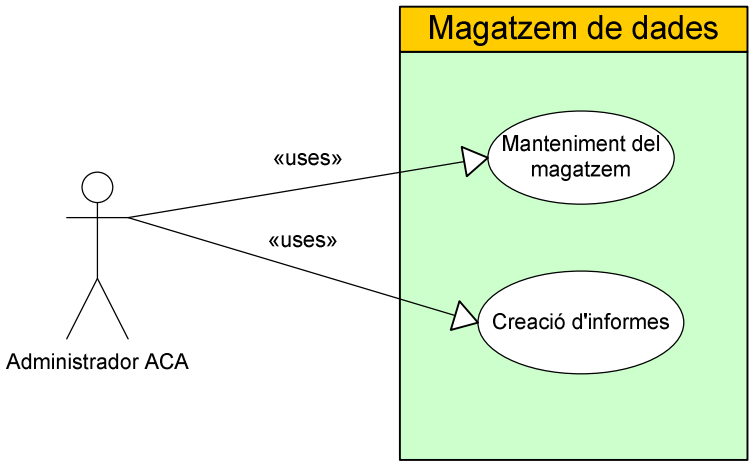
4.3. Diagrama de casos d'ús

El següent diagrama de casos d'ús representa els usos dels diferents usuaris del magatzem de dades i les funcions que poden realitzar.



Bàsicament, els usuaris ACA són els encarregats d'introduir les dades en la base de dades relacional que posteriorment aniran al magatzem de dades.
A banda, són els que executen els actuals informes per tal d'explotar les dades i en cas necessari, s'encarregarien de demanar nous informes segons les necessitats als administradors del magatzem.

D'altra banda, tenim a l'administrador/s del magatzem de dades.
L'administrador és l'encarregat del manteniment general del magatzem de dades. Que els processos de càrrega de dades funcionin diàriament sense cap problema, que les dades siguin consistents, etc...
També és l'encarregat de gestionar els usuaris, realitzar còpies de seguretat, encarregar-se de que el rendiment del sistema sigui òptim, etc...
Finalment, i en cas que sigui necessari, també és l'encarregat de realitzar nous informes si així els usuaris del magatzem ho demanen.



4.4. Diagrama del model conceptual

Per tal de poder donar sortida a tots els requeriments sol·licitats per l'ACA i al mateix temps esmenar aquestes deficiències generals de com emmagatzenem les dades en l'actualitat, pensem que seria bo el fet de generar un magatzem de dades que sigui capaç de donar solució tant a les necessitats actuals com les que es puguin plantejar en un futur.

Bàsicament, un magatzem de dades és una estratègia tecnològica que ens permet una còmoda anàlisi de les dades. Essencialment, pot definir-se com una col·lecció de dades classificades per temes, integrada, variable en el temps i no volàtil que s'utilitza com ajuda al procés de la presa de decisions per part de qui ha de dirigir les organitzacions.

Després d'estudiar les dades d'entrada decidim que el més adient és generar una base de dades multidimensional basada en un disseny en estrella.
Aquest disseny es basa en tenir una taula de fets que conté totes les dades necessàries per a dur terme les anàlisis rodejada de taules de totes les dimensions que contenen la informació complementària dels registres d'aquesta taula de fets.

Aquest disseny en estrella amb la taula de fets i els diferents fitxers satèl·lits oferint la informació relacionada ens oferirà una simplicitat i un rendiment molt adequat pels requeriments als quals hem de donar suport.

4.4.1. Dissenyar una estrella

Un disseny en estrella és aquell que disposa d'una taula central. Aquesta taula normalment és anomenada taula de fets. Conté les dades de l'anàlisi i es troba rodejada de totes les taules que marquen les dimensions. En aquest aspecte de taula

central rodejada de radis que s'uneixen amb les taules més petites és el que fa que aquest disseny s'assembli a una estrella.

Aquest esquema resulta ideal per la seva simplicitat i velocitat a l'hora de ser usat. Permet accedir tant a dades agregades com a dades de detall facilitant el drill-down i el drill-across, a més també permet reduir el número de joins entre les taules i permet establir diversos nivells entre les diferents dimensions.

Finalment, aquest disseny afavoreix molt el bon rendiment ja que permet indexar les dimensions de forma individualitzada sense que repercuteixi en el rendiment de la base de dades com a conjunt.

4.4.2. Triar el fet

A grans trets, la taula de fet ha de contenir els valors pels quals volem analitzar les nostres dades.
Basant-nos en aquesta senzilla definició, seleccionarem els camps que ens serviran com a base per tal de classificar les mètriques i que seran els que utilitzarem en els informes, els documents i les diferents anàlisis.

En el nostre cas, la taula de fet quedaria conceptualment de la següent manera:

Taula de fets
Pais
Pelicula
Resultat
Entitat
Categoria
Certamen
EdicioCertamen
Nominacions

4.4.3. Trobar el grànul escaient

El grànul és l'individu últim que volem analitzar, la cel·la més petita que volem tenir disponible.

A partir d'aquest moment, ja no parlem de temes abstractes, ja ens hem de definir per la unitat mínima de registre que serà objecte d'estudi.

Normalment el grànul correspon a la transacció diària del registre de la taula de fets.

- Línies de factura si la taula de fets fos una taula de vendes.
- Moviments de magatzem si la taula de fets fos una taula d'stocks.

En el nostre cas, el grànul seran les pròpies nominacions. Cada registre de nominació serà la unitat mínima d'estudi.

Ser acurat a l'hora de seleccionar la grandària d'aquest grànul pot marcar definitivament l'èxit de les dades que pot oferir la taula de fets. Si el grànul és molt petit, pot alentir molt els càlculs d'acumulats dels seus nivells superiors, i si és molt gran pot no oferir tota la informació desitjada.

4.4.4. Escollir les dimensions i mesures

Per escollir correctament les dimensions, resulta molt important haver seleccionat correctament tant la taula de fets com el nivell de granulitat. Si això és així, trobar el conjunt de dimensions d'anàlisi resulta immediat a partir de la mateixa definició de la taula de fets.

En el nostre cas les dimensions seran,

Dimensions
Pais
Pelicula
Resultat
Entitat
Categoria
Certamen
EdicioCertamen
Nominacions

Per a seleccionar les mesures o mètriques normalment ens quedem amb tots aquells imports o números que són motiu d'estudi/anàlisi, imports (bruts, nets, marges), unitats, stocks, etc.... En el nostre cas serà el comptador de registres de les pròpies transaccions de la taula de fets.

Aquesta base de dades ens proporcionarà els avantatges típics dels models multidimensionals expressat en cubs OLAP.

Aquestes característiques són:

- Consolidació: Implica l'agregació de les dades, tals com totalitzacions simples o expressions complexes que impliquen dades relacionals.
- Profundització: Operació inversa a la consolidació. Implica mostrar les dades de detall que es troben en les dades consolidades.
- Navegació (pivotatges): Capacitat d'examinar les dades des de diferents punts de vista, amb la finalitat d'analitzar tendències i trobar patrons de comportament.

5. Disseny

Per tal de convertir les nostres dades inicials compostes de 5 Excels amb un magatzem de dades amb aquesta estructura d'estrella dissenyarem una estratègia de transformació ben diferenciada per tres fases.

5.1. Fase I Model temporal seqüencial

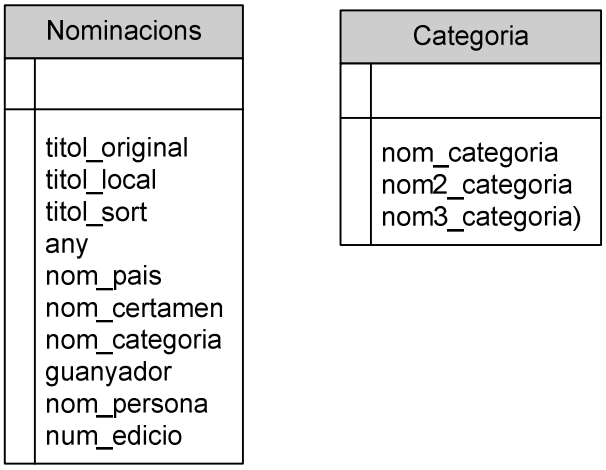
Traspàs de les dades de Excel a una base de dades temporal.

Per tal de disposar d'un entorn heterogeni, més íntegre, més consistent i més estàndard, el primer que farem serà migrar tots els Excels a una base de dades seqüencial.

