

Construcción y explotación de un almacén de datos de información cinematográfica

Fernando Cuenca Cabezas

ITIS

Consultor

Pere Juanola Juanola

Junio de 2011

*Para Feli, que sin su paciencia y
ayuda durante estos años
estos estudios nunca hubieran sido posibles.
Para Fer, nuestro proyecto de vida,
que nos ha llenado de felicidad.*

Muchas gracias :)

Resumen y Palabras clave

Resumen

Este proyecto tiene como finalidad el diseño, desarrollo y puesta en marcha de un almacén de datos destinado, tanto al almacenamiento de la información que nuestro cliente la Academia de Cine Andorrana dispone de los diferentes premios de los Festivales cinematográficos más importantes, como al análisis y explotación de los datos.

El proyecto parte del conjunto de documentos en formato Microsoft Excel que la Academia de Cine Andorrana utiliza como repositorio de información cinematográfica, los cuales tras un proceso de análisis y diseño darán lugar a modelo conceptual, lógico y físico de un cubo multidimensional de procesamiento Analítico Relacional o ROLAP en esquema de copo de nieve.

Así mismo, el proyecto incluye el diseño y desarrollo de los procesos necesarios de extracción, transformación y carga (ETL) de las fuentes operacionales de información, previamente exportadas a formato CSV, sobre la base de datos multidimensional.

Por último, el proyecto contempla la construcción de un área de negocio que sirve como base y entorno de trabajo para el desarrollo de los informes de extracción y explotación de los datos requeridos por la Academia de Cine Andorrana.

Palabras clave

Almacén de datos, Data Warehouse, DW, OLAP, ROLAP, cubo multidimensional, Business Intelligence, BI, Oracle Database 10g Express Edition, Oracle Business Intelligence Discoverer, ETL, PL/SQL, dimensiones, tabla de hechos.

Índice

1	Introducción	7
1.1	Justificación	7
1.2	Objetivos	7
1.2.1	Objetivos del TFC	7
1.2.2	Objetivos del proyecto	8
1.3	Enfoque y metodología	9
1.3.1	Fase 1: Definición de requisitos	9
1.3.2	Fase 2: Diseño y modelización	9
1.3.3	Fase 3: Implementación	9
1.3.4	Fase 4: Revisión	9
1.4	Planificación	10
1.4.1	Fases del TFC	10
1.4.2	Hitos del proyecto	11
1.4.3	Diagrama de Gantt	12
1.4.4	Incidencias y riesgos	13
1.5	Productos	13
1.6	Otros capítulos	14
2	Análisis	15
2.1	Datos de partida	15
2.1.1	Descripción de las fuentes	15
2.1.2	Recuento de datos	15
2.1.3	Análisis de los datos de partida	16
2.2	Diagramas de casos de uso	19
2.3	Diagrama del modelo conceptual	19
2.3.1	Definición de las dimensiones	19
2.3.2	Definición de los atributos de cada dimensión	19
2.3.3	Definición del Hecho	20
2.3.4	Definición de las medidas	20
2.3.5	Diseño Conceptual	20
3	Diseño	21
3.1	Arquitectura de Hardware	21
3.2	Arquitectura de Software	21
3.3	Diseño Lógico	22
3.3.1	Entidades	22
3.3.2	Relaciones	22
3.4	Diseño Físico	23
3.4.1	Tabla Pelicula	24
3.4.2	Tabla Pais	24
3.4.3	Tabla PeliculaPais	24

3.4.4	Tabla Nominado.....	24
3.4.5	Tabla Categoría	24
3.4.6	Tabla CategoríaTipos	25
3.4.7	Tabla Edición.....	25
3.4.8	Tabla Festival	25
3.4.9	Tabla Premios.....	25
3.4.10	Tabla temporal Temp.....	26
3.4.11	Tabla temporal ETLPaíses.....	26
3.4.12	Otros objetos.....	26
3.5	Procesos ETL	27
3.5.1	Procesos de Extracción	27
3.5.2	Procesos de Transformación	27
3.5.3	Procesos de Carga.....	29
3.5.4	Otros procesos	31
3.6	Entorno de Trabajo	32
3.6.1	Construcción del Modelo con Oracle® BI Discoverer Administrator.....	32
3.6.2	Instalación y creación de Capa de Usuario Final (EUL).....	32
3.6.3	Carga y diseño del Área de Negocio.....	32
3.6.4	Configuración del Área de Negocio.....	33
3.7	Construcción de los Informes con Oracle® BI Discoverer Desktop.....	35
3.7.1	Informe 1: Premios por actor o actriz, año y festival	35
3.7.2	Informe 2: Premios por actor o actriz, año, festival y categoría	35
3.7.3	Informe 3: Premios por director, año, festival y película.	36
3.7.4	Informe 4: Premios por año, festival, categoría y película.	36
3.7.5	Informe 5: Número de ediciones del festival, total de actores nominados, total de actrices nominadas y total de películas nominadas por tipo de festival, dada una fecha a seleccionar.....	36
3.7.6	Informe 6: Premios por año, película y festival.....	37
3.7.7	Informe 7: Definición de "película exitosa" y "película fracasada", por año.	38
3.7.8	Informe 8: País con más premios ganados. Número de premios por país y categoría.....	39
3.7.9	Informe 9: Relación entre premios en distintos festivales.	40
4	Conclusiones.....	40
5	Líneas de evolución futura.....	41
6	Glosario	42
7	Bibliografía	43
7.1	Web	43
8	Anexos.....	44
8.1	Entorno de Producción	44
8.1.1	Requisitos de servidor	44
8.1.2	Requisitos en cliente.....	44

8.2	Entorno de Desarrollo	44
8.2.1	<i>Recursos de software</i>	44
8.2.2	<i>Recursos Hardware</i>	44
8.3	Localización de Entregables del Proyecto	45
8.4	Scripts de procesos ETL.....	47
8.4.1	<i>Proceso de Extracción</i>	47
8.4.2	<i>Script de Transformación</i>	47
8.4.3	<i>Script de Carga</i>	48
8.4.4	<i>Script de actualización de estadísticas</i>	49
8.5	Carpetas personalizadas del Área de Negocio	50
8.5.1	<i>Script Pelicula_Pais</i>	50
8.5.2	<i>Script PremiosEntreCategorias</i>	50

Índice de figuras

Figura 1	- Metodología de mejora continua.....	10
Figura 2	- Diagrama de Gantt.....	12
Figura 3	- Diagrama de casos de uso.....	19
Figura 4	- Diseño conceptual	20
Figura 5	- Diagrama de arquitectura de hardware	21
Figura 6	- Diagrama de arquitectura de software.....	22
Figura 7	- Diseño lógico.....	23
Figura 8	- Diseño físico.....	23
Figura 9	- Asistente de carga: Paso 2	32
Figura 10	- Asistente de carga: Paso 3	32
Figura 11	- Área de Negocio: Datos, clases de elementos y jerarquías	34
Figura 12	- Informe 1.....	35
Figura 13	- Informe 2.....	35
Figura 14	- Informe 3.....	36
Figura 15	- Informe 4.....	36
Figura 16	- Informe 5: Número de ediciones	37
Figura 17	- Informe 5: Número de películas	37
Figura 18	- Informe 5: Número de actores	37
Figura 19	- Informe 5: Número de actrices.....	37
Figura 20	- Informe 6.....	37
Figura 21	- Informe 7: Películas exitosas	38
Figura 22	- Informe 8: Películas fracasadas	38
Figura 23	- Informe 8: Premios por país y categoría	39
Figura 24	- Informe 8: Premios por país	39
Figura 25	- Informe 9.....	40

1 Introducción

1.1 Justificación

Este proyecto surge ante la solicitud de la Academia de Cine Andorrana de construir un almacén de datos que permita, no solamente almacenar la información cinematográfica de los diferentes premios y Festivales, si no también de cruzarla entre sí y explotarla. Actualmente, la Academia mantiene la información mediante documentos Microsoft Excel. Estas hojas de cálculo contienen, cada una y de forma independiente, la información histórica relacionada con los siguientes Festivales Cinematográficos: Oscar de la Academia, Festival Internacional de Cine de Berlín, Festival Internacional de Cine de Cannes, Premios César y MTV Movie Awards.

Esta información, se encuentra disgregada, desestructurada, y separada entre sí, por lo que la gestión, explotación y utilización de la misma, requiere de procesos manuales, laboriosos y muy costosos.

Es por tanto necesario, integrar y almacenar dichas fuentes de datos en un único repositorio de información estructurado, organizado y explotable. La solución propuesta a la problemática de nuestro cliente la Academia de Cine Andorrana, es el diseño y construcción de un Data Warehouse (DW) que almacene la información del cliente, el cual dispondrá de las siguientes características:

- Orientado a temas, porque los datos estarán organizados para que los elementos relacionados estén unidos entre sí.
- Variante en el tiempo: para que las variaciones de los datos durante el tiempo, se reflejen el DW.
- No volátil, ya que la información del DW será de sólo lectura.
- Integrado, del tal forma que contenga toda la información de todos los sistemas operacionales necesarios, de forma consistente y relacional.

1.2 Objetivos

Podemos definir los objetivos como la suma de los objetivos finales en varias vertientes: los objetivos propios del proyecto solicitado por la Academia de Cine Andorrana y los objetivos propios del Trabajo Fin de Carrera y la propia asignatura.

1.2.1 Objetivos del TFC

El objetivo principal de este TFC es adquirir experiencia en el diseño, construcción y explotación de un almacén de datos a partir de fuentes de datos operacionales. Este objetivo principal, podemos descomponerlo en objetivos puntuales como:

- Adquisición y consolidación de experiencia en el diseño y construcción de bases de datos.
- Obtención y consolidación de conocimientos en sentencias DDL y DML, mediante lenguaje de programación PL/SQL, para la construcción, actualización, selección y optimización de un almacén de datos.

- Obtención de experiencia en el uso y tratamiento de herramientas que faciliten la explotación de datos y generación de informes.
- Obtención de conocimientos acerca de los Data Warehouse y cubos multidimensionales.
- Asimilar y consolidar conceptos relacionados con el Data Warehouse, como dimensiones, atributos, hechos, ...

Así mismo, consideramos igualmente objetivos de este TFC, las tareas relacionadas con la gestión del propio proyecto, tales como:

- Obtención y consolidación de conocimientos relacionados con el tratamiento de la documentación relativa a un proyecto de índole técnica.
- Obtención de experiencia en la planificación y consecución de proyectos.

1.2.2 Objetivos del proyecto

El objetivo del proyecto solicitado por la Academia de Cine Andorrana, consiste en el diseño y construcción de un almacén de datos que mantenga su información y permita su fácil explotación mediante informes. Podemos dividir este objetivo en otros más puntuales, cómo:

- Diseño y construcción de un almacén de datos estructurado según las fuentes operacionales que gestiona la Academia de Cine, que facilite su explotación mediante informes ágiles y efectivos.
- Diseño y desarrollo de un proceso automatizado de importación de datos, el cual permita alimentar el almacén desarrollado de forma sencilla y transparente al usuario final. Este proceso deberá ser capaz de incorporar sobre el modelo final los nuevos datos previo tratamiento y optimización de los mismos.
- El modelo final deberá ser fácilmente explotable. Para ello, se desarrollan los siguientes informes:
 - o Número de premios conseguidos por un actor/actriz por año y festival
 - o Número de premios conseguidos por un actor/actriz por año, festival y categoría.
 - o Número de premios conseguidos por un director por año, festival y película. Informe totalizado.
 - o Número de premios conseguidos por año, festival, categoría y película. Se podrá elegir si mostrar la categoría estándar o la propia.
 - o Numero de ediciones, número de actores nominados, número de actrices nominadas y número total de películas nominadas, por festival y dada una fecha.
 - o Número de premios ganados por cada película, festival y año. Informe subtotalizado por año y película.
 - o Se definirá película exitosa y película fracasada. Se listarán la película más exitosa y fracasada por año.
 - o Informe de con los premios ganados por país. Se mostrarán, para cada país, el número de premios por categoría.
 - o Informe comparativo entre premios por categoría y festival.

1.3 Enfoque y metodología

Podemos dividir el proceso de creación e implantación de un proyecto de Data Warehouse en cuatro fases o grupos de tareas, las cuales ocurren dentro de un ciclo iterativo de mejoras continuas. Estas fases son:

1.3.1 Fase 1: Definición de requisitos

Durante esta fase, se identifican y recogen todas las fuentes operacionales de datos de los usuarios implicados en el proyecto. Con ellas se definen los requisitos del mismo y su alcance. En esta fase podremos definir una planificación inicial del proyecto.

1.3.2 Fase 2: Diseño y modelización

Fase de estudio y análisis de las fuentes operacionales, y de diseño del Data Warehouse. La salida de esta fase será un modelo lógico y un diseño técnico a implementar, el cual se desarrollará en la fase siguiente.

1.3.3 Fase 3: Implementación

Fase de desarrollo y construcción del proyecto. Es una fase puramente técnica, en la cual podemos definir varias subfases o grupos de tareas:

- Desarrollo de procesos de Extracción, Transformación y Carga (ETL). Comprende los trabajos de extracción de los datos necesarios de las fuentes operacionales, las tareas de limpieza y transformación de los mismos, y la carga en el nuevo repositorio.
- Desarrollo del modelo OLAP. Tareas de construcción de la base de datos.
- Desarrollo de los informes. Construcción de los métodos de consulta y explotación de los datos.

1.3.4 Fase 4: Revisión

Esta es la fase de cierre del ciclo, la cual incluye tareas como: depuración de errores, ajuste y optimización del rendimiento de los nuevos procesos analíticos construidos, verificación de los informes obtenidos asegurando que los datos resultantes tiene la calidad esperada.

Una vez cerrado el ciclo, comienza una nueva iteración del mismo con nuevos requisitos, añadiendo nuevas fuentes operacionales y nuevos informes, o bien, asegurando y mejorando la calidad de los procesos existentes.



Figura 1 - Metodología de mejora continua

1.4 Planificación

1.4.1 Fases del TFC

Según el calendario y las fases del TFC, encontramos 3 fases definidas y relativas a la construcción del Almacén de datos:

1.4.1.1 PEC1: Plan de trabajo y análisis preliminar de requerimientos

Fase de definición de requisitos del proyecto de construcción de un DW. Analizamos las fuentes de datos operacionales y definimos los elementos analíticos identificados. Además diseñamos un plan de trabajo para el proyecto.

1.4.1.2 PEC2: Análisis de requerimientos, diseño conceptual y técnico

Se diseñará un documento técnico con el análisis del modelo que soportará los requisitos. En esta etapa se identifican las tareas definidas en la fase 2, diseño y modelización, del proyecto de construcción de un DW.

1.4.1.3 PEC3: Implementación

Fase de construcción del almacén de datos y de los informes, y analizaremos la información extraída. Se identifican con esta etapa las fases 3 y 4 del modelo de proceso de construcción del DW.

1.4.1.4 Memoria Final. Cierre de entregables

Aunque no corresponde propiamente con una fase del proceso de implantación de un proyecto de DW, en esta tarea se asemejaría a la generación de documentación, manuales, documentos de *Win & Lost* o conclusiones finales, presentación ejecutiva del producto o preparación de documentación de cierre de proyecto. Es decir, parte de la fase 4 de revisión junto con el cierre de entregables y paso a producción.

Obviamente, queda fuera del alcance de este TFC y por lo tanto del proyecto, el proceso cíclico de mejora continua del almacén de datos.

1.4.2 Hitos del proyecto

Se establecen como hitos del proyecto los siguientes:

HITO	OBJETIVO
PEC1 (2011/03/05 – 2011/03/16)	
DEP - Documento de Estudio Preliminar <ul style="list-style-type: none"> Entrega 	Revisión y firma del documento.
DPP - Documento de Planificación del Proyecto <ul style="list-style-type: none"> Entrega 	Este documento. Definición de planificación y tareas del proyecto.
DADP - Documento de Análisis de Datos de Partida <ul style="list-style-type: none"> Entrega 	Proceso conjunto entre Cliente y Analistas donde se definan los procesos a realizar de forma más detallada. Definición de los diferentes requisitos resultantes del análisis conjunto.
PEC2 (2011/03/17 – 2011/04/20)	
DDT - Documento de Diseño Técnico <ul style="list-style-type: none"> Entrega 	Documentación técnica del proyecto. Diseño conceptual. Diseño Físico. Diseño Lógico. Definición de procesos ETL.
DPP - Documento de Revisión de la Planificación del Proyecto <ul style="list-style-type: none"> Revisión y entrega 	Revisión y entrega, en caso de ser necesarios, de la planificación y tareas del proyecto.
PEC3 (2011/04/21 – 2011/05/25)	
Implementación <ul style="list-style-type: none"> Desarrollo Testeo y validación Entrega 	Tareas de desarrollo e implementación del proyecto. Diseño técnico del proyecto. Programación de scripts. Desarrollo de informes. Pruebas y validación del desarrollo.
DDT - Documento de Desarrollo e Implantación del Proyecto <ul style="list-style-type: none"> 	Definición y documentación de las tareas y procesos realizados en la fase de desarrollo y construcción del Proyecto.
OUT - Documento de Entregables del Proyecto <ul style="list-style-type: none"> 	Documentar todos aquellos entregables y salidas resultantes del proceso de desarrollo y construcción del Proyecto.
ENTREGA FINAL (2011/05/26 – 2011/06/13)	
MEMORIA - Memoria Final del Proyecto <ul style="list-style-type: none"> Entrega 	Documento final del TFC. Cierre de entregables. Desarrollo de presentación multimedia.

1.4.3 Diagrama de Gantt

Represento gráficamente las tareas en un diagrama de Gantt:

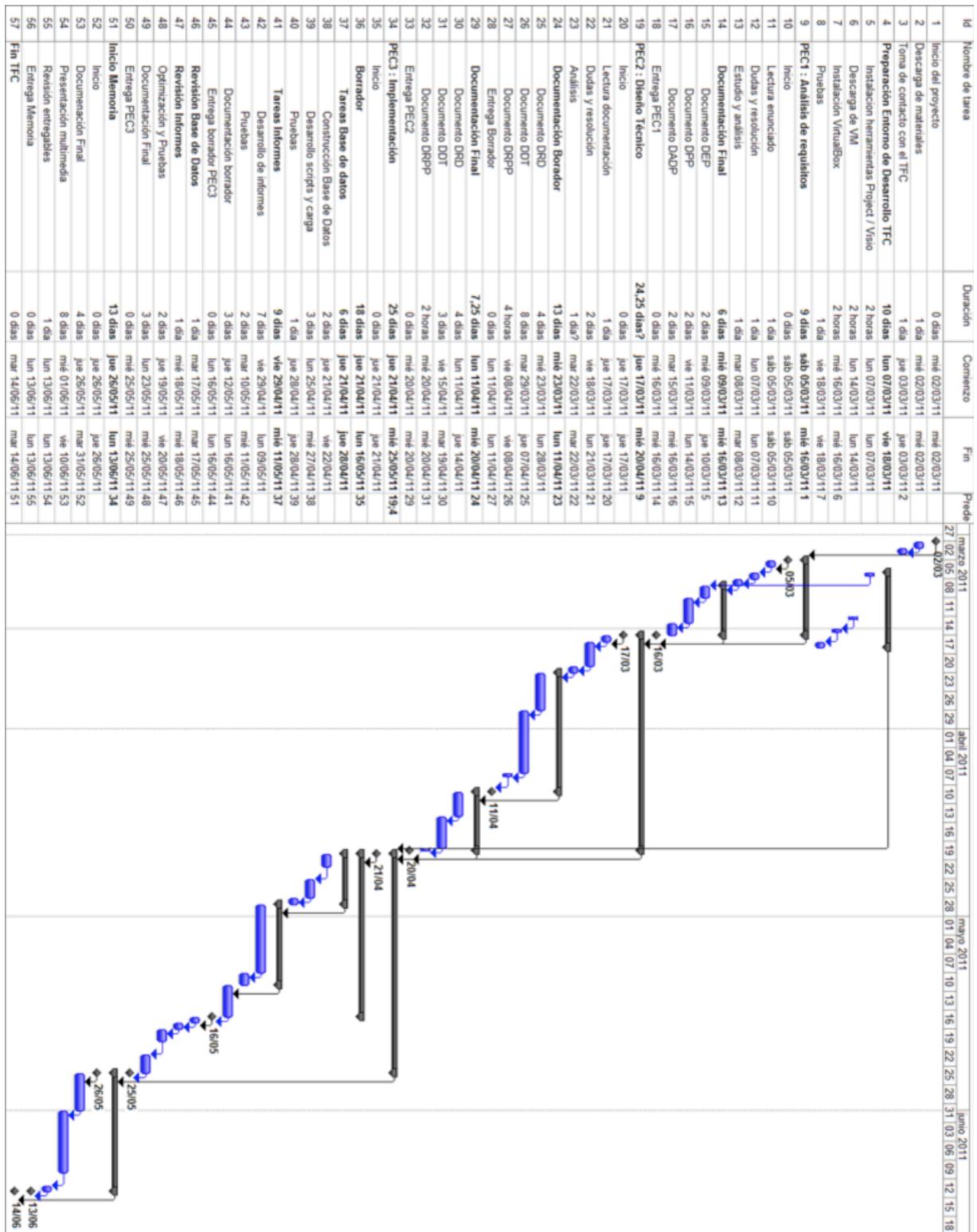


Figura 2 - Diagrama de Gantt

1.4.4 Incidencias y riesgos

Durante la fase de planificación, prevemos la posibilidad de la existencias de incidencias y riesgos en la ejecución del proyecto. Ante esta problemática, se prevé posibles soluciones como plan de contingencia:

- Incidencias causadas por motivos laborales

Ante posibles desfases en la planificación las tareas pueden ser replanificadas asegurando siempre el cumplimiento de los hitos de entrega de las PEC.

Para ello y como media prioritaria, se reducirá el tiempo libre personal. Como medida alternativa, podemos eliminar la entrega de los borradores de cada fase. Esto dará margen suficiente para la ejecución en tiempo y forma de del compromiso de entrega de PEC.

- Incidencias causadas por motivos personales o por problemas de salud.

Al igual que en el caso anterior, se reducirá el tiempo libre personal para acomodar la planificación de tareas.

- Incidencias causadas por motivos técnicos.

Aunque no se prevé problemática en este respecto, se disponen de recursos hardware y software suficientes para replicar el escenario de trabajo rápidamente. Se habilitan medidas de réplica de datos, mediante servicios externos como *DropBox* y *Apple mobileMe*, que eliminen el riesgo de desastre.

1.5 Productos

Durante la ejecución del proyecto, se han generado el siguiente listado de documentos y archivos:

- Documento de Estudio Preliminar.
- Documento de Análisis de Datos de Partida.
- Documento de Planificación del Proyecto.
- Documento de Diseño Técnico.
- Documento de Desarrollo e Implantación del Proyecto.
- Documento de Entregables del Proyecto.
- Revisión del Documento de la Planificación del Proyecto.
- Revisión del Documento de Diseño Técnico.
- Memoria final del proyecto.
- Presentación multimedia del proyecto.
- Entorno de desarrollo.
- Proceso BATCH de automatización de las tareas ETL y actualización de estadísticas del sistema.
- Scripts de programación DDL de construcción de los objetos de la base de datos.
- Ficheros de control (CTL), fuente (CSV) y de registros (LOG), utilizamos en el proceso de extracción de datos.
- Scripts de procesos de transformación de los datos.
- Scripts de procesos de carga de los datos transformados.

- Scripts y lanzador manual del proceso de actualización de estadísticas de los objetos de Oracle (índices y tablas).
- Proceso manual de backup de los datos del esquema TFC de Oracle para recuperación manual ante posibles desastres.
- Backup generado desde *Oracle® BI Discoverer Administrator*, para recuperación ante posibles desastres.
- Programación DML de las sentencias de selección utilizadas en el Área de Negocio de *Oracle® BI Discoverer Administrator*.
- Informes de consulta propietarios de *Oracle® BI Discoverer Desktop*:
 - o Informe 1: Premios por actor o actriz, año y festival
 - o Informe 2: Premios por actor o actriz, año, festival y categoría
 - o Informe 3: Premios por director, año, festival y película.
 - o Informe 4: Premios por año, festival, categoría y película.
 - o Informe 5: Número de ediciones del festival, total de actores nominados, total de actrices nominadas y total de películas nominadas por tipo de festival, dada una fecha a seleccionar.
 - o Informe 6: Premios por año, película y festival.
 - o Informe 7: Definición de "película exitosa" y "película fracasada", por año.
 - o Informe 8: País con más premios ganados. Número de premios por país y categoría.
 - o Informe 9: Relación entre premios en distintos festivales.

1.6 Otros capítulos

En capítulos posteriores de esta memoria, se refleja el proceso de análisis y diseño del proyecto.

El capítulo de análisis contiene el estudio de las fuentes operacionales o datos de partida y su resultado, el modelo conceptual del mismo.

En lo relativo al apartado de diseño, se estudiará el modelo físico y lógico de nuestro almacén de datos, junto con la definición de los procesos de Extracción, Transformación y Carga de datos y el diseño de los informes finales.

2 Análisis

Esta fase tiene como objetivo definir y analizar las fuentes operacionales del proyecto. Es el punto de partida del proyecto, y su salida será un modelo conceptual del mismo, el cual contendrá los elementos constituyentes del almacén de datos final.

2.1 Datos de partida

2.1.1 Descripción de las fuentes

El cliente nos facilita cinco fuentes de datos, en formato Microsoft Excel 2003 o anterior (.xls).

Cada uno de los ficheros contienen los datos de cada uno de los cinco Festivales:

- Oscar de la Academia: Academy_Awards_2006.xls
- Festival Internacional de Cine de Berlín: Berlinale_2004.xls
- Festival Internacional de Cine de Cannes: Cannes_Awards_2005.xls
- Premios César: Cesar_Awards_2005.xls
- MTV Movie Awards: MTV_Movie_Awards_2006.xls

Los datos contenidos en las fuentes operacionales, se encuentran almacenada bajo una estructura común :

- Columna [*Original Title*]: Almacena el título original de la película.
- Columna [*English Title*]: Contiene el título de la película traducido al inglés.
- Columna [*Sort title*].
- Columna *Year*: Contiene el año de la edición del festival.
- Columna *Country*: Contiene el país de origen de la película.
- Columna *Award*: Almacena el nombre del festival.
- Columna *Category*: almacena la categoría del premio.
- Columna *Winner?*: indica si el nominado es ganador.
- Columna *Nominee(s)*: almacena el nominados/ganador.
- Columna [*Academy Awards*]: contiene la número de edición del festival.
- Columna *Item*: identificador del registro único y correlativo dentro del fichero de datos Excel.