En aquesta base de dades temporal crearem dues taules sense cap tipus d'índex i sense cap tipus de restricció d'integritat. Una que contindrà les dades generals, i una altra que contindrà les diferents nomenclatures que fan servir els diferents certàmens a l'hora de designar les categories dels premis (diferents nomenclatures, diferents idiomes, diferents tipus de premis, etc....)

Això farà que puguem migrar totes les dades contingudes en els Excels tot i les mancances que aquestes ens oferien. Un cop disposem de totes les dades migrades en aquestes taules, aquestes ens serviran de punt de partida per dissenyar una estratègia de transformació i aconseguir un model més íntegre i relacional.

Gràfic del model de dades de la BBDD temporal



5.2. Fase II Model Relacional

Traspàs de les dades de la base de dades temporal a la base de dades relacional.

Tot i que la base de dades temporal és un bon punt de partida, no cal dir que encara tenim per solucionar tots els problemes d'integritat, consistència, etc... abans comentats.

Per tal de començar a depurar dades i normalitzar-les procedirem a generar un model de dades relacional que englobi totes les dades amb les normes d'integritat i consistència adequades.

A fi i efecte de resoldre tots aquests problemes i donar-li una sol·lució definitiva, procedim a generar les següents taules.

A continuació breu descripció de les taules relacionades que marcaran les dimensions de la taula de fets.

Dimensió País

Nom de la taula	Camps
Pais	**Id_pais**, nom_pais

Dimensió país. Taula que designarà els països de manera inequívoca.

Dimensió Pel·lícula

Nom de la taula	Camps
Pelicula	**Id_pelicula**, titol_original, titol_local, titol_sort, id_pais, rati_exit

Dimensió pel·lícula. Taula que designarà les pel·lícules de manera inequívoca.

Trobem que també té una clau forana que apunta a la taula de països. Hem de tenir en compte que les pel·lícules han estat produïdes en un país concret. Finalment hi ha un camp de rati. Aquest camp conté el percentatge d'èxit de les pel·lícules. El càlcul d'aquest camp es obtingut de la següent manera. $\frac{Nom.Guanyades}{Nom.Totals} * 100$ per pel·lícula.

Dimensió Resultat

Nom de la taula	Camps
Resultat	**Id_resultat**, nom_resultat

Dimensió Resultat. Taula que designarà els resultat de la nominació i si efectivament el guardonat ha estat guanyador o no.
Tot i que en principi no faria falta, creiem que en la taula de fets multidimensionals és important que aquest atribut quedi ben definit per una clau única.

Dimensió Entitat

Nom de la taula	Camps
Entitat	**Id_entitat,** nom_entitat

Dimensió Entitat. Taula que designarà les entitats de manera inequívoca.
Entenem per entitats les entitats físiques o persones que seran objecte de les nominacions.

Dimensió Categoria

Nom de la taula	Camps
Categoria	**Id_certamen, Id_categoria**, nom_categoria, nom_angles

Dimensió Categoria. Taula que designarà les categories de manera inequívoca.
Aquesta taula és un tant especial. Degut a les diferents maneres que hi ha de categoritzar els premis amb diferents nomenclatures, diferents idiomes, etc... hem arribat a la conclusió que cada certamen té unes categories diferents, potser molt semblants, però diferents, per això hem creat una taula on cada certamen disposa de les seves categories.

Dimensió Certamen

Nom de la taula	Camps
Certamen	**Id_certamen**, nom_certamen

Dimensió Certamen. Taula que designarà els certàmens de manera inequívoca.

Dimensió EdicioCertamen

Nom de la taula	Camps
EdicioCertamen	**Id_certamen, id_edicio**, any_edicio

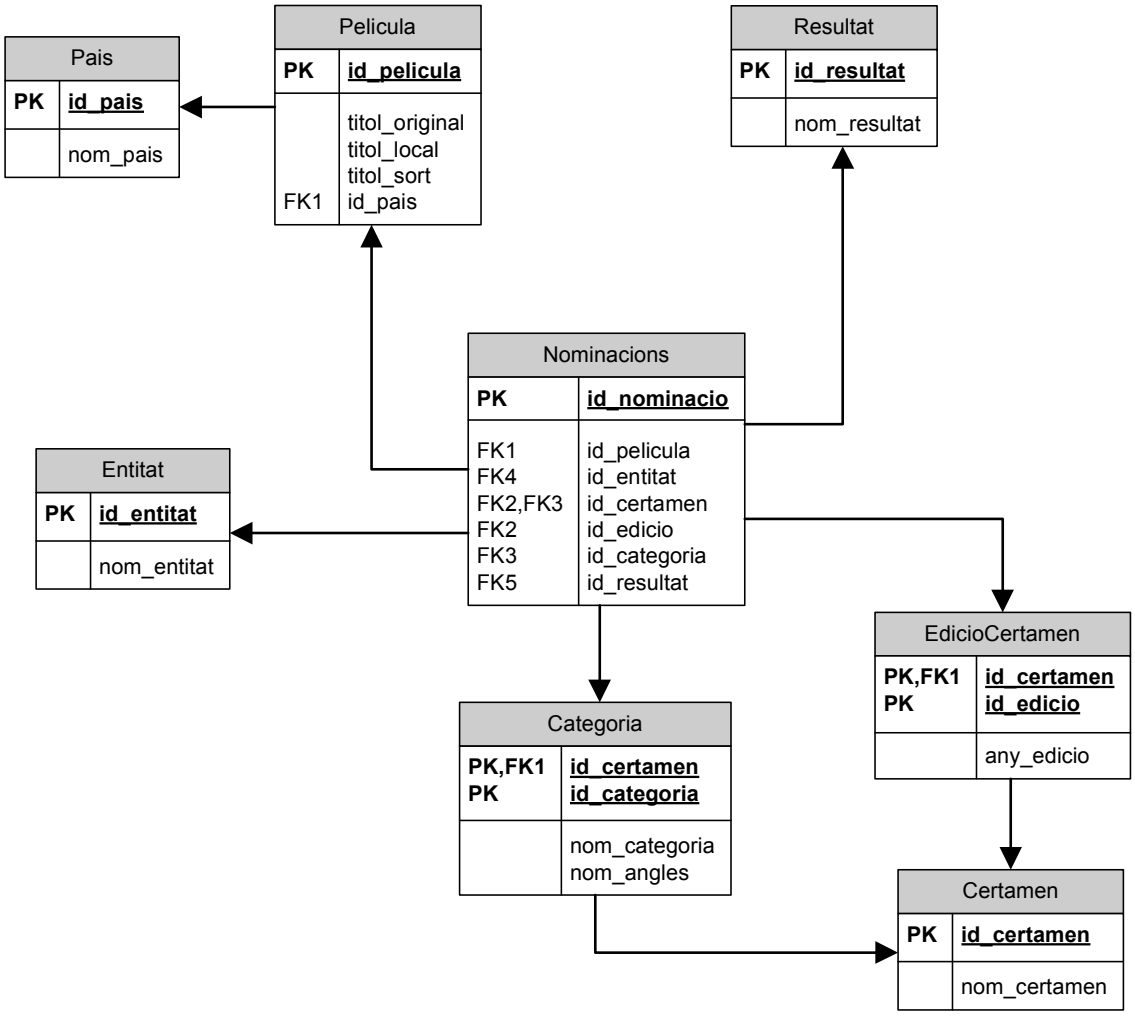
Dimensió EdicioCertamen. Taula que designarà els certàmens de manera inequívoca.
En ella, destaquem que cada edició ordinal de cada certamen es produeix en un any determinat que designem amb l'atribut any_edició.

Taula de Fets Nominacions

Nom de la taula	Camps
Nominacions	**Id_nominacio**, id_pelicula, id_entitat, id_certamen, id_edicio, id_resultat, id_categoria

La taula de fets, donarà suport a totes les transaccions que es produeixin. La seva clau primària i única ens assegurarà la seva unicitat de registre i totes les seves claus foranes que apunten de manera relacionada cap a les taules de les dimensions ens asseguraran la integritat referencial que estem buscant.