2.1.2 Recuento de datos

Para cada uno de las fuentes de datos, encontramos el siguiente número de registros:

- Academy_Awards_2006.xls: 8.832 registros.
- Berlinale_2004.xls: 404 registros.
- Cannes_Awards_2005.xls: 675 registros.
- Cesar_Awards_2005.xls: 2.457 registros.
- MTV_Movie_Awards_2006.xls: 911 registros.

Recuento total de 13.279 registros.

2.1.3 Análisis de los datos de partida

Dado que todas las fuentes mantienen una estructura y formato común, realizaremos una unión de todas ellas en un único fichero, con el fin de disponer de una visión global de todos ellos a la vez, y no de forma separada.

De esta forma, detectaremos más rápidamente la amplitud de aquellos campos con valores discretos, posibles errores de ortográficos, de formato, valores fuera de rango, etc.

Analizamos detenidamente cada una de las columnas:

2.1.3.1 Columna [Original Title]

De los 13,279 registros, se identifican a priori, 6.092 registros con valores únicos.

Es una variable continua de 108 caracteres de longitud como mínimo, correspondiente al valor: *"Those Magnificent Men in Their Flying Machines or How I Flew From London to Paris in 25 hours and 11 minutes"*.

Incidencias:

- Los valores serán revisados detenidamente ya que, registros que deberían ser únicos, no lo son debido a "errores" ortográficos. Estas diferencias son debidas, por ejemplo, a errores como: `(espacio):(espacio)` y `:(espacio)`, o `&` y `and`.
- Existencia de registros vacíos (en blanco).
- Existencia de registros especiales ('no specific film title').

2.1.3.2 Columna [English Title]

Se identifican, a priori, 5.898 registros con valores únicos.

Es una variable continua de 82 caracteres de longitud como mínimo, correspondiente al valor: *"Borat Cultural Learnings of America for Make Benefit Glorious Nation of Kazakhstan"*.

Se revisarán detenidamente los valores, ya que no se descartan posibles incidencias, aunque este atributo no es relevante para el resultado final de los informes.

2.1.3.3 Columna [Sort title]

Se identifican, a priori, 6.147 registros con valores únicos.

Es una variable continua de 108 caracteres de longitud como mínimo, correspondiente al valor: *"Those Magnificent Men in Their Flying Machines or How I Flew From London to Paris in 25 hours and 11 minutes"*.

Se revisarán detenidamente los valores, ya que no se descartan posibles incidencias, aunque este atributo no es relevante para el resultado final de los informes.

2.1.3.4 Columna Year

Después de las transformaciones y operaciones de preparación de datos realizadas con los registros, una vez ya solucionadas las incidencias, obtenemos 80 registros con valores únicos.

Es una variable numérica entera, que puede ser discretizada en el rango 1928, 2006. Si contemplamos futuros eventos, podemos representarla en el rango de 100 valores 1928, 2027. De esta forma, podríamos definir el valor con dos números: sus dos últimos dígitos.

Incidencias:

- Como ya se informa en el documento de requisitos inicial del proyecto, podemos encontrar el valor con el siguiente formato: YYYY₁-YYYY₂. En este caso se tomará como valor YYYY₂.

2.1.3.5 Columna Country

Se identifican inicialmente 244 registros con valores únicos.

Es una variable discreta alfanumérica que toma valores del conjunto de nombres de países, y que identifica el país de origen de la película.

Incidencias:

- Como ya se indica en el documento de requisitos inicial del proyecto, el valor '0' (cero) representa también el valor 'USA'.
- Encontrar registros que contienen más de una valor, separados mediante caracteres separadores como '/' o ','. Estos registros deberán ser normalizados en 1 valor por registro.
- Encontramos varios valores que identifican un mismo país, como por ejemplo: 'UK' y 'United Kingdom', o 'USA' y 'United States'. Estos registros deberán ser normalizados.
- Encontramos que contienen el mismo valor representado en diferentes idiomas, como por ejemplo: 'Yougoslavie' o 'Yugoslavia'. Estos registros deberán ser normalizados.
- Existen otros errores ortográficos que deberán ser corregidos. Como por ejemplo: 'Süd Korea' o 'Süd-Korea'.
- Existen registros sin valor o con valor vacío, como por ejemplo todos los registros relativos al Festival 'MTV Movie Awards'. Por lo tanto, debemos crear un valor 'sin país especificado'.

2.1.3.6 Columna Award

Se identifican, 5 registros con valores únicos.

Es una variable alfanumérica discreta que toma valores del siguiente conjunto: {'Academy Award', 'César', 'MTV Movie Awards', 'Cannes Festival', 'Berlinale'}.

Este conjunto de valores podrá verse ampliado en un futuro, añadiendo datos de nuevos festivales de cine.

2.1.3.7 Columna Category

Se identifican, a priori y después de solucionar la incidencia identificada relativa a la existencia de espacios en la variable (ver más a bajo), 241 registros con valores únicos.

Es una variable continua alfanumérica de, al menos, 61 caracteres de longitud, correspondiente a la categoría con valor "AGICOA-Preis Der Blaue Engel für den besten europäischen Film".

Incidencias:

- La categoría contiene espacios al principio y final de la cadena de texto que deberán ser eliminados (trim()).

2.1.3.8 Columna Winner?

Se identifican 3 registros con valores únicos.

Aunque es inicialmente es una variable discreta que toma valores del siguiente conjunto {'0', 'X', 'Y'}, resolviendo las incidencias se convierte en una variable binaria {1, 0}, que indica si el registro es ganador o no.

Incidencias:

- Si encontramos un valor vacío en el registro, indica que no es ganador.

2.1.3.9 Columna Nominee(s)

De los 13,279 registros, se identifican a priori y tras una limpieza de espacios al principio y final de la cadena de texto, 7.875 registros con valores únicos.

Es una variable continua alfanumérica de, al menos, 255 caracteres de longitud.

Se observa que existen registros con varios valores separados por caracteres como ',' o ';'. Inicialmente se descarta la posibilidad de ser normalizados en registros independientes.

Debido a la heterogeneidad en formatos, se descarta casi cualquier proceso automatizado de tratamiento de valores.

Incidencias:

- Los valores contiene espacios al principio y final de la cadena de texto que serán eliminados (trim()).

2.1.3.10 Columna [Academy Awards]

Después de las transformaciones y operaciones de preparación de datos realizadas con los registros y una vez solucionadas las incidencias, obtenemos 79 registros con valores únicos.

Es una variable ordinal, que puede ser discretizada en el rango 1, 79. Si contemplamos futuros eventos, podemos representarla en el rango de 100 valores 1, 99.

Incidencias:

- Se normalizarán todos los registro pertenecientes al festival "César". Estos valores contienen texto complementario que será eliminado.

2.1.3.11 Columna Item

Esta columna, al no contener información relevante, será descartada en el proceso de traspaso a la base de datos.

2.2 Diagramas de casos de uso

Este diagrama definirá cómo el usuario interactuará con el sistema. Podemos identificar dos roles diferentes de usuarios o actores: el usuario administrador que se encargará de convertir previamente los ficheros de datos de origen a formato CSV, de los procesos ETL y de gestionar los accesos del resto de usuarios al Data Warehouse; y el rol de usuario operador, que será quien realice las consultas a través de los informes y de mantener los ficheros Excel.

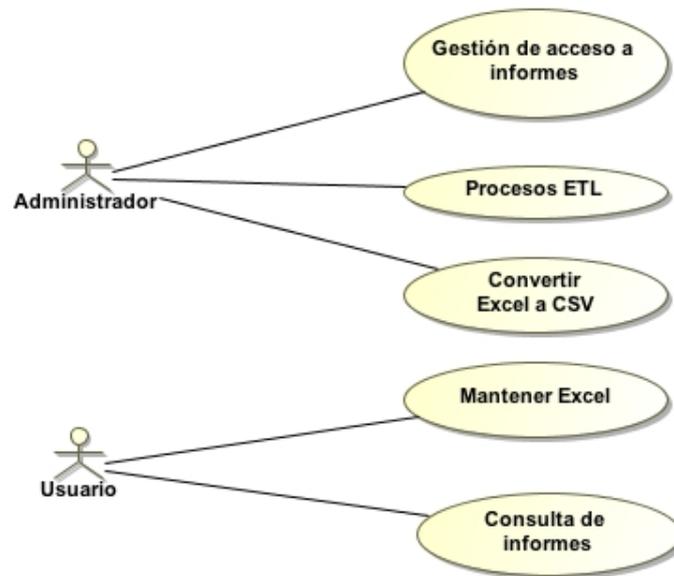


Figura 3 - Diagrama de casos de uso

2.3 Diagrama del modelo conceptual

El diseño conceptual define el almacén de datos desde el punto de vista más abstracto y próximo al usuario, y sin vinculación con la tecnología final a utilizar.

2.3.1 Definición de las dimensiones

Definimos las siguientes dimensiones para nuestro Almacén de Datos:

- Dimensión Película que contendrá el conjunto de características de la película nominada o premiada.
- Dimensión Categoría, a la que está nominada.
- Dimensión Nominado contendrá el conjunto de nominados, bien sea persona o personas, o entidad nominada al premio.
- Dimensión Edición representará el conjunto de características relativas a la edición del Festival.

2.3.2 Definición de los atributos de cada dimensión

Definimos los atributos de cada dimensión:

2.3.2.1 Dimensión Película

- El atributo Original contendrá el nombre original de la película.

- El atributo País contendrá el nombre del país o de los países origen de la película.

2.3.2.2 Dimensión Nominado

- El atributo Nombre contendrá el nombre del nominado.

2.3.2.3 Dimensión Categoría

- El atributo Nombre contendrá el nombre de la categoría a la que pertenece la nominación.
- El atributo Categoría_Unificada, contendrá el nombre de la categoría unificada a la que pertenece la categoría. En caso de pertenecer a una categoría unificada concreta, se asignará la categoría a una categoría unificada genérica.

2.3.2.4 Dimensión Edición

- El atributo Año contendrá el año de la edición del festival.
- El atributo Número almacenará el número de edición del festival.
- El atributo Nombre contendrá el nombre del Festival

2.3.3 Definición del Hecho

Partiendo de que todos los datos representan nominaciones, y no candidatos a nominados, definiremos nuestro hecho, como "ser ganador o no". En esta tabla de hechos, interseccionan todas las dimensiones definidas anteriormente, mediante las cuales se podrá definir un solo hecho: solo existirá un nominado para una categoría determinada, en una película concreta y en la misma edición de premios.

2.3.4 Definición de las medidas

El Hecho tendrá como medida el campo *esGanador* que indicará si el nominado al premio es ganador o solamente nominado.

2.3.5 Diseño Conceptual

Representamos el diseño conceptual de nuestro modelo, siguiendo las características descritas anteriormente en un diagrama en forma de estrella:

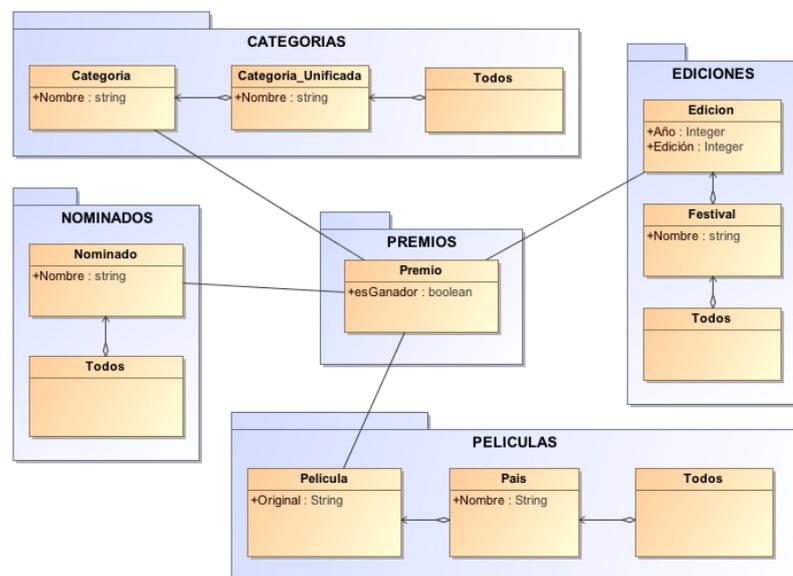


Figura 4 - Diseño conceptual

3 Diseño

3.1 Arquitectura de Hardware

Arquitectura de hardware final del proyecto. Se observan los siguientes elementos:

- Administrador/es del sistema: Equipo o equipos de los usuarios administradores del sistema.
- Usuarios del sistema: equipos que se conectarán con el repositorio de información y explotarán el DW a través de informes de consulta
- Servidor/es de ficheros Excel: Servidor o servidores de archivos que contendrán los documentos Excel, orígenes de los datos.
- Servidor de Oracle: Equipo servidor de base de datos que soportará las tareas de Extracción, Transformación y Carga de las fuentes de datos, así como su proceso y alojamiento bajo el SGBD de Oracle.
- Red de trabajo: red informática que proveerá de conectividad a los elementos involucrados, siendo igualmente válida independientemente de su alcance: WAN, LAN, VLAN, móvil,...

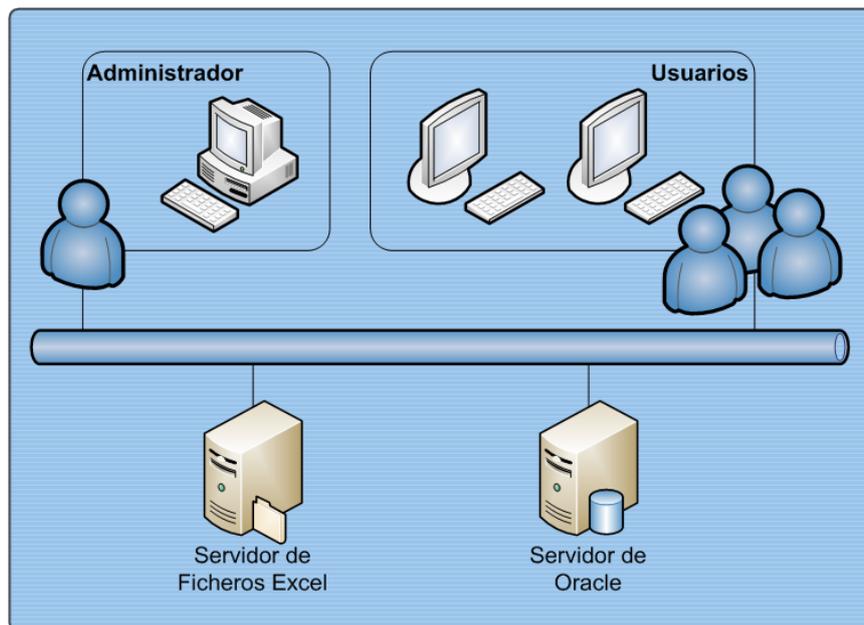


Figura 5 - Diagrama de arquitectura de hardware

3.2 Arquitectura de Software

Se observan los siguientes elementos:

- Ficheros Microsoft Excel: Ficheros fuente de los datos. El usuario administrador será el encargado de transformar los datos a formato .csv para una posterior lectura.
- Oracle SQL*Loader y Oracle SQL Developer: El usuario administrador será el encargado en ejecutar los procesos desarrollados en este proyecto y bajo estas herramientas.
- Oracle Database 10g XE: Sistema Gestor de Base de Datos que contendrá el almacén de datos.

- Oracle Business Intelligence Discoverer: esta herramienta dará acceso a los informes desarrollados para explotar la información contenida en el almacén de datos.

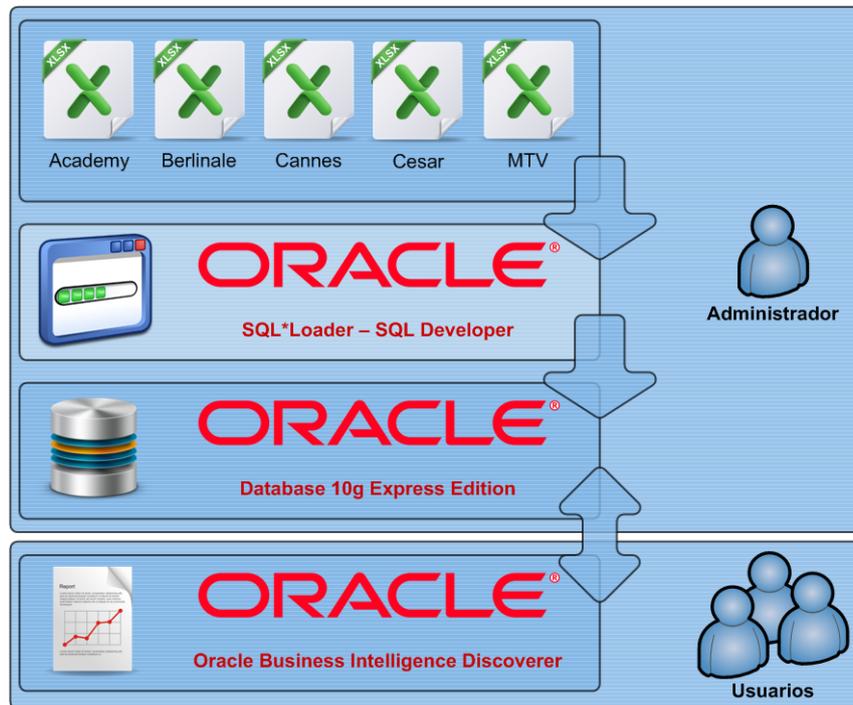


Figura 6 - Diagrama de arquitectura de software

3.3 Diseño Lógico

Partiendo del diseño conceptual definido anteriormente, creamos el diseño lógico de nuestro almacén de datos, donde identificamos las siguientes entidad y relaciones:

3.3.1 Entidades

Los valores entre paréntesis representan los campos, y el subrayado el/los campos clave de la entidad:

- *Pais* (IdPais, Nombre)
- *PeliculaPais* (IdPelicula, IdPais)
- *Pelicula* (IdPelicula, Original)
- *Festival* (IdFestival, Nombre)
- *Edición* (IdEdicion, Ano, IdFestival, Edicion)
- *CategoríaTipos* (IdTipo, Nombre)
- *Categoría* (IdCategoría, Nombre, IdTipo)
- *Nominado* (IdNominado, Nombre)
- *Premios* (IdPelicula, IdEdicion, IdCategoría, IdNominado, EsGanador)

3.3.2 Relaciones

- *FK_Pais_PeliculaPais*: relación 1:n entre *Pais.IdPais* y *PeliculaPais.IdPais*.
- *FK_Pelicula_PeliculaPais*: relación 1:n entre *Pelicula.IdPelicula* y *PeliculaPais.IdPelicula*.
- *FK_Edicion_Festival*: Relación 1:n entre *Festival.IdFestival* y *Edicion.IdFestival*.

- *FK_Premios_Pelicula*: Relación 1:n entre *Pelicula.IdPelicula* y *Premios.IdPelicula*.
- *FK_Premios_Edicion*: Relación 1:n entre *Edicion.IdEdicion* y *Premios.IdEdicion*.
- *FK_Premios_Categoria*: Relación 1:n entre *Categoria.IdCategoria* y *Premios.IdCategoria*.
- *FK_Premios_Nominado*: Relación 1:n entre *Nominado.IdNominado* y *Premios.IdNominado*.

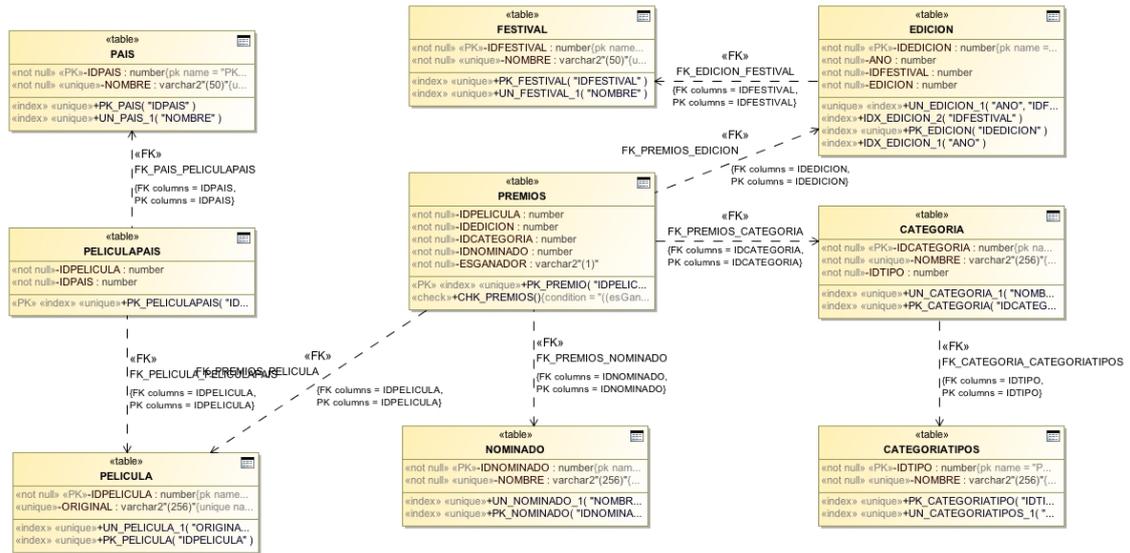


Figura 7 - Diseño lógico

3.4 Diseño Físico

La siguiente figura recoge el diseño físico de la base de datos relacional del proyecto. Podemos observar finalmente cómo en nuestro modelo, alguna de las dimensiones están formadas por varias tablas relacionadas, por lo tanto, el diseño corresponde a un modelo de datos en formato de copo de nieve.

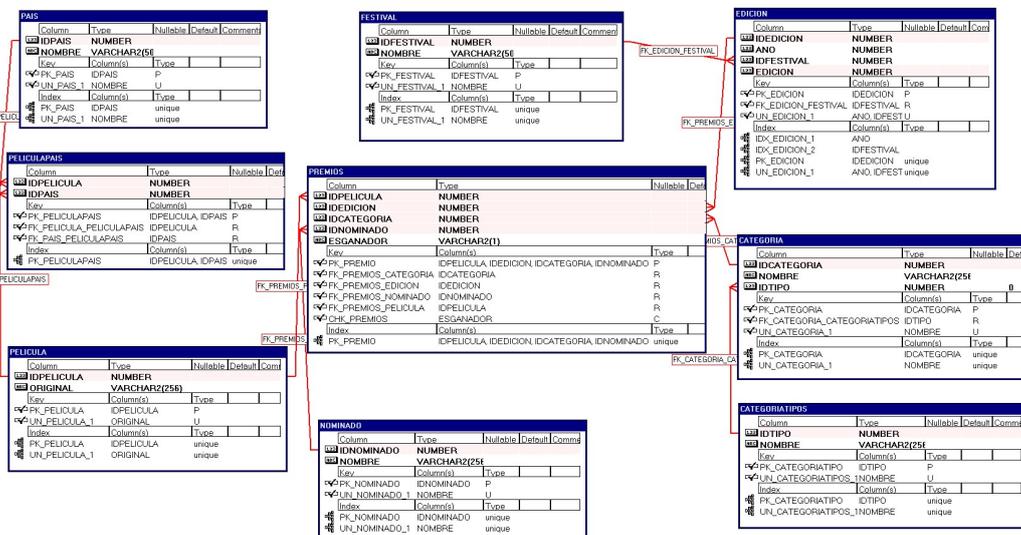


Figura 8 - Diseño físico

Se describe brevemente los objetos que forman la base de datos:

3.4.1 Tabla *Pelicula*

Formada por los siguientes objetos:

- Campo *idPelicula*. Tipo NUMBER. No Nulo. Contiene el identificador asignado a la película.
- Campo *Original*. Tipo VARCHAR2(256). No Nulo. Contiene el título original de la película.
- Clave primaria de la tabla *PK_IdPelicula*. Formada por el campo *idPelicula*.
- Índice único *UN_Pelicula_1*. Formado por el campo *Original*.

3.4.2 Tabla *Pais*

Formada por los siguientes objetos:

- Campo *idPais*. Tipo NUMBER. No Nulo. Contiene el identificador asignado al País.
- Campo *Nombre*. Tipo VARCHAR2(50). No Nulo. Contiene el nombre del País.
- Clave primaria *PK_Pais*. Formada por el campo *idPais*.
- Índice único *UN_Pais_1*. Formado por el campo *Nombre*.

3.4.3 Tabla *PeliculaPais*

Formada por los siguientes objetos:

- Campo *idPelicula*. Tipo NUMBER. No Nulo. Contiene el identificador asignado a la Película.
- Campo *idPais*. Tipo NUMBER. No Nulo. Contiene el identificador asignado al País.
- Clave primaria *PK_PeliculaPais*. Formada por ambos campos.
- Clave foránea *FK_Pais_PeliculaPais*. Relaciona la tabla *PeliculaPais* con *Pais* mediante el campo *idPais*.
- Clave foránea *FK_Pelicula_PeliculaPais*. Relaciona la tabla *PeliculaPais* con *Pelicula* mediante el campo *idPelicula*.

3.4.4 Tabla *Nominado*

Formada por los siguientes objetos:

- Campo *idNominado*. Tipo NUMBER. No Nulo. Contiene el identificador asignado al nominado.
- Campo *Nombre*. Tipo VARCHAR2(256). No Nulo. Contiene el nombre del nominado.
- Clave primaria *PK_Nominado*. Formada por el campo *idNominado*.
- Índice único *UN_Nominado_1*. Formado por el campo *Nombre*.

3.4.5 Tabla *Categoria*

Formada por los siguientes objetos:

- Campo *idCategoria*. Tipo NUMBER. No Nulo. Contiene el identificador asignado a la categoría.
- Campo *idTipo*. Tipo NUMBER. No Nulo. Valor por defecto 0. Contiene el identificador de la Categoría única a la que pertenece.
- Campo *Nombre*. Tipo VARCHAR2(256). No Nulo. Contiene el nombre de la categoría.
- Clave primaria *PK_Categoria*. Formada por el campo *Categoria*.
- Índice único *UN_Categoria_1*. Formado por el campo *Nombre*.

- Clave foránea *FK_Categoria_CategoriaTipos*. Relaciona la tabla *Categoria* con *CategoriaTipos* mediante el campo *idTipo*.

3.4.6 Tabla *CategoriaTipos*

Formada por los siguientes objetos:

- Campo *idTipo*. Tipo NUMBER. No Nulo. Contiene el identificador asignado a la categoría tipo o categoría unificada.
- Campo *Nombre*. Tipo VARCHAR2(256). No Nulo. Contiene el nombre de la categoría unificada.
- Clave primaria *PK_CategoriaTipo*. Formada por el campo *idTipo*.
- Índice único *UN_CategoriaTipos_1*. Formado por el campo *Nombre*.