Gràfic del model relacional

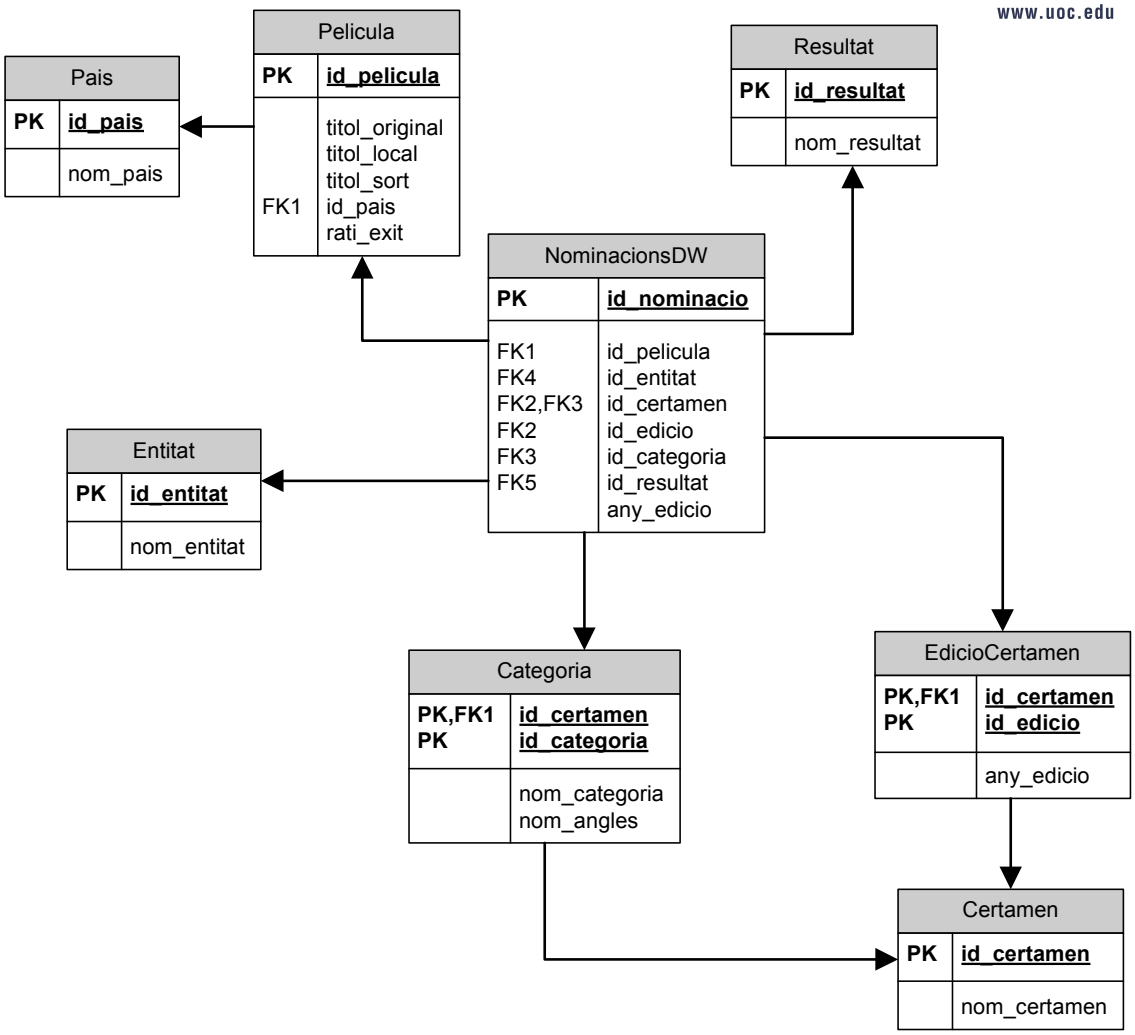


Un cop tinguem aquesta base de dades relacional creada, aquesta servirà a l'ACA com a Base de dades de gestió en la qual podrà registrar de manera normalitzada totes les dades relacionades amb tots els certàmens de manera regular i regulada.

5.3. Fase III Model Multidimensional

Traspàs de les dades de la base de dades relacional a la base de dades multidimensional.

Finalment, un com tenim la base de dades relacional en perfecte estat, procedim a dissenyar una base de dades multidimensional que passi a ser el sistema d'informació de l'ACA i doni suport a totes les necessitats d'informació que han estat plantejades en els requeriments originals.



Bàsicament, el model de dades és molt similar a la base de dades relacional. En principi, només canviarà la taula de fets.

- Afegim el camp any_edicio com a propietat de la taula general de fets NominacionsDW. Correspon a l'any que es produeix el certamen al qual la pel·lícula es presenta i resulta nominada.

Cal esmentar que podríem substituir dins la taula de fets alguns identificadors per les seves descripcions. En alguns casos ens pot aportar millores de rendiment, encara que en altres casos perdríem el fet que dos actors poguessin dir-se igual.

5.4. Depuració i Normalització de les dades

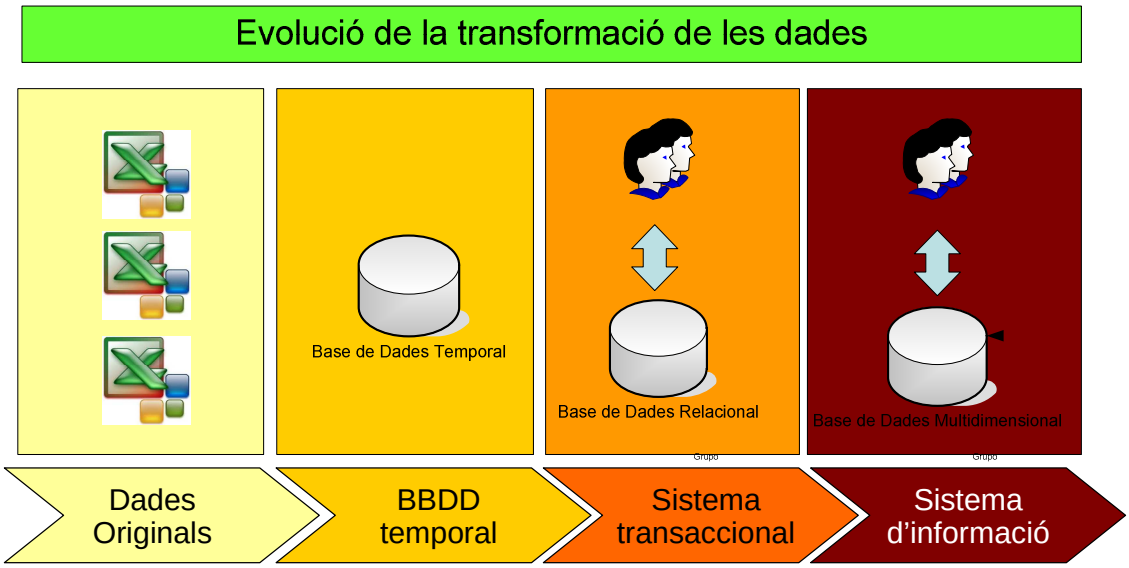
Per tal de donar consistència i robustesa a la base de dades relacional i multidimensional, s'ha hagut de realitzar algunes modificacions de les dades a la base de dades temporal per tal de normalitzar-les.

Alguns d'aquests criteris de normalització són els següents:

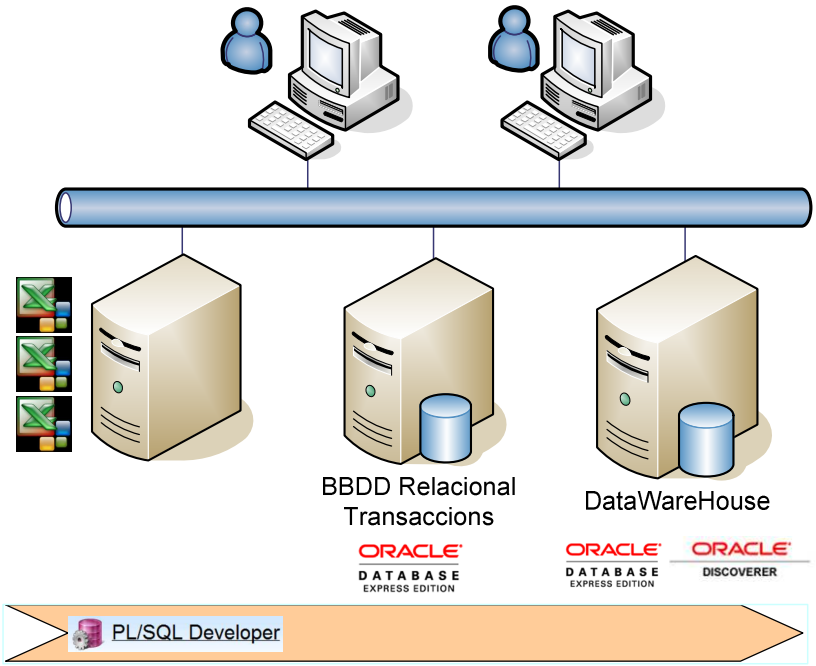
- Totes les nominacions que tenien l'any amb un format 1929-1930, han estat assignats a l'any 1930.
- Tots les nominacions que el país el tenien en blanc o amb un 0, han estat assignats al país EUA.
- Totes les nominacions amb el camp guanyador 0 o blanc s'assignaran a perdedor, en canvi, el camp guanyador serà X.
- En les categories de les cançons, s'ha suprimit la cançó, i només s'ha deixat la persona a la taula d'entitats.
- Els nominats que estaven en blanc, s'ha assignat el valor "Persona no assignada" en la taula d'entitats.
- Els nominats que estaven amb un 0, s'ha assignat el valor "Persona no assignada" en la taula d'entitats.
- Si l'edició del certamen comença amb 2 dígits, assignem els dos dígits.
- Si l'edició del certamen comença amb 1 dígit, assignem només un dígit.
- Si el titol_original és blanc, s'assigna el titol_local.
- Si el titol_local és blanc, s'assigna el titol_original.
- Si el titol_sort és blanc, s'assigna el titol_original.

5.5. Evolució gràfica de la transformació de les dades

A continuació veiem de manera il·lustrada l'evolució que farem de les dades des dels Excels originals fins a obtenir el magatzem de dades multidimensional.



A continuació mostrem l'estructura de software i hardware que donarà suport a tota la infraestructura del nostre magatzem de dades.



Procés ETL. Oracle Loader

5.6. Disseny i descripció dels informes creats

A continuació es mostra el disseny i la descripció de cadascun dels informes demanats. De la mateixa manera s'explica les característiques del filtre aplicat i les dimensions i mètriques utilitzades.

També es mostra en cada apartat de cada informe una il·lustració del propi informe per ajudar a la seva comprensió.

5.6.1. Informe 1

Nombre de premis aconseguits per un actor/actriu per any i festival. L'informe ha d'estar totalitzat per saber el nombre total de premis.

Per tal de facilitar l'explotació de les dades, hem posat l'any de la edició del certamen com a element de pàgina. D'aquesta manera poder seleccionar l'any que ens interessa saber sense que ens destorbin les altres dades.

Elementos de Página: Ano Edicio: 2003

		Id Valor
Academy Award		2
	Charlize Theron	1
	Sean Penn	1
Berlinale		3
	Catalina Sanino Moreno	1
	Charlize Theron	1
	Daniel Hendler	1
Cannes Festival		2
	Marie Josée Croze	1
	Muzaffer Ozdemir, Mehmet Emin Toprak	1
César		4
	Mathieu Amalric	1
	Omar Sharif	1
	Sylvie Testud	1
	Yolande Moreau	1
MTV Movie Awards		2
	Eminem	1
	Kirsten Dunst	1
Total de l'Any		13

Dimensions i mètriques

Per aquest informe, s'han seleccionat les següents dimensions i mètriques.

Dimensions	Mètriques
Certamen	Id_valor (recompte registres)
Entitat	Any Edicio

I s'han distribuït per la matriu amb la següent disposició:

Filtres

Per tal d'aconseguir les dades necessàries hem hagut de crear dos filtres.
El primer per escollir només les nominacions (NominacionsDW) que han estat guanyadores (NominacionsDW.id_resultat = 2).

Recordem la taula de resultats:

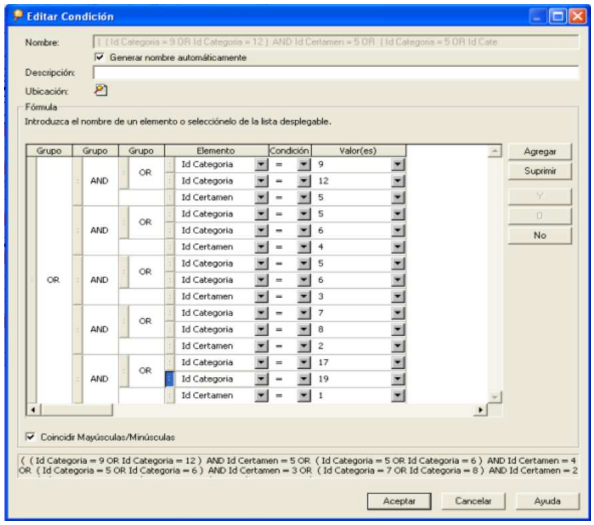
Id_resultat	Resultat
1	Nominació no guanyadora
2	Nominació guanyadora

I el segon, per tal de seleccionar només els registres que guanyen els actors i les actrius hem hagut d'anar per tots els certàmens i esbrinar quines categories corresponen a aquests premis per cada categoria.

Establint el següent criteri:

Id_certamen	Certamen	Id_categoria	Categoria
1	Academy Award	17	Best Actor
1	Academy Award	19	Best Actress
2	MTV Movie Awards	7	Best Male Performance
2	MTV Movie Awards	8	Best Female Performance
3	Berlinale	5	Silberner Bär für die beste Darstellerin
3	Berlinale	6	Silberner Bär für den besten Darsteller
4	Cannes Festival	5	Best Actress
4	Cannes Festival	6	Best Actor
5	César	9	Meilleur Acteur
5	César	12	Meilleure Actrice

Seguint el criteri establim la següent condició:



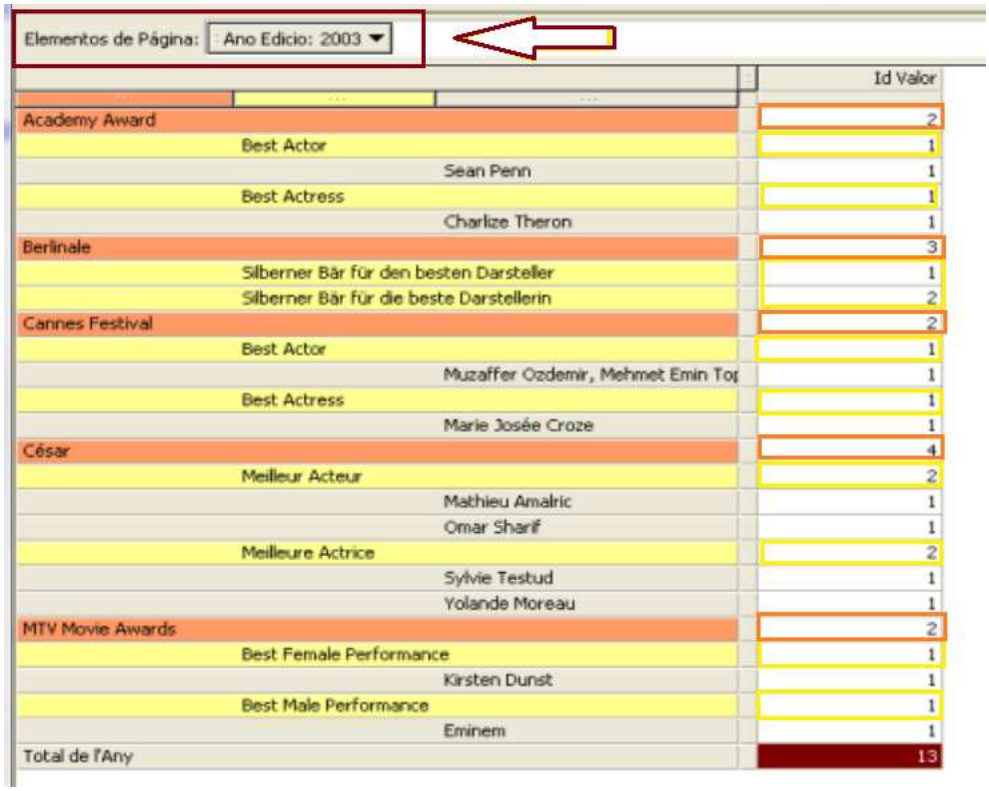
5.6.2. Informe 2

Nombre de premis aconseguits per un actor/actriu per any, festival i categoria.

Aquest informe és molt similar al anterior.
L'única diferència és que en aquest cas hi afegim la categoria com a dimensió.

A l'igual que en l'informe anterior, hi hem afegit l'any de l'edició del certamen com a element de pàgina.

A nivell de filtres, l'informe 2 és exactament igual que l'informe 1.