3.4.7 Tabla *Edicion*

Formada por los siguientes objetos:

- Campo *idEdicion*. Tipo NUMBER. No Nulo. Contiene el identificador asignado a la edición.
- Campo *Edicion*. Tipo NUMBER. No Nulo. Contiene el número de edición para la relación entre *idFestival* y *Ano*.
- Campo *Ano*. Tipo NUMBER. No Nulo. Contiene el año de la edición.
- Campo *idFestival*. Tipo NUMBER. No Nulo. Contiene el identificador del Festival al que hace referencia la edición.
- Clave primaria *PK_Edicion*. Formada por el campo *idEdicion*.
- Clave foránea *FK_Edicion_Festival*. Relaciona la tabla *Festival* con *Edicion* mediante el campo *idFestival*.
- Índice único *UN_Edicion_1*. Formado por los campos *Ano*, *idFestival* y *Edicion*.
- Índice *IDX_Edicion_1*. Formado por el campo *Ano*.
- Índice *IDX_Edicion_2*. Formado por el campo *idFestival*.

3.4.8 Tabla *Festival*

Formada por los siguientes objetos:

- Campo *idFestival*. Tipo NUMBER. No Nulo. Contiene el identificador asignado al Festival.
- Campo *Nombre*. Tipo VARCHAR2(50). No Nulo. Contiene el nombre del Festival.
- Clave primaria *PK_Festival*. Formada por el campo *idFestival*.
- Índice único *UN_Festival_1*. Formado por el campo *Nombre*.

3.4.9 Tabla *Premios*

Formada por los siguientes objetos:

- Campo *idPelicula*. Tipo NUMBER. No Nulo. Contiene el identificador de la película.
- Campo *idEdicion*. Tipo NUMBER. No Nulo. Contiene el identificador de la edición.
- Campo *idCategoria*. Tipo NUMBER. No Nulo. Contiene el identificador de la categoría.
- Campo *idNominado*. Tipo NUMBER. No Nulo. Contiene el identificador del nominado.
- Campo *esGanador*. Tipo VARCHAR2(1). No Nulo. Contiene valores 'S' o 'N' indicando si el registro es ganador o no.

- Clave primaria *PK_PREMIO*. Formada por los campos *idPelicula*, *idEdicion*, *idCategoria* e *idNominado*.
- Clave foránea *FK_Premios_Pelicula*. Relaciona la tabla *Pelicula* con *Premios* mediante el campo *idPelicula*.
- Clave foránea *FK_Premios_Edicion*. Relaciona la tabla *Edicion* con *Premios* mediante el campo *idEdicion*.
- Clave foránea *FK_Premios_Categoria*. Relaciona la tabla *Categoria* con *Premios* mediante el campo *idCategoria*.
- Clave foránea *FK_Premios_Nominado*. Relaciona la tabla *Nominado* con *Premios* mediante el campo *idNominado*.
- Restricción check *CHK_Premios*. Restringe el valor almacenado en el campo *esGanador* a 'S' o 'N'.

3.4.10 Tabla temporal *Temp*

Tabla temporal utilizada en los procesos ETL. En esta tabla se almacenan todos los datos extraídos de las fuentes operacionales. Formada por los siguientes objetos:

- Campo *Original*. Tipo VARCHAR2(500). Permite Nulo.
- Campo *English*. Tipo VARCHAR2(500). Permite Nulo.
- Campo *Sort*. Tipo VARCHAR2(500). Permite Nulo.
- Campo *Year*. Tipo VARCHAR2(500). Permite Nulo.
- Campo *Country*. Tipo VARCHAR2(500). Permite Nulo.
- Campo *Awards*. Tipo VARCHAR2(500). Permite Nulo.
- Campo *Category*. Tipo VARCHAR2(500). Permite Nulo.
- Campo *Winner*. Tipo VARCHAR2(500). Permite Nulo.
- Campo *Nominee*. Tipo VARCHAR2(500). Permite Nulo.
- Campo *Academy*. Tipo VARCHAR2(500). Permite Nulo.
- Campo *Item*. Tipo VARCHAR2(500). Permite Nulo.
- Campo *Error*. Tipo VARCHAR2(500). Permite Nulo.
- Campo *IdRow*. Tipo NUMBER. No Nulo.
- Índice *IDX_Temp_1*. Formado por los campos *Original*, *Year*, *Awards*, *Academy*, *Category* y *Nominee*.

3.4.11 Tabla temporal *ETLPaises*

Tabla utilizada en el proceso de Transformación y Carga en el modelo final. Usada para convertir los países con nombres en diferentes idiomas en único.

Formada por los siguientes objetos:

- Campo *PaisOrigen*. Tipo VARCHAR2(100). No Nulo.
- Campo *PaisDestino*. Tipo VARCHAR2(100). No Nulo.
- Clave primaria *PK_ETLPAISES*. Formada por el campo *PaisOrigen*

3.4.12 Otros objetos

Otros objetos de la base de datos no listados:

- Secuencia *SEQ_idPais*: Usada para la generación de código auto numérico para asignación al campo *Pais.idPais*.
- Secuencia *SEQ_idPelicula*: Usada para la generación de código auto numérico para asignación al campo *Pelicula.idPelicula*.
- Secuencia *SEQ_idNominado*: Usada para la generación de código auto numérico para asignación al campo *Nominado.idNominado*.
- Secuencia *SEQ_idCategoria*: Usada para la generación de código auto numérico para asignación al campo *Categoria.idCategoria*.
- Secuencia *SEQ_idEdicion*: Usada para la generación de código auto numérico para asignación al campo *Edicion.idEdicion*.
- Secuencia *SEQ_idFestival*: Usada para la generación de código auto numérico para asignación al campo *Festival.idFestival*.
- Paquete *misFunciones*: Funciones y procedimientos utilizados en el proceso de transformación y carga de los datos.

3.5 Procesos ETL

Clasificamos los procesos en los 3 subprocesos lógicos:

3.5.1 Procesos de Extracción

El proceso de extracción toma como datos de origen los ficheros de fuentes operaciones en formato CSV, y mediante la herramienta SQL*Loader, importa dichos datos en la tabla de trabajo *TEMP*. La importación se realiza mediante los siguientes procesos BATCH y ficheros de configuración.

El proceso BATCH carga, de uno en uno, los ficheros CSV gracias a los ficheros de control y configuración CTL, uno por cada fuente operacional.

En este fichero de configuración, se indica que el proceso debe leer el fichero CSV correspondiente de datos de origen y añadir los datos en la tabla *Temp*. También se indica al proceso, que los campos en el fichero de origen estarán separados por ';' y los valores, opcionalmente, pueden venir calificados entre '"' (comilla doble).

De igual manera, se indica al proceso el orden de inserción de los valores y la utilización de un valor numérico secuencial e incrementado de 1 en 1 a partir del máximo valor existente para el campo *Temp.idRow*.

3.5.2 Procesos de Transformación

Para realizar el proceso de transformación de los datos se utiliza un script que, de forma masiva, prepara los datos para cargarlos en el modelo final.

Este script, realiza las siguientes actualizaciones:

3.5.2.1 Campo Original

- Elimina los posibles espacios en blanco existentes antes y después del valor.

- Actualiza el valor a mayúsculas.
- En caso de ser valor nulo, actualiza dicho valor con el del campo english.
- En caso de seguir siendo nulo, actualiza el valor a 'NO SPECIFIC FILM TITLE'.
- Reemplaza la cadena " : " por ": ".
- Reemplaza la cadena "&" por "AND"
- Reemplaza la cadena ", AND", por " AND".

3.5.2.2 Campo Year

- Actualiza su valor con la subcadena existente a partir del carácter "-"

3.5.2.3 Campo Country

- Convertimos el valor a mayúsculas
- Actualizamos el valor a "NO COUNTRY" para todos aquellos con valor nulo.
- Se reemplaza la cadena ",," por "/".

3.5.2.4 Campo Awards

- Convertimos el valor a mayúsculas.

3.5.2.5 Campo Category

- Convertimos el valor a mayúsculas.
- Elimina los posibles espacios en blanco existentes antes y después del valor.

3.5.2.6 Campo Winner

- Actualizamos el valor según la siguiente lógica:
 - o Si el valor es "X" lo actualizamos a "S"
 - o Si el valor es "0" lo actualizamos a "N"
 - o Para todos los demás casos, lo actualizamos a "N"

3.5.2.7 Campo Nominee

- Convertimos el valor a mayúsculas.
- Elimina los posibles espacios en blanco existentes antes y después del valor.
- Reemplazamos el valor "" (comillas dobles) por valor nulo
- Reemplazamos el valor a "NO NOMINEE" para todos los valores nulos.

3.5.2.8 Campo Academy

- Convertimos el valor a mayúsculas.
- Elimina los posibles espacios en blanco existentes antes y después del valor.
- Reemplazamos en los valores la cadena "ÈRE NUIT DES CÉSARS" por cadena vacía.
- Reemplazamos en los valores la cadena "ÈME NUIT DES CÉSARS" por cadena vacía.
- Reemplazamos en los valores la cadena "ST" por cadena vacía.
- Reemplazamos en los valores la cadena "ND" por cadena vacía.
- Reemplazamos en los valores la cadena "RD" por cadena vacía.
- Reemplazamos en los valores la cadena "TH" por cadena vacía.

3.5.3 Procesos de Carga

El script de carga se encargará de volcar los datos sobre el modelo físico y relacional. El proceso se realiza de manera ordenada, cargando primeramente la tablas básicas de datos y posteriormente las tablas relacionadas con otras, siendo la ultima tabla a carga la tabla de hechos. De esta forma, se elimina el riesgo de incidencias relacionadas con la ausencia de valores en las claves foráneas.

Para los procesos de carga más complejos se utilizan procedimientos (*PROCEDURE*) definidos en el paquete de programación *misFunciones*. Podremos dividir el script de carga en fases:

3.5.3.1 Carga de datos en la tabla *ETLPaises*

Mediante procesos semi-automatizados, se ha creado un script que contiene 68 sentencias de inserción sobre la tabla *ETLPaises*, donde se relaciona el nombre origen de los países en diversos idiomas con el nombre destino o final del país. De esta forma, logramos identificar el país con un nombre unívoco e independiente del idioma utilizado.

Dado que este proceso de relación no puede automatizarse completamente, el mantenimiento de esta tabla deberá realizarse manualmente.

3.5.3.2 Carga de datos en la tabla *CategoriaTipos*

Mediante un proceso manual, se ha creado un script que contiene 15 sentencias de inserción en la tabla *CategoriaTipos*, con la relación de categorías unificadas de los premios. Dado que este proceso de creación y definición de categorías tipo no puede automatizarse, el mantenimiento y actualización de dicha tabla deberá realizarse manualmente.

3.5.3.3 Carga de datos en la tabla *Pelicula*

Se cargan en el sistema solamente aquellas películas que no existan. Como identificador de la película se asigna el siguiente valor de la secuencia *SEQ_IDPELICULA*.

3.5.3.4 Carga de datos en la tabla *Pais*

El script hace una llamada al procedimiento *misfunciones.cargaPaises*, el cual realiza el siguiente proceso:

Trocea el valor del campo *Country* a partir del carácter "/" con el fin de obtener los posibles países para cada uno de los registros de la tabla *Temp*.

Para cada uno de los valores obtenidos y mediante la función *misfunciones.damePaisDestino* y la tabla *ETLPaises*, se realiza la transformación del nombre del país a un nombre de país unívoco. El nombre del país resultante será cargado en el sistema siempre que no exista previamente.

3.5.3.5 Carga de datos en la tabla *PeliculaPais*

El script hace una llamada al procedimiento *misfunciones.cargaPeliculaPais*, el cual realiza el siguiente proceso:

Trocea el valor del campo *Country* a partir del carácter "/", para obtener los posibles países para cada uno de los registros de la tabla *Temp*.

Para cada uno de estos valores, y al igual que el caso anterior, sirviéndonos de la función *misfunciones.damePaisDestino*, se obtiene el identificador correspondiente utilizando la función *misfunciones.dameIdPais*.

Así mismo, y mediante el valor del campo *Original* del registro afectado, obtenemos el identificador de película mediante la función *misfunciones.dameIdPelicula*.

A partir de ambos identificadores, insertamos el registro solamente si esta relación no existe previamente.

3.5.3.6 Carga de datos en la tabla *Nominado*

Se cargan solamente aquellos nominados que no existen en el sistema. Como identificador de Nominado se asigna el siguiente valor de la secuencia *SEQ_IDNOMINADO*.

3.5.3.7 Carga de datos en la tabla *Festival*

Se cargan solamente aquellos festivales que no existen en el sistema. Como identificador de Festiva se asigna el siguiente valor de la secuencia *SEQ_IDFESTIVAL*.

3.5.3.8 Carga de datos en la tabla *Edicion*

Se carga únicamente aquellas ediciones que no existen en el sistema. Como identificador de Edicion se asigna el siguiente valor de la secuencia *SEQ_IDEDICION*.

3.5.3.9 Carga de datos en la tabla *Categoria*

Se carga únicamente aquellas categorías que no existen en el sistema. Como identificador de Categoría se asigna el siguiente valor de la secuencia *SEQ_IDCATEGORIA*.

3.5.3.10 Carga de datos en la tabla de hechos *Premios*

El script hace una llamada al procedimiento *misfunciones.cargaPremios*. Este procedimiento realizará las siguientes operaciones por cada uno de los registros existentes en la tabla *Temp*:

Obtendrá el identificador *idPelicula* a partir del valor del campo *Original* mediante la función *misfunciones.dameIdPelicula*.

Obtendrá el identificador *idEdicion* a partir de los valores de los campos *Year*, *Awards* y *Academy* mediante la función *misfunciones.dameIdEdicion*.

Obtendrá el identificador *idCategoria* a partir del valor del campo *Category* mediante la función *misfunciones.dameIdCategoria*.