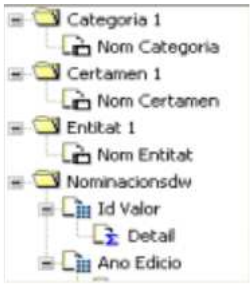


Elementos de Página: Ano Edicio: 2003

			Id Valor
Academy Award			2
	Best Actor		1
		Sean Penn	1
	Best Actress		1
		Charlize Theron	1
Berlinale			3
	Silberner Bär für den besten Darsteller		1
	Silberner Bär für die beste Darstellerin		2
Cannes Festival			2
	Best Actor		1
		Muzaffer Ozdemir, Mehmet Emin Top	1
	Best Actress		1
		Marie Josée Croze	1
César			4
	Meilleur Acteur		2
		Mathieu Amalric	1
		Omar Sharif	1
	Meilleure Actrice		2
		Sylvie Testud	1
		Yolande Moreau	1
MTV Movie Awards			2
	Best Female Performance		1
		Kirsten Dunst	1
	Best Male Performance		1
		Eminem	1
Total de l'Any			13

Dimensions i mètriques

Per aquest informe s'han seleccionat les següents dimensions i mètriques.



Dimensions	Mètriques
Certamen	Id_valor (recompte registres)
Categoria	Any Edicio
Entitat	

I s'han distribuït per la matriu amb la següent disposició:

5.6.3. Informe 3

Nombre de premis aconseguits per un director per any, festival i pel·lícula L'informe ha d'estar totalitzat per saber el nombre total de premis.

A l'igual que en anteriors informes, optem per posar l'any del certamen com a camp d'element de pàgina. Això facilita l'explosió de les dades.

Dimensions i mètriques

Per aquest informe s'han seleccionat les següents dimensions i mètriques.

Dimensions	Mètriques
Entitat	Id_valor (recompte registres)
Certamen	Any Edicio
Pel·lícula	

I s'han distribuït per la matriu amb la següent disposició:

Filtres

Per tal d'aconseguir les dades necessàries hem hagut de crear dos filtres.
El primer per escollir només les nominacions (NominacionsDW) que han estat guanyadores (NominacionsDW.id_resultat = 2).

Recordem la taula de resultats:

Id_resultat	Resultat
1	Nominació no guanyadora
2	Nominació guanyadora

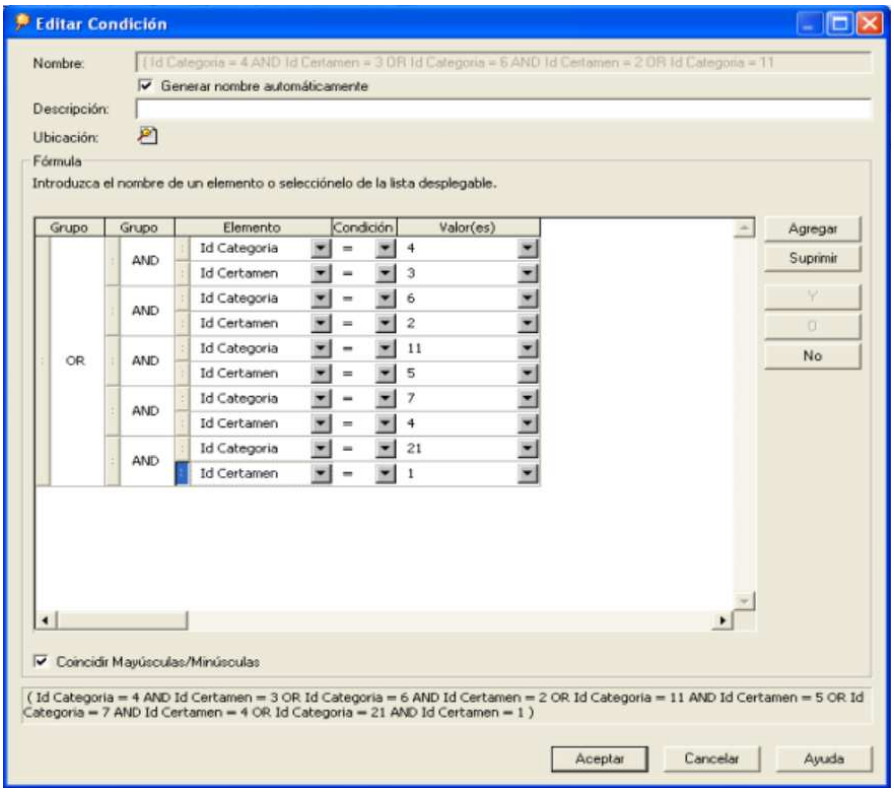
I el segon, per tal de seleccionar només els registres que guanyen els directors hem hagut d'anar per tots els certàmens i esbrinar quines categories corresponen a aquests premis per cada categoria.

Establint el següent criteri:

Id_certamen	Certamen	Id_categoria	Categoria
1	Academy Award	21	Best Director
2	MTV Movie Awards	6	Best Movie
3	Berlinale	4	Silberner Bär für die beste Regie
4	Cannes Festival	7	Direction
5	César	11	Meilleur Réalisateur

Hem de tenir en compte que el certamen MTV Movie Awards no disposa d'una categoria per als millors directors. Per aquest cas ens hem decantat per la categoria Best Movie. Aquesta categoria creiem que pot englobar la feina d'un director.

Seguint el criteri establim la següent condició:



5.6.4. Informe 4

Nombre de premis aconseguits per any, festival, categoria (s'ha de poder triar si es vol veure l'estàndard o la pròpia) i pel·lícula.

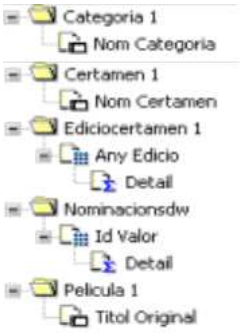
A l'igual que en anteriors informes, optem per posar l'any del certamen com a camp d'element de pàgina. Això facilita la explosió de les dades.

Elementos de Página: Any Edicio: 1932

			Id Valor
Academy Award			14
	Best Actor		2
		Dr. Jekyll and Mr. Hyde	1
		The Champ	1
	Best Actress		1
		The Sin of Madelon Claudet	1
	Best Art Direction		1
		Transatlantic	1
	Best Cinematography		1
		Shanghai Express	1
	Best Director		1
		Bad Girl	1
	Best Picture		1
		Grand Hotel	1
	Best Short Subject, Cartoon		1
		Flowers and Trees	1
	Best Short Subject, Comedy		1
		The Music Box	1
	Best Short Subject, Novelty		1
		Wrestling Swordfish	1
	Best Sound Recording		1
		[no specific film title]	1
	Best Writing, Adapted Screenplay		1
		Bad Girl	1
	Best Writing, Original Story		1
		The Champ	1
	Special Award		1
		[no specific film title]	1
Total de Pel·lícules guanyadores per Any			14

Dimensions i mètriques

Per aquest informe, s'han seleccionat les següents dimensions i mètriques.



Dimensions	Mètriques
Certamen	Id_valor (recompte registres)
Categoria	Any Edicio
Pel·lícula	

I s'han distribuït de per la matriu amb la següent disposició:

Filtres

Per tal d'aconseguir les dades necessàries hem hagut de crear un filtre.
En aquest cas hem hagut d'escollir només les nominacions (NominacionsDW) que han estat guanyadores (NominacionsDW.id_resultat = 2).

Recordem la taula de resultats:

Id_resultat	Resultat
1	Nominació no guanyadora
2	Nominació guanyadora

5.6.5. Informe 5

Donada una data a seleccionar per l'ACA, per cada tipus de festival mostrar el nombre d'edicions realitzades del festival, nombre total d'actors que han estat nominats, nombre total d'actrius que han estat nominades i nombre total de pel·lícules nominades.

En aquest informe veiem cal destacar:

- Amb un rectangle taronja el nombre d'edicions que s'han produït fins a la data introduïda i en el certamen assenyalat.
- Amb un rectangle groc el nombre d'actors i actrius que han estat nominats en el certamen assenyalat i fins a la data introduïda.
- Amb un rectangle verd el nombre de vegades que els actors i actrius han estat nominats en el certamen assenyalat i fins a la data introduïda.
- Senyalat amb vermell tenim el Total de pel·lícules nominades des de la data inicial fins a la introduïda per l'usuari.

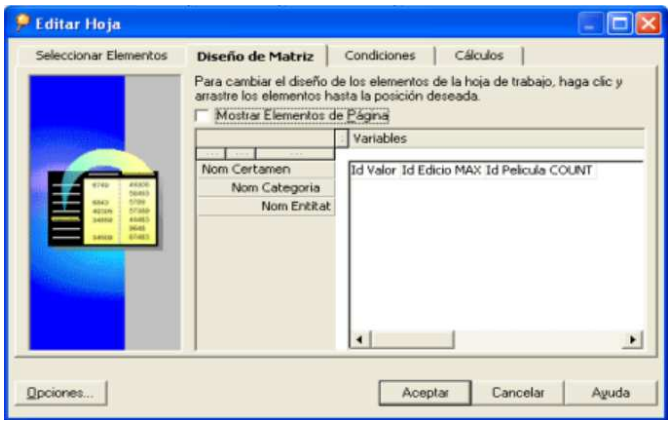
	Id Valor	Id Edicio MAX	Id Pelicula COUNT
Academy Award	26	3	35
Best Actor	12	3	17
Chester Morris	1	2	1
Emil Jannings	1	1	2
George Arliss	1	3	2
George Bancroft	1	2	1
Lawrence Tibbett	1	3	1
Lewis Stone	1	2	1
Maurice Chevalier	1	3	2
Paul Muni	1	2	1
Richard Barthelmess	1	1	2
Ronald Colman	1	3	2
Wallace Beery	1	3	1
Warner Baxter	1	2	1
Best Actress	14	3	18
Bessie Love	1	2	1
Betty Compson	1	2	1
Corinne Griffith	1	2	1
Gloria Swanson	2	3	2
Greta Garbo	1	3	2
Janet Gaynor	1	1	3
Jeanne Eagels	1	2	1
Louise Dresser	1	1	1
Mary Pickford	1	2	1
Nancy Carroll	1	3	1
Norma Shearer	1	3	2
Ruth Chatterton	2	3	2
Suma Pelicules Nominades			35

Dimensions i mètriques

Per aquest informe, s'han seleccionat les següents dimensions i mètriques.

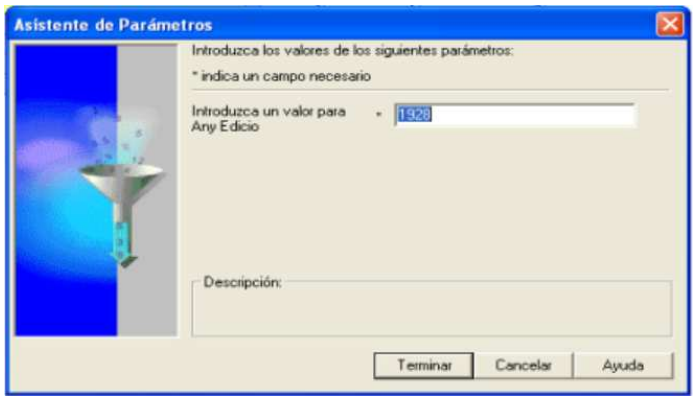
Dimensions	Mètriques
Certamen	Id_valor (recompte registres)
Categoria	id_edicio:MAX
Entitat	Id_pelicula:COUNT

I s'han distribuït de per la matriu amb la següent disposició:



Filtres

En aquest informe hem d'introduir una nova varietat de filtre. Es tracta d'un filtre que l'introdueixi l'usuari en temps d'execució. D'aqueta manera, quan executem la consulta sempre ens preguntarà l'any fins on volem fer la consulta.



En funció de l'any introduït per l'usuari, s'aplicarà el següent filtre a la consulta.

A banda, hem posat també un filtre de condició per tal de filtrar només les categories de millor actor i millor actriu pels diferents certàmens.

Per veure la condició d'aquest filtre i els criteris establerts veure l'informe 1.

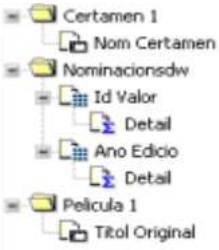
5.6.6. Informe 6

Per a cada any, nombre de premis guanyats per cada pel·lícula i festival (i subtotal per any/pel·lícula).

			Id Valor
1928			18
	[no specific film title]		1
		Academy Award	1
	Street Angel		1
		Academy Award	1
	Sunrise		3
		Academy Award	3
	Tempest		1
		Academy Award	1
	The Circus		1
		Academy Award	1
	The Dove		1
		Academy Award	1
	The Jazz Singer		1
		Academy Award	1
	The Last Command		1
		Academy Award	1
	The Way of All Flesh		1
		Academy Award	1
	Two Arabian Knights		1
		Academy Award	1
	Underworld		1
		Academy Award	1
	Wings		2
		Academy Award	2
	7th Heaven		3
		Academy Award	3
1929			7
	Coquette		1
		Academy Award	1
	In Old Arizona		1
		Academy Award	1
	The Bridge of San Luis Rey		1

Dimensions i mètriques

Per aquest informe, s'han seleccionat les següents dimensions i mètriques.



Dimensions	Mètriques
Edició	Id_valor (recompte registres)
Pel·lícula	
certamen	

I s'han distribuït de per la matriu amb la següent disposició:

Filtres

A l'igual que en altres casos, cada informe que precisa dades de nominacions guanyades hem d'escollir només les nominacions (NominacionsDW) que han estat guanyadores (NominacionsDW.id_resultat = 2).

Recordem la taula de resultats:

Id_resultat	Resultat
1	Nominació no guanyadora
2	Nominació guanyadora

5.6.7. Informe 7

Addicionalment hem de proporcionar una definició de "pel·lícula exitosa" i "pel·lícula fracassada". A partir d'aquesta definició, i per a cada any, hem de llistar la pel·lícula més exitosa i més fracassada.

Per aquest informe hem escollit la modalitat de taula.

	Id Pelicula	Titol Original	Rati Exit	Total Nominacions	Ano Edicio
1	6047	Wings	100,00	2	1928
2	6054	The Way of All Flesh	100,00	1	
3	6058	Two Arabian Knights	100,00	1	
4	6072	The Circus	100,00	1	
5	6071	Tempest	100,00	1	
6	6061	Underworld	100,00	1	
7	83	[no specific film title]	90,65	3	
8	6050	Sunrise	75,00	4	
9	6045	7th Heaven	60,00	5	
10	6051	The Last Command	50,00	2	
11	1695	The Dove	50,00	1	
12	781	Street Angel	33,33	1	
13	748	The Jazz Singer	25,00	3	
14	6055	Sadie Thompson	,00	2	
15	6049	The Crowd	,00	2	
16	6056	A Ship Comes In	,00	1	
17	6060	Glorious Betsy	,00	1	
18	6048	Chang: A Drama of the Wilderness	,00	1	
19	6053	The Patent Leather Kid	,00	1	
20	6062	The Private Life of Helen of Troy	,00	1	
21	6057	Speedy	,00	1	
22	6059	Sorrell and Son	,00	1	
23	6063	The Devil Dancer	,00	1	
24	6070	The Magic Flame	,00	1	
25	6046	The Racket	,00	1	
26	6052	The Noose	,00	1	
27	6077	Coquette	100,00	1	1929
28	782	White Shadows in the South Seas	100,00	1	

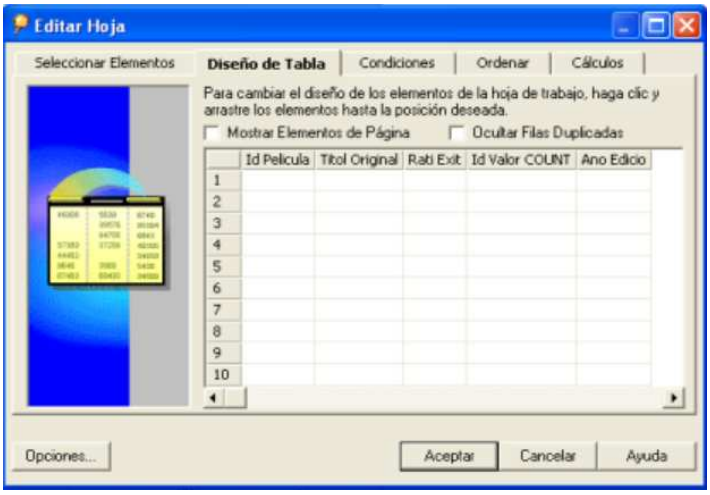
En aquesta taula hem fet ordenacions per any de l'edició, rati d'èxit i per pel·lícula. D'aquesta manera podrem agrupar les dades per any. De la mateixa manera, agrupem per pel·lícula per obtenir el total de nominacions que ha obtingut la pel·lícula. Finalment obtenim el % d'èxit de les pel·lícules. Recordem que el càlcul d'èxit surt de la següent fórmula $\frac{Nom.Guanyades}{Nom.Totals} * 100$.