Obtendrá el identificador *idNominado* a partir del valor del campo *Nominee* mediante la función *misfunciones.dameIdNominado*.

A partir de dichos identificadores, junto con el valor del campo *Winner*, abrirá una transacción y se realizará una inserción en la tabla *Premios*. Posteriormente el registro afectado de la tabla *Temp* será eliminado y desechado para posibles nuevos procesos de carga. Si todo fue bien, la transacción será confirmada.

En caso de error, se registrara el valor de la variable de sistema *SQLERRM* en el campo *Error* de la tabla *Temp* del registro afectado y de realizará *ROLLBACK* de la transacción activa, quedan así registrado el error provocado para su posible depuración.

3.5.3.11 Actualización de datos en la tabla *Categoria*

Mediante procesos manuales, se ha creado un script que contienen 42 sentencias de actualización de la tabla *Categoria*. Estas sentencias sobrescriben el valor del campo *idTipo* con el valor del campo clave *CategoriaTipo.*, relacionando dichas tablas y de esta forma, asignando una categoría unificada a las categorías existentes.

Dado que este proceso de relación no puede automatizarse completamente, el mantenimiento de dicha unión deberá realizarse manualmente.

3.5.4 Otros procesos

Tras la carga e incorporación de los datos sobre el modelo físico de nuestro Data Warehouse, se lanzarán los procesos de estimación y cálculo de estadísticas, cuya finalidad es recopilar y actualizar el catálogo de datos estadísticos del motor de base de datos.

Mediante estas estadísticas, Oracle dispondrá de los datos necesarios para elaborar el plan de ejecución más óptimo durante la ejecución de las sentencias, y por lo tanto, se mejorará el rendimiento a la hora de ejecutar los informes de explotación de nuestra base de datos.

3.6 Entorno de Trabajo

3.6.1 Construcción del Modelo con Oracle® BI Discoverer Administrator

Una vez llegada esta fase, ya contamos con el modelo físico relacional donde se alojan, de forma coherente y estructurada, los datos que vamos a explotar con los informes. Pero previamente a la fase de desarrollo de informes, debemos instalar el entorno y escenario de trabajo de Business Intelligence y su área de negocio.

3.6.2 Instalación y creación de Capa de Usuario Final (EUL)

El EUL es la capa de funcionalidades que abstrae al usuario final de las tareas y procesos necesarios para el correcto funcionamiento de BI Discoverer. Esta capa está formada por tablas, vistas, funciones y metadatos que deberá ser instalada bajo nuestro esquema de datos de Oracle XE.

3.6.3 Carga y diseño del Área de Negocio

Después de la instalación del entorno EUL, debemos crear un área de negocio. El proceso de creación o carga, se realizará a través de un sencillo asistente.



Figura 9 - Asistente de carga: Paso 2

Selección del usuario/s cuyos objetos formarán parte del área de negocio.

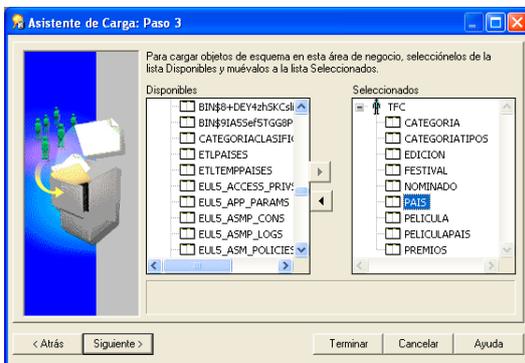


Figura 10 - Asistente de carga: Paso 3

Selección final de los objetos que formarán parte del área de negocio.

Durante la creación del Área de Negocio, concretamente en el paso 4 del asistente, indicaremos al proceso se deben crear las uniones a partir de las restricciones de clave primaria/ajena.

Una vez terminado el asistente, el espacio de trabajo generado será similar al que observamos en la imagen lateral.

3.6.4 Configuración del Área de Negocio

Realizaremos la siguiente configuración en nuestro Área de Negocio como paso previo al desarrollo de los informes:

- Ocultaremos los elementos no necesarios.
- Seleccionaremos *detalle* como valor por defecto para los campos *Edicion.Año* y *Edicion.Edicion*.
- Creamos la carpeta personalizada *Pelicula_Pais*.
- Creamos la carpeta personalizada *PremiosEntreCategorias*.
- Creamos las siguientes clases de elementos:
 - o *Países*
 - o *Año*
 - o *Edición*
 - o *Festival*
 - o *Categoría*
 - o *Categoría Unificada*
 - o *Es ganador*
- Creamos las siguientes jerarquías que relacionen los diferentes elementos:
 - o *Categoría Unificada*
 - o *Festival_Edicion*
 - o *Año_Festival*
 - o *Pelicula_Pais*

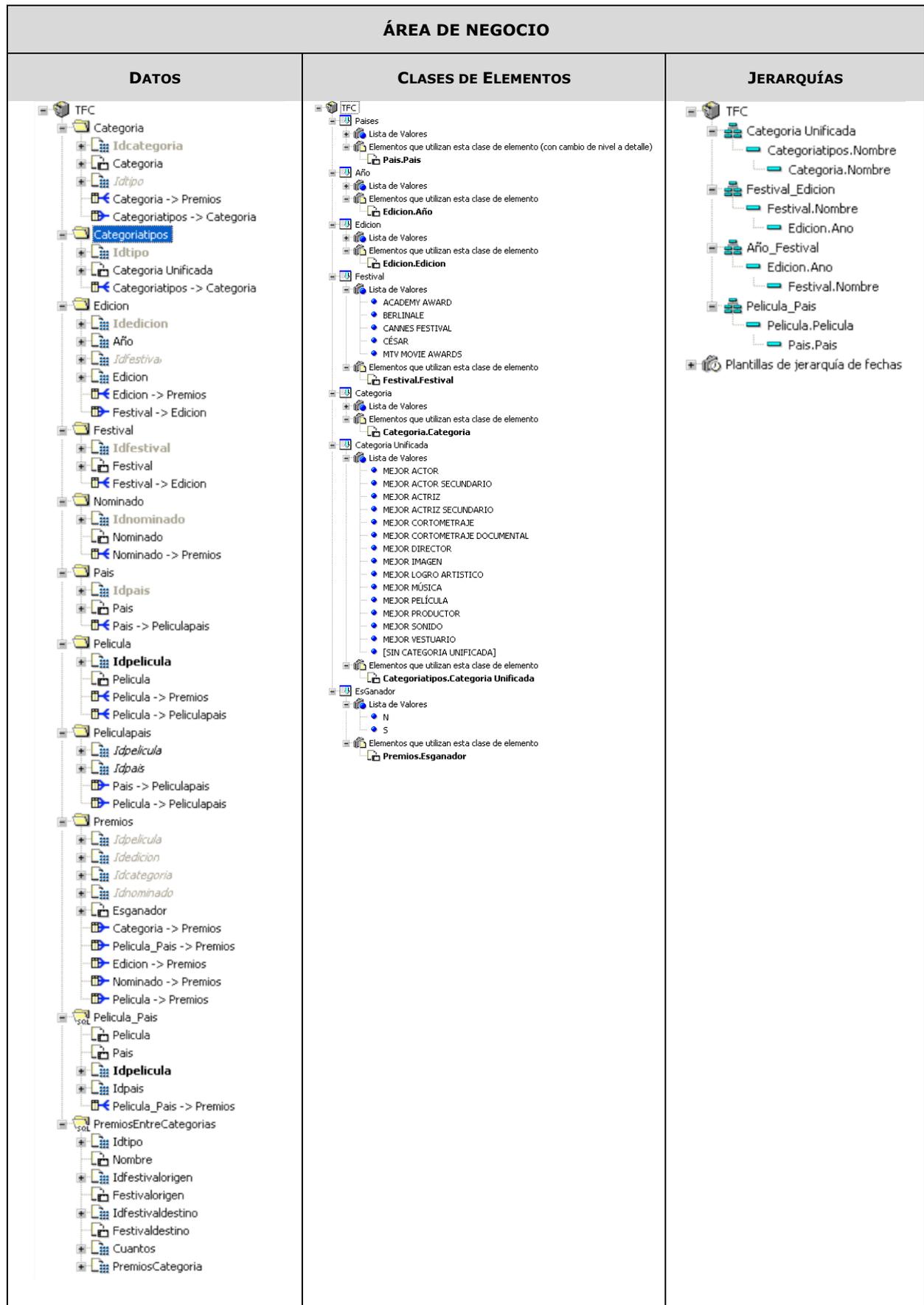


Figura 11 - Área de Negocio: Datos, clases de elementos y jerarquías

3.7 Construcción de los Informes con Oracle® BI Discoverer Desktop

3.7.1 Informe 1: Premios por actor o actriz, año y festival

Se crea un informe en formato tabla página-detalle, donde se muestra el número de premios por actor o actriz por año y festival. Utilizamos las clases de elementos "Categoría Unificada" y "es Ganador" para filtrar los resultados.

Se crea un nuevo informe con formato matriz y utilizando los mismos criterios de filtro.

Nominado	Año	Festival	Cuantos Premios
ADRIEN BRODY	2002	ACADEMY AWARD	1
		CÉSAR	1
			Recuento: 2
AGNÈS JAQUI	1997	CÉSAR	1
			Recuento: 1
AL PACINO	1992	ACADEMY AWARD	1
			Recuento: 1
ALAIN DELON	1984	CÉSAR	1
			Recuento: 1
ALAN ARKIN	2006	ACADEMY AWARD	1
			Recuento: 1
ALBERT FINNEY	1984	BERLINALE	1
			Recuento: 1
ALBERTO SORDI	1972	BERLINALE	1
			Recuento: 1
ALEC GUINNESS	1957	ACADEMY AWARD	1
			Recuento: 1
ALFREDO LANDA, FRANCISCO RABAL	1984	CANNES FESTIVAL	1
			Recuento: 1
ALICE BRADY	1937	ACADEMY AWARD	1
			Recuento: 1

Figura 12 - Informe 1

3.7.2 Informe 2: Premios por actor o actriz, año, festival y categoría

Se crea un informe en formato tabla página-detalle, donde se muestran el número de premios conseguidos por un actor o actriz por año, festival y categoría. A diferencia del anterior informes, y únicamente con motivo de cambiar de método, utilizamos condiciones con operadores *LIKE* e "igual" para filtrar los resultados.

Al igual que el informe anterior, se crear un informe paralelo con formato matriz utilizando los mismos criterios de filtro.

Nominado	Año	Festival	Categoría	Cuantos Premios
ADRIEN BRODY	2002	ACADEMY AWARD	BEST ACTOR	1
		CÉSAR	MEILLEUR ACTEUR	1
				Total: 2
AGNÈS JAQUI	1997	CÉSAR	MEILLEURE ACTRICE DANS UN SECOND RÔLE	1
				Total: 1
AL PACINO	1992	ACADEMY AWARD	BEST ACTOR	1
				Total: 1
ALAIN DELON	1984	CÉSAR	MEILLEUR ACTEUR	1
				Total: 1
ALAN ARKIN	2006	ACADEMY AWARD	BEST SUPPORTING ACTOR	1
				Total: 1
ALBERT FINNEY	1984	BERLINALE	SILBERNER BÄR FÜR DEN BESTEN DARSTELLER	1
				Total: 1
ALBERTO SORDI	1972	BERLINALE	SILBERNER BÄR FÜR DEN BESTEN DARSTELLER	1
				Total: 1
ALEC GUINNESS	1957	ACADEMY AWARD	BEST ACTOR	1
				Total: 1
ALFREDO LANDA, FRANCISCO RABAL	1984	CANNES FESTIVAL	BEST ACTOR	1
				Total: 1
ALICE BRADY	1937	ACADEMY AWARD	BEST SUPPORTING ACTRESS	1
				Total: 1
ANA BEATRIZ NOGUEIRA	1986	BERLINALE	SILBERNER BÄR FÜR DIE BESTE DARSTELLERIN	1
				Total: 1
ANDRÉ DUSSOLLIER	1992	CÉSAR	MEILLEUR ACTEUR DANS UN SECOND RÔLE	1
	1997	CÉSAR	MEILLEUR ACTEUR	1
	2001	CÉSAR	MEILLEUR ACTEUR DANS UN SECOND RÔLE	1
				Total: 3

Figura 13 - Informe 2

3.7.3 Informe 3: Premios por director, año, festival y película.

Se crear un informe en formato tabla página-detalle, que muestre el número de premios conseguidos por director, año, festival y película. Se utilizan las clases de elementos "Categoría Unificada" y "es Ganador" para filtrar los resultados.

Al igual que en ocasiones anteriores, se crea un informe alternativo en formato matriz y utilizando los mismos criterios para filtrar los resultados.

Nominado	Año	Festival	Película	Cuantos Premios
ABDEL KECHICHE	2003	CÉSAR	L'ESQUIVE	1
				Suma: 1
AKIRA KUROSAWA	1959	BERLINALE	KAKUSHI TORIDE NO SAN-AKUNIN	1
				Suma: 1
ALAIN CAVALIER	1986	CÉSAR	THÉRÈSE	1
				Suma: 1
ALAIN CORNEAU	1991	CÉSAR	TOUS LES MATINS DU MONDE	1
				Suma: 1
ALAIN RESNAIS	1977	CÉSAR	PROVIDENCE	1
	1993	CÉSAR	SMOKING/NO SMOKING	1
				Suma: 2
ANDRÉ TÉCHINÉ	1994	CÉSAR	LES ROSEAUX SAUVAGES	1
				Suma: 1
ANDREW BIRKIN	1992	BERLINALE	THE CEMENT GARDEN	1
				Suma: 1
ANDRZEJ WAJDA	1983	CÉSAR	DANTON	1
				Suma: 1
ANG LEE	2005	ACADEMY	BROKEBACK MOUNTAIN	1
				Suma: 1
ANTHONY MINGHELLA	1996	ACADEMY	THE ENGLISH PATIENT	1
				Suma: 1

Figura 14 - Informe 3

3.7.4 Informe 4: Premios por año, festival, categoría y película.

Se crear un informe en formato tabla página-detalle, que muestre el número de premios conseguidos por película, año, festival y categoría. Se utilizan la clase de "es Ganador" para filtrar los resultados.

Película	Año	Festival	Categoría	
A LITTLE NIGHT MUSIC	1977	ACADEMY AWARD	BEST MUSIC, ADAPTATION / ORIGINAL SONG SCORE	Total: 1
A LITTLE ROMANCE	1979	ACADEMY AWARD	BEST MUSIC, ORIGINAL SCORE	Total: 1
A MAN AND A WOMAN	1966	ACADEMY AWARD	BEST WRITING, ORIGINAL SCREENPLAY	Total: 1
A MAN FOR ALL SEASONS	1966	ACADEMY AWARD	BEST WRITING, ADAPTED SCREENPLAY	
	1966	ACADEMY AWARD	BEST PICTURE	
	1966	ACADEMY AWARD	BEST DIRECTOR	
	1966	ACADEMY AWARD	BEST COSTUME DESIGN, COLOR	
	1966	ACADEMY AWARD	BEST CINEMATOGRAPHY, COLOR	
	1966	ACADEMY AWARD	BEST ACTOR	Total: 6
A MIDSUMMER NIGHT'S DREAM	1935	ACADEMY AWARD	BEST FILM EDITING	
	1935	ACADEMY AWARD	BEST CINEMATOGRAPHY	Total: 2
A NOS AMOURS	1983	CÉSAR	MELLEUR FILM	
	1983	CÉSAR	MELLEUR ESPOIR FÉMININ	Total: 2
A NOTE OF TRIUMPH: THE GOLDEN AGE	2005	ACADEMY AWARD	BEST DOCUMENTARY, SHORT SUBJECT	Total: 1
A PASSAGE TO INDIA	1984	ACADEMY AWARD	BEST SUPPORTING ACTRESS	
	1984	ACADEMY AWARD	BEST MUSIC, ORIGINAL SCORE	Total: 2

Figura 15 - Informe 4

3.7.5 Informe 5: Número de ediciones del festival, total de actores nominados, total de actrices nominadas y total de películas nominadas por tipo de festival, dada una fecha a seleccionar.

Este informe, aunque podía haberse construido mediante varias selecciones unidas en un sentencias común con el fin de mostrar todos los datos en una misma tabla página-detalle, he preferido construir cuatro tablas página-detalle y presentarlas en diferentes pestañas, en el mismo informe final, aunque separadas entre si.

Por este motivo, el informe presenta cuatro pestañas:

- Número de Ediciones por tipo de festival y año.
- Número de Películas por tipo de festival y año.

- Número de actores por tipo de festival y año.
- Número de actrices por tipo de festival y año.

Para filtrar los datos, he utilizado la clase de elementos "Categoría Unificada" y "es Ganador".

Elementos de Página: Año: 1.985	
	Cuantas Ediciones
ACADEMY AWARD	1
BERLINALE	2
CANNES FESTIVAL	1
CÉSAR	2

Figura 16 - Informe 5: Número de ediciones

Elementos de Página: Año: 1983	
	Cuantas Películas
ACADEMY AWARD	46,00
BERLINALE	8,00
CANNES FESTIVAL	10,00
CÉSAR	33,00

Figura 17 - Informe 5: Número de películas

Elementos de Página: Año: 2001	
	Cuantos Actores
ACADEMY AWARD	10
BERLINALE	1
CANNES FESTIVAL	1
CÉSAR	10

Figura 18 - Informe 5: Número de actores

Elementos de Página: Año: 1998	
	Cuantas Actrices
ACADEMY AWARD	10
BERLINALE	1
CANNES FESTIVAL	1
CÉSAR	10

Figura 19 - Informe 5: Número de actrices

3.7.6 Informe 6: Premios por año, película y festival.

Se crear un informe en formato tabla página-detalle que muestra el número de premios ganados por película y festival dado un año. Mostramos un subtotal de premios ganados por año y película.

Utilizo la clase de elementos "es Ganador" junto con un operador "distinto de" ('<>') para eliminar de los resultados aquellos premios sin película definida.

Película	Festival	Cuantos Premios
A PASSAGE TO INDIA	ACADEMY AWARD	2
		Suma: 2
AMADEUS	ACADEMY AWARD	8
	CÉSAR	1
		Suma: 9
ANOTHER COUNTRY	CANNES FESTIVAL	1
		Suma: 1
CAL	CANNES FESTIVAL	1
		Suma: 1
CARMEN	CÉSAR	1
		Suma: 1
CHARADE	ACADEMY AWARD	1
		Suma: 1
CYKELSYMPFONIEN	BERLINALE	1
		Suma: 1
FLIRT	BERLINALE	1
		Suma: 1

Figura 20 - Informe 6

3.7.7 Informe 7: Definición de "película exitosa" y "película fracasada", por año.

Antes de crear el informe, debemos proveer una definición a película exitosa y película fracasada:

- Película exitosa: Aquella película con más premios ganados. Ante igual número de premios ganados, será más exitosa cuantas más nominaciones tenga.
- Película fracasada: Aquella película con menor número de premios y mayor diferencia entre nominaciones y premios ganados.

Para obtener este informe, crearemos dos coeficientes en función de las definiciones anteriores y crearemos dos pestañas, una por definición, con dos tablas pagina-detalle ordenados de mayor a menor coeficiente por película para un año dado.

Además, nos serviremos de dos funciones definidas en el paquete de funciones *misFunciones*:

- función *CuantasNominaciones*: devuelve el número de nominaciones dado un año y el identificador de la película.
- función *CuantosGanados*: devuelve el número de premios ganados dado un año y el identificador de la película.

Los coeficientes de éxito y fracaso se obtienen del cálculo:

- $\text{Éxito} = ((\text{núm. premios} * 10) + \text{núm de nominaciones})$
- $\text{Fracaso} = ((\text{núm. nominaciones} - \text{núm. premios}) + (\frac{1}{\text{núm. nominaciones}}))$

	Película	Nominaciones	Premios	Éxito
▶ 1	THE DEER HUNTER	9,00	5,00	59,00
▶ 2	COMING HOME	9,00	4,00	49,00
▶ 3	L'ALBERO DEGLI ZOCCOLI	3,00	3,00	33,00
▶ 4	MIDNIGHT EXPRESS	6,00	2,00	26,00
▶ 5	L'ARGENT DES AUTRES	5,00	2,00	25,00
▶ 6	MOLIÈRE	5,00	2,00	25,00
▶ 7	VIOLETTE NOZIÈRE	5,00	2,00	25,00
▶ 8	LE DOSSIER 51	4,00	2,00	24,00
▶ 9	PRÉPAREZ VOS MOUCHOIRS	2,00	2,00	22,00
▶ 10	UNE HISTOIRE SIMPLE	11,00	1,00	21,00
▶ 11	HEAVEN CAN WAIT	9,00	1,00	19,00
▶ 12	AN UNMARRIED WOMAN	4,00	1,00	14,00
▶ 13	DAYS OF HEAVEN	4,00	1,00	14,00
▶ 14	SUPERMAN	4,00	1,00	14,00
▶ 15	THE BUDDY HOLLY STORY	3,00	1,00	13,00
▶ 16	CALIFORNIA SUITE	3,00	1,00	13,00
▶ 17	L'ETAT SAUVAGE	2,00	1,00	12,00
▶ 18	THE SHOUT	1,00	1,00	11,00
▶ 19	THE FLIGHT OF THE GOSSAMER CONDOR	1,00	1,00	11,00

Figura 21 - Informe 7: Películas exitosas

	Película	Nominaciones	Premios	Fracaso
▶ 1	UNE HISTOIRE SIMPLE	11,00	1,00	10,09
▶ 2	HEAVEN CAN WAIT	9,00	1,00	8,11
▶ 3	INTERIORS	5,00	0,00	5,20
▶ 4	LE SUCRE	5,00	0,00	5,20
▶ 5	COMING HOME	9,00	4,00	5,11
▶ 6	SAME TIME, NEXT YEAR	4,00	0,00	4,25
▶ 7	THE WIZ	4,00	0,00	4,25
▶ 8	MIDNIGHT EXPRESS	6,00	2,00	4,17
▶ 9	THE DEER HUNTER	9,00	5,00	4,11
▶ 10	THE BOYS FROM BRAZIL	3,00	0,00	3,33
▶ 11	AN UNMARRIED WOMAN	4,00	1,00	3,25
▶ 12	DAYS OF HEAVEN	4,00	1,00	3,25
▶ 13	SUPERMAN	4,00	1,00	3,25
▶ 14	L'ARGENT DES AUTRES	5,00	2,00	3,20
▶ 15	MOLIÈRE	5,00	2,00	3,20
▶ 16	VIOLETTE NOZIÈRE	5,00	2,00	3,20
▶ 17	PERCEVAL LE GALLOIS	2,00	0,00	2,50
▶ 18	JUDITH THERPAUVE	2,00	0,00	2,50
▶ 19	AUTUMN SONATA	2,00	0,00	2,50

Figura 22 - Informe 8: Películas fracasadas

3.7.8 Informe 8: País con más premios ganados. Número de premios por país y categoría.

Este informe utilizará la carpeta personalizada *Pelicula_Pais* del Área de Negocio.

Dividiremos el informe en dos pestañas diferentes:

- Premios por País:

Informe en formato tabla página-detalle que muestra el número de premios ganados por País y los ordenados de forma descendente.

Para filtrar los datos utilizado la clase de elementos "es Ganador" y un operador "distinto de" para eliminar los premios sin país asignado.

- Premios por País y Categoría:

Informe en formato Matriz, que muestra por cada País, el número de premios ganados por categoría.

Para filtrar los datos utilizado la clase de elementos "es Ganador" y un operador "distinto de" para eliminar los premios sin país asignado.

	Número de Premios
USA	2352
▶ MEJOR ACTOR	129
▶ MEJOR ACTOR SECUNDARIO	74
▶ MEJOR ACTRIZ	125
▶ MEJOR ACTRIZ SECUNDARIO	73
▶ MEJOR CORTOMETRAJE	6
▶ MEJOR CORTOMETRAJE DOCUMENTAL	66
▶ MEJOR DIRECTOR	93
▶ MEJOR IMAGEN	5
▶ MEJOR LOGRO ARTISTICO	10
▶ MEJOR MÚSICA	76
▶ MEJOR PELÍCULA	110
▶ MEJOR SONIDO	47
▶ MEJOR VESTUARIO	45
▶ [SIN CATEGORIA UNIFICADA]	1493
FRANCE	772
▶ MEJOR ACTOR	47
▶ MEJOR ACTOR SECUNDARIO	31
▶ MEJOR ACTRIZ	54
▶ MEJOR ACTRIZ SECUNDARIO	33
▶ MEJOR CORTOMETRAJE	10
▶ MEJOR CORTOMETRAJE DOCUMENTAL	1
▶ MEJOR DIRECTOR	36
▶ MEJOR IMAGEN	32
▶ MEJOR LOGRO ARTISTICO	1
▶ MEJOR MÚSICA	33
▶ MEJOR PELÍCULA	35

Figura 23 - Informe 8: Premios por país y categoría

	▶ País	Número de premios
▶ 1	USA	2352
▶ 2	FRANCE	772
▶ 3	ITALY	157
▶ 4	UNITED KINGDOM	146
▶ 5	FRANKREICH	83
▶ 6	SPAIN	48
▶ 7	USSR	44
▶ 8	GERMANY	43
▶ 9	SWEDEN	41
▶ 10	DEUTSCHLAND	39
▶ 11	JAPAN	37
▶ 12	POLAND	36
▶ 13	HUNGARY	30
▶ 14	CANADA	29
▶ 15	DENMARK	26
▶ 16	CHINA	25
▶ 17	SWITZERLAND	20
▶ 18	INDIA	18
▶ 19	MEXICO	18
▶ 20	AUSTRALIA	15
▶ 21	BELGIUM	15
▶ 22	BRAZIL	14
▶ 23	TAIWAN	14
▶ 24	CZECHOSLOVAKIA	13
▶ 25	YUGOSLAVIA	13
▶ 26	AUSTRIA	12
▶ 27	ROMANIA	12
▶ 28	HONG KONG	11

Figura 24 - Informe 8: Premios por país

3.7.9 Informe 9: Relación entre premios en distintos festivales.

Este informe utilizará la carpeta personalizada *PremiosEntreCategorias* del Área de Negocio.

Este informe muestra la relación porcentual existente para que una película, concursando en la misma categoría unificada, gane el premio en dos festivales diferentes.

Para ello, obtenemos un conjunto de datos formado por las películas ganadoras en categorías unificadas, y lo contrastamos con el mismo conjunto de datos pero en festivales diferentes. Así mismo. El resultado, será el número de películas ganadoras en la misma categoría unificada en festivales diferentes. Así mismo, obtendremos el subconjunto de datos formado por el número total de películas ganadoras por cada categoría unificada. La relación porcentual de relación entre premios en distinto festivales, vendrá dado por el cociente entre el primer resultado obtenido entre el número de premios en distintos festivales.