Dimensions i mètriques

Per aquest informe, s'han seleccionat les següents dimensions i mètriques.

Dimensions	Mètriques
Pel·lícula	Id_valor (recompte registres)
	Any_edicio
	Rati_exit

I s'han distribuït per la taula amb la següent disposició:



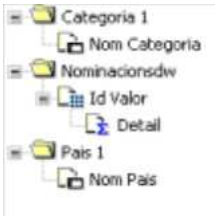
5.6.8. Informe 8

També hem de proporcionar una manera de veure el país (Estats Units a banda) que més premis ha guanyat. Per aquest país, també voldrem veure el nombre de premis per categoria guanyats.

	Id Valor
France	580
Adventure and Detective Film	1
Artistic Contribution	1
Best Actor	5
Best Actress	7
Best Director	2
Best Documentary, Feature	5
Best Foreign Language Film	9
Best Short Film	1
Best Short Film, Live Action	1
Best Short Subject, Live Action	4
Best Writing, Original Screenplay	1
Catholic Film Office Award	4
Contemporary Cinema	1
Country Selection: France	4
Décor	1
Direction	9
Film Debut	1
Fipresci Award	2
Fipresci Prize	4
Golden Camera	1
Golden Camera for Film Debut	4
Golden Palm	4
Grand Prize	3
Grand Prize Du Cinema De Creation	1
Homage	1
Honorary Award	1
Jury Prize	5
Jury Prize for Short Film	1
Jury Prize for Short Film (Fiction)	1
Meilleur Acteur	26
Meilleur Acteur dans un second rôle	27

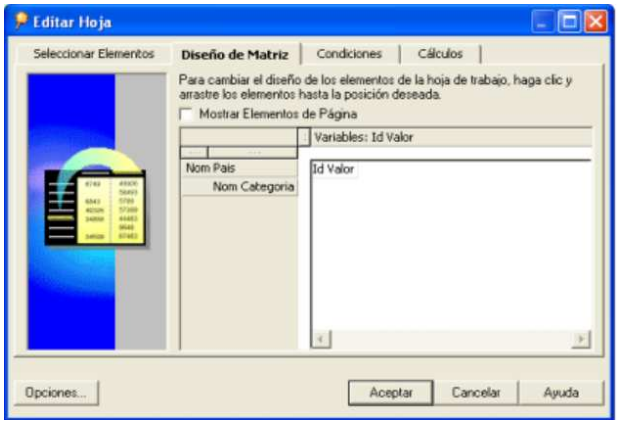
Dimensions i mètriques

Per aquest informe, s'han seleccionat les següents dimensions i mètriques.



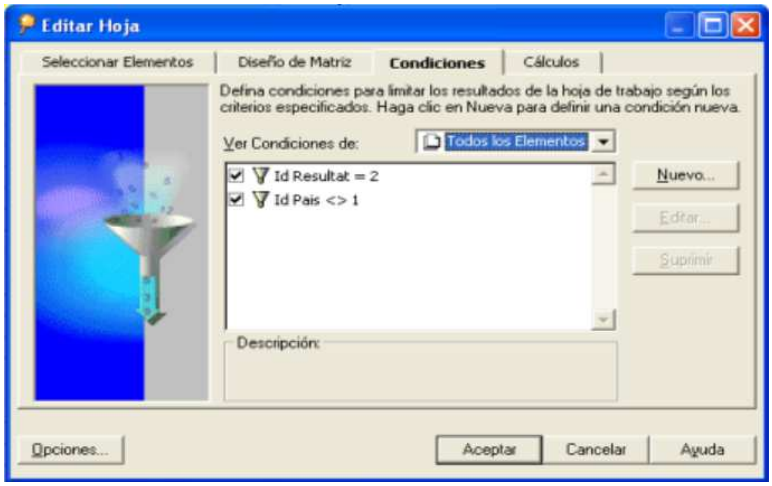
Dimensions	Mètriques
País	Id_valor (recompte registres)
Categoria	

I s'han distribuït per la matriu amb la següent disposició:



Filtres

Per aquest informe hem fet el filtre de nominacions guanyades (NominacionsDW.id_resultat = 2) i un altre per tal de descartar totes les nominacions de les pel·lícules americanes (NominacionsDW.id_pais <> 1).



Recordem la taula de resultats:

Id_resultat	Resultat
1	Nominació no guanyadora
2	Nominació guanyadora

I que en la taula de països, estats units correspon al codi 1.

6. Conclusions

Crec que el compliment dels objectius establerts tant a nivell personal com a nivell del projecte s'han assolit completament i fins i tot podríem dir que han estat sobrepassats, per tant la valoració global del TFC es pot considerar com molt positiva.

L'objectiu principal del projecte era adquirir experiència en el disseny, construcció i explotació d'un magatzem de dades a partir de la informació disponible primer a uns Excels i després en una base de dades transaccional. En aquest sentit, l'objectiu s'ha complert de manera satisfactòria.

Altres objectius també molt importants i que també considero com a aconseguits han estat el fet de conèixer els productes Oracle, tant els productes de base de dades (Oracle DataBase Express Edition), com els productes orientats al Business Intelligence (Oracle Discoverer Administration i Oracle Discoverer Desktop).

Altre objectiu important també assolit és el tema de les transformacions de les dades mitjançant els procesos ETL.

També ha estat molt positiu tot l'esforç dedicat a la planificació del projecte. Tot aquest apartat m'ha ajudat molt a seguir un mètode i un procediment i m'ha ajudat a anticipar-me als possibles problemes fent que no s'hagi produït cap tipus de desviació. Aquest apartat de la planificació segur que em servirà i l'aplicaré en la meva carrera professional a partir d'ara.

Per acabar, comentar que ha estat una gran experiència que m'ha satisfet totalment, primerament perquè els objectius plantejats originalment han estat satisfets completament, i perquè professionalment, aquesta experiència em servirà molt com a punt de partida.

7. Línies d'evolució futur

En aquest apartat intentarem tractar possibles millores que es podrien realitzar en el sistema i també diferents problemes trobats que s'haurien de millorar cara el futur.

El tema d'integració i unificació de les categories crec que és un tema fonamental que ha de millorar. Cada festival premia a unes categories concretes, i no per a tots els festivals les categories coincideixen. S'hauria de fer una taula de conversió on quedessin relacionades totes les categories de tots els festivals i quedés clar com cassen les unes amb les altres i l'equivalència entre elles. Només així podríem consolidar de manera veraç diferents categories amb diferents festivals.

Altra cosa a millorar seria el tractament dels països en el que fa referència a les coproduccions. En el projecte i en la transformació de dades, he tractat cada coproducció com un propi país amb el seu codi. En principi no està malament, però tampoc descarto que hi hagi un codi de país amb una coproducció italo-alemana i un altre alemano-italiana. Si volem donar una bona qualitat i veracitat a les estadístiques s'hauria ser més exigent en aquest sentit.

En el que fa referència als informes, crec que a priori són correctes. Els futurs informes, han d'anar relacionats a les futures peticions que puguin realitzar els usuaris de l'ACA per tal de satisfer les noves necessitats que puguin tenir.

8. Glossari

- **Business Intelligence**: Denominem Intel·ligència Empresarial, o Intel·ligència de Negocis o BI (de l'angès Business Intelligence) al conjunt d'estratègies i eines enfocades a l'administració i la creació de coneixement mitjançant l'anàlisi de les dades existents en una organització o una empresa.
- **ACA**: Acadèmia de Cinema d'Andorra.
- **DataWareHouse**: En el context de la informàtica, un magatzem de dades (del Anglès DataWareHouse) és una col·lecció de dades orientada a un determinat àmbit (empresa, organització, etc...), integrat, no volàtil i variable en el temps, que ajuda a la presa de decisions en l'entitat que s'utilitza.
- **Dimensió**: Les dimensions d'un cub són els atributs relatius a les variables, són les diferents perspectives d'anàlisi de les variables. Són els catàlegs d'informació complementària necessària per la presentació de les dades als usuaris, com per exemple: descripcions, noms, zones, rangs de temps, etc.. És a dir, la informació general complementària a cadascun dels registres de la taula de fets.
- **Mètrica**: També anomenades "indicadors de gestió", són les dades que estan sent analitzades. Formen part de la taula de fets. Més formalment, les variables

representen algun aspecte quantificable o mesurable dels objectes o dels fets a analitzar.

- **Model Relacional**: El model relacional considera la base de dades com una col·lecció de relacions. De manera simple, una relació representa una taula que no és més que un conjunt de files, cada una és un conjunt de camps i cada camp representa un valor que interpretat descriu el món real.
- **Magatzem de dades**: Veure DataWareHouse.
- **Oracle DataBase Express Edition**: Versió gratuïta del gestor de base de dades relacional desenvolupat per Oracle Corporation.
- **Oracle Discoverer Administration/Desktop**: Conjunt d'eines orientades a l'anàlisi i l'explotació de dades desenvolupat per Oracle Corporation.
- **PL/SQL**: De l'angès (Procedural Language/Structured Query Language) és un llenguatge de programació incrustat en Oracle que dóna suport a totes les consultes, ja que la manipulació de les dades que s'usa és la mateixa que en SQL, però incloent noves característiques com la utilització de variables, estructures modulars, estructures de control de fluxe i presa de decisions i control d'excepcions.
- **OLAP**: Sigles d'On-Line Analitycal Processing. És una de les solucions utilitzades en el que actualment s'anomena intel·ligència de negoci i que consisteix en la creació i utilització d'estructures de dades multidimensionals per poder realitzar anàlisis de forma adhoc.
- **ETL**: De l'anglès, (Exctract, Transform and Load). Extreure, transformar i carregar, freqüentment abreviat ETL. És el procés que permet a les organitzacions moure les dades des de diferents fonts, reformatejar-los i netejar-los per carregar-los en una altra base de dades per tal d'analitzar i donar suport als processos de negoci.

9. Bibliografia

Tota la biblioigrafia consultada ha estat bibliografia web.

- Wikipedia. Business Intelligence.
  - http://es.wikipedia.org/wiki/Inteligencia_empresarial
- Wikipedia. Tablas de hechos.
  - http://es.wikipedia.org/wiki/Tabla_de_hechos
- Wikipedia. Esquema en estrella.
  - http://es.wikipedia.org/wiki/Esquema_en_estrella
- Wikipedia. ETL
  - http://es.wikipedia.org/wiki/Extract,_transform_and_load
- Wikipedia. OLAP
  - http://es.wikipedia.org/wiki/OLAP
- Wikipedia. PL/SQL
  - http://es.wikipedia.org/wiki/PL/SQL
- Oracle Faq. Merge.
  - [http://www.orafaq.com/wiki/Merge] 15/05/2011.

- Oracle Faq. Insert.
  - [http://www.orafaq.com/wiki/Insert] 14/05/2011.

- Oracle Faq. Drop.
  - [http://www.orafaq.com/wiki/Drop] 14/05/2011.

- Oracle WebSite. Datatypes
  - [http://download.oracle.com/docs/cd/B10501_01/appdev.920/a96624/03_types.htm#LNPLS003] 10/05/2011.

10. Annexos

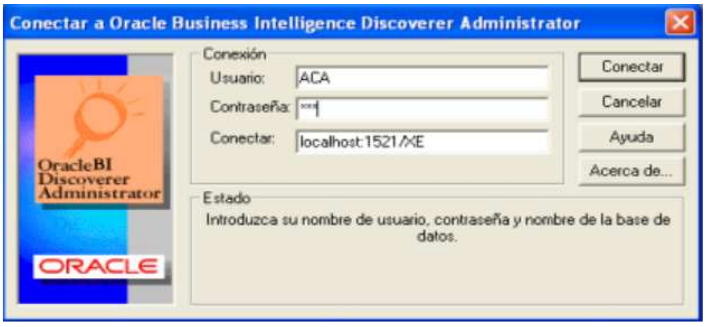
10.1. Annex 1. Credencials d'accés als diferents sistemes

Oracle Data Base

Usuari: ACA
Password: uoc

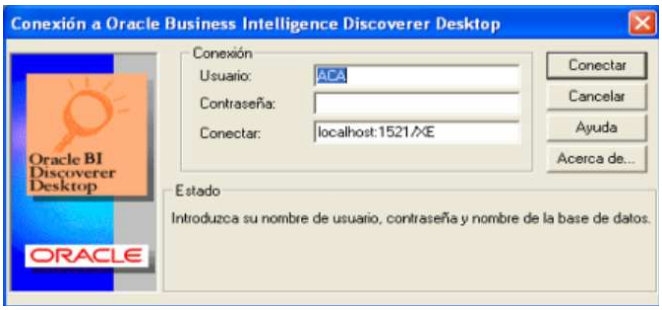
Oracle Business Intelligence Discoverer Administration

Usuari: ACA
Password: uoc
Conectar: localhost:1521/XE



Oracle Discoverer Desktop

Usuari: ACA
Password: uoc
Conectar: localhost:1521/XE



10.2. Annex 2. Ubicació de fitxers

Ubicació d'informes

La ubicació dels informes dins la màquina virtual és:

E:\informes\informe1.dis
E:\informes\informe2.dis
E:\informes\informe3.dis
E:\informes\informe4.dis
E:\informes\informe5.dis
E:\informes\informe6.dis
E:\informes\informe7.dis
E:\informes\informe8.dis

Ubicació d'Excels

La ubicació dels Excels dins la màquina virtual és:

E:\excel

Ubicació de CSV

La ubicació dels csv creats per fer les càrregues

E:\datos

10.3. Annex 3. Procés ETL

Per tal d'anar actualitzant el nostre magatzem de dades a partir de les dades que es van introduint en el model relacional, generarem un procés que s'executi cada dia. D'aquesta manera la base da dades relacional i la multidimensional estarà contínuament actualitzada.

```
MERGE INTO NOMINACIONS D
      USING (SELECT * FROM NOMINACIONSDW) S
      ON (D.id_nominacio = S.id_nominacio)
      WHEN MATCHED THEN UPDATE SET D.ano_edicio =
        (select ano_edicio from ediciocertamen
         where ediciocertamen.id_certamen = s.id_certamen
          and ediciocertamen.id_edicio = s.id_edicio)
      WHEN NOT MATCHED THEN INSERT (D.id_nominacio,
                                    D.id_pelicula,
                                    D.id_entitat,
                                    D.id_certamen,
                                    D.id_edicio,
                                    D.id_categoria,
                                    D.id_resultat,
                                    D.id_valor,
                                    D.ano_edicio)
                              VALUES (S.id_nominacio,
                                    S.id_pelicula,
                                    S.id_entitat,
                                    S.id_certamen,
                                    S.id_edicio,
                                    S.id_categoria,
                                    S.id_resultat,
                                    S.id_valor,
                                    S.ano_edicio)
```

Mitjançant PL/SQL, Oracle permet crear tasques programades. Aquestes ens permetran executar el nostre procés d'actualització entre la base de dades relacional i el magatzem de dades de manera periòdica.

```
BEGIN
DBMS_SCHEDULER.CREATE_JOB(
     job_name          =>  'ETL',
     job_type          =>  'PLSQL_BLOCK',
     job_action        =>  'MERGE INTO NOMINACIONS D
                            USING (SELECT * FROM NOMINACIONSDW) S
                            ON (D.id_nominacio = S.id_nominacio)
                            WHEN MATCHED THEN UPDATE SET D.ano_edicio =
                           (select ano_edicio from ediciocertamen
                             where ediciocertamen.id_certamen = s.id_certamen
                              and ediciocertamen.id_edicio = s.id_edicio)
                             WHEN NOT MATCHED THEN INSERT (D.id_nominacio,
                                  D.id_pelicula,
                                  D.id_entitat,
                                  D.id_certamen,
                                  D.id_edicio,
                                  D.id_categoria,
                                  D.id_resultat,
                                  D.id_valor,
```

```
                                        D.ano_edicio)
                                VALUES (S.id_nominacio,
                                        S.id_pelicula,
                                        S.id_entitat,
                                        S.id_certamen,
                                        S.id_edicio,
                                        S.id_categoria,
                                        S.id_resultat,
                                        S.id_valor,
                                        S.ano_edicio)
                                       ',
     start_date        =>  SYSDATE,
     repeat_interval   =>  'FREQ = DAILY; INTERVAL = 1');
END;

PL/SQL procedure successfully completed.
```