Nombre	FestivalOrigen	FestivalDestino	Cuantos	PremiosCategoria	Relacion (%)
MEJOR ACTOR	ACADEMY AWARD	BERLINALE	2	214	0,93
MEJOR ACTOR	ACADEMY AWARD	CANNES FESTIVAL	3	214	1,40
MEJOR ACTOR	ACADEMY AWARD	CÉSAR	1	214	0,47
MEJOR ACTOR	BERLINALE	ACADEMY AWARD	2	214	0,93
MEJOR ACTOR	CANNES FESTIVAL	ACADEMY AWARD	3	214	1,40
MEJOR ACTOR	CANNES FESTIVAL	CÉSAR	1	214	0,47
MEJOR ACTOR	CÉSAR	ACADEMY AWARD	1	214	0,47
MEJOR ACTOR	CÉSAR	CANNES FESTIVAL	1	214	0,47
MEJOR ACTRIZ	ACADEMY AWARD	BERLINALE	2	225	0,89
MEJOR ACTRIZ	ACADEMY AWARD	CANNES FESTIVAL	4	225	1,78
MEJOR ACTRIZ	BERLINALE	ACADEMY AWARD	2	225	0,89
MEJOR ACTRIZ	BERLINALE	CÉSAR	1	225	0,44
MEJOR ACTRIZ	CANNES FESTIVAL	ACADEMY AWARD	4	225	1,78
MEJOR ACTRIZ	CÉSAR	BERLINALE	1	225	0,44
MEJOR DIRECTOR	ACADEMY AWARD	BERLINALE	2	160	1,25
MEJOR DIRECTOR	ACADEMY AWARD	CÉSAR	1	160	0,63
MEJOR DIRECTOR	BERLINALE	ACADEMY AWARD	2	160	1,25
MEJOR DIRECTOR	BERLINALE	CÉSAR	1	160	0,63
MEJOR DIRECTOR	CÉSAR	ACADEMY AWARD	1	160	0,63
MEJOR DIRECTOR	CÉSAR	BERLINALE	1	160	0,63
MEJOR MÚSICA	BERLINALE	CÉSAR	1	111	0,90
MEJOR MÚSICA	CÉSAR	BERLINALE	1	111	0,90
MEJOR VESTUARIO	ACADEMY AWARD	CÉSAR	1	64	1,56
MEJOR VESTUARIO	CÉSAR	ACADEMY AWARD	1	64	1,56

Figura 25 - Informe 9

De esta manera y viendo la relación porcentual obtenida, podemos concluir que no existe relación entre ganar ser premiado en la misma categoría en festivales diferentes.

4 Conclusiones

Analizando los resultados obtenidos durante las diferentes fases del proyecto, y contrastándolos con los objetivos propios e iniciales del mismo, puedo extraer las siguientes conclusiones:

- Se ha satisfecho los objetivos del proyecto: se ha construido un almacén de datos partiendo de la información disponible y se ha facilitado su explotación mediante la construcción de informes diseñados para tal efecto.
- Se ha construido un proceso automatizado que realice las tareas de extracción, transformación y carga de forma transparente para el usuario final.

- El desarrollo de los informes finales de explotación del almacén de datos, satisfacen la solicitud de información requerida. Así mismo, se proporciona la una definición de "película exitosa" y "película fracasada" como se solicita en los requisitos iniciales del proyecto.
- Se han satisfecho en tiempo y forma los compromisos iniciales adquiridos en la planificación, definidos como hitos de proyecto.
- Se ha trabajado siguiendo la metodología definida en el inicio de proyecto, respetando las fases del proyecto.

Así mismo, también afirmo la consecución de los propósitos definidos como objetivos del TFC, consolidando mi experiencia en la construcción y programación bajo Sistemas Gestores de Base de Datos y entornos Data Warehouse.

Por lo tanto, puedo afirmar que se han cumplido los requisitos del proyecto.

5 Líneas de evolución futura

Como se define en la metodología, el proceso de construcción de un almacén de datos es un proceso iterativo de mejora continua, mejoras que permitan tanto una evolución en las funcionalidades, añadiendo nuevas fuentes de datos, o bien, afinando y depurando los datos ya existentes.

Por lo tanto, podemos definir múltiples líneas de mejora del producto, como por ejemplo:

- Evolución y enriquecimiento de los datos
Nuevos procesos que mejoren la calidad de las fuentes operacionales, depurando en mayor medida los datos que los usuarios cargan en el sistema.
Estos procesos se pueden realizar mediante la construcción de herramientas ad-hoc, que mejoren la inserción de datos, eliminando posibles incoherencias o fallos en la carga.
- Nuevas fuentes operacionales
Incorporación de nuevos datos de origen, ampliando las dimensiones del modelo o los atributos de las existentes, enriqueciendo los informes.
- Optimización y mejora de los procesos de extracción, transformación y carga
Existen múltiples formas de actuación para la mejora de este punto. Sin duda, mi propuesta iría enfocada hacia la utilización de lenguajes de programación de alto nivel, como por ejemplo .NET, que permitan construir una herramienta robusta, escalable y tolerante a fallos, que realice los procesos de manera desatendida y transparente al usuario.
- Nuevos procesos de minería de datos
Mejora del modelo añadiendo procesos de minería de datos, los cuales permitan extraer información predictiva o de tipo estadístico.

- Optimización y mejoras del rendimiento. Escalabilidad horizontal y vertical.
Aunque con el conjunto de datos existente actualmente no es necesario, se puede plantear en un futuro la necesidad de optimizar, mejorar los tiempos de respuesta de las consultas.

6 Glosario

Almacén de datos: Colección de datos orientada a un determinado ámbito, integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza.

Atributo: Un atributo representa un tipo básico de información en una dimensión.

Business Intelligence: Conjunto de estrategias y herramientas enfocadas a la administración y creación de conocimiento mediante el análisis de datos existentes en una organización o empresa.

Cubo OLAP: Subconjuntos de datos de un almacén de datos, organizado y sumariado dentro de una estructura multidimensional. Se compone de hechos llamados medidas que se clasifican por dimensiones. El cubo de metadatos se suele crear a partir de un esquema en estrella o copo de nieve, esquema de las tablas en una base de datos relacional. Las medidas se obtienen de los registros de una tabla de hechos y las dimensiones se derivan de la dimensión de los cuadros.

Data Warehouse: Almacén de datos.

Dataminig: Minería de Datos.

DDL: Lenguaje de definición de datos (Data Definition Language) es un lenguaje proporcionado por el sistema de gestión de base de datos que permite a los usuarios de la misma llevar a cabo las tareas de definición de las estructuras que almacenarán los datos así como de los procedimientos o funciones que permitan consultarlos.

DML: Lenguaje de Manipulación de Datos (Data Manipulation Language) es un lenguaje proporcionado por el sistema de gestión de base de datos que permite a los usuarios de la misma llevar a cabo las tareas de consulta o manipulación de los datos, organizados por el modelo de datos adecuado.

Dimensión: Una estructura que representa una de las caras de un cubo. Cada dimensión representa una categoría diferente, como región, tiempo, inmueble, etc., son las perspectivas de análisis de las variables (forman parte de las tablas de dimensiones).

Esquema en copo de nieve: Se da cuando alguna de las dimensiones se implementa con más de una tabla de datos. La finalidad es normalizar las tablas y así reducir el espacio de almacenamiento al eliminar la redundancia de datos; pero tiene la contrapartida de generar peores rendimientos al tener que crear más tablas de dimensiones y más relaciones entre las tablas lo que tiene un impacto directo sobre el rendimiento.

Esquema en estrella: Modelo de datos que tiene una tabla rodeada de las tablas de dimensiones. Este aspecto, de tabla de hechos (o central) más grande rodeada de radios o tablas más pequeñas, es lo que asemeja a una estrella.

ETL: Procesos de extracción, transformación y carga (Extraction, Transformation and Loading).

Jerarquía: Cada una de las dimensiones de un cubo OLAP puede resumirse mediante una jerarquía que define una estructura de navegación en función de un mayor o menor detalle sobre la dimensión estudiada.

Medida: Atributo numérico de una celda. Las medidas se obtienen de los registros de una tabla de hechos.

Minería de Datos: Análisis de los datos para descubrir relaciones, patrones, o asociaciones desconocidas.

Modelo dimensional: Un tipo de modelo de datos empleado en el Almacén de Datos. En un modelo dimensional existen dos tipos de tablas de dimensiones y tablas de hechos.

OLAP: procesamiento analítico en línea (On-Line Analytical Processing). Se trata de una solución utilizada en el campo de la llamada Inteligencia empresarial (o *Business Intelligence*) cuyo objetivo es agilizar la consulta de grandes cantidades de datos.

Tabla de dimensiones: Elementos que contienen atributos (o campos) que se utilizan para restringir y agrupar los datos almacenados en una tabla de hechos cuando se realizan consultas sobre dicho datos en un entorno de almacén de datos.

Tabla de hechos: Tabla central de un esquema dimensional que contiene los valores de las medidas de negocio.

7 Bibliografía

7.1 Web

- <http://es.wikipedia.org>
- <http://www.gestiopolis.com/canales8/ger/olap-online-analytic-processing.htm>
- <http://www.dataprix.com/blogs/respinosamilla/cubos-olap-line-analytic-processing>
- <http://www.jaspersoft.com/jaspersoft-olap>
- <http://www.wikioracle.es>
- <http://www.orafaq.com>
- <http://www.oracleutilities.com/OSUtil/sqlldr.html>
- <http://cisnet.baruch.cuny.edu/holowczak/oracle/sqlplus/>
- <http://www.oracle.com/technetwork/developer-tools/discoverer/overview/index.html>

8 Anexos

8.1 Entorno de Producción

8.1.1 Requisitos de servidor

El servidor será el encargado de almacenar y gestionar el repositorio central de datos, así como de servir los datos solicitados por los informes alojados en el *lado del cliente*. Para ello, se instalará el SGBD Oracle Database 10g Express Edition. Para un óptimo funcionamiento de este software, se recomienda una plataforma de producción con, al menos, los siguientes requisitos hardware/software:

- Sistema Operativo Windows 32-bit
 - o Windows 2000 Server SP4 o superior
 - o Windows 2003 Server
 - o Windows XP Professional SP1 o superior
- 1,6 Gb. de espacio en disco
- 512 Mb. de RAM
- Microsoft Windows Installer v2.0 o superior

8.1.2 Requisitos en cliente

El lado *cliente* será el encargado de solicitar al servidor los datos utilizando la aplicación Oracle Discoverer Desktop, incluida dentro del entorno de desarrollo Oracle Developer Suite

8.2 Entorno de Desarrollo

8.2.1 Recursos de software

El entorno de desarrollo dispondrá de, al menos, las siguientes herramientas de trabajo:

- VirtualBox
- Oracle Discoverer
- Oracle SQL*Loader
- PL/SQL Developer
- Microsoft Excel
- Microsoft Visio

8.2.2 Recursos Hardware

El entorno de desarrollo Host:

- iMac 2.93 GHz Core i7, 8Gb DDR3 RAM con sistema operativo Mac OS X 10.6
- Backup: sistemas de copia NAS bajo software TmeMachine.
- Servicio de almacenamiento en la nube con DropBox y Apple mobileMe .

Entorno de desarrollo virtualizado:

- Windows XP, 3GB RAM bajo sistema VirtualBOX VM

8.3 Localización de Entregables del Proyecto

Listado y localización de los entregables resultantes del proceso de desarrollo y construcción del proyecto, alojados en la unidad de almacenamiento virtual en (E:) formato *VirtualBox*, que contiene el entorno en producción del proyecto: *fuencia_vm.zip*.

Proceso principal ETL – Fichero por lotes

Automatiza las tareas ETL y optimización de índices de la base de datos tras la carga.

Este proceso, toma como datos CSV de origen los almacenados en E:\TFC\DATOS_CSV

- E:\TFC\ETL.bat.

Construcción de la Base de Datos

Scripts de programación DDL de construcción de los objetos de la base de datos:

- E:\TFC\0_DDL\DDL.sql
- E:\TFC\0_DDL\DDL_Sequence.sql (creación de objetos *sequence*)
- E:\TFC\0_DDL\MisFunciones.pck (definición de cabecera y cuerpo del paquete *misFunciones*)

Proceso ETL - Extracción de datos

Ficheros de control (CTL), fuente (CSV) y de registros tras la extracción (LOG) almacenados en la ruta E:\TFC\1_Extraccion:

- loader_Academy.ctf, Academy_Awards_2006.csv, sqlloader_Academy.log
- loader_Berlinale.ctf, Berlinale_2004.csv, sqlloader_Berlinale.log
- loader_Cannes.ctf, Cannes_Awards_2005.csv, sqlloader_Cannes.log
- loader_Cesar.ctf, Cesar_Awards_2005.csv, sqlloader_Cesar.log
- loader_MTV.ctf, MTV_Movie_Awards_2006.csv, sqlloader_MTV.log

Proceso ETL – Transformación de datos

Scripts de los procesos de transformación de los datos:

- E:\TFC\2_Transformacion\transformaciones.sql

Proceso ETL – Carga de datos

Scripts de los procesos de carga de los datos transformados:

- E:\TFC\3_Carga\Carga.sql
- E:\TFC\3_Carga\Carga_Categoria_CategoriaTipos.sql
- E:\TFC\3_Carga\Carga_CategoriaTipos.sql
- E:\TFC\3_Carga\Carga_ETLPaises.sql

Otros procesos – Actualización de estadísticas del esquema

Scripts y lanzador manual del proceso de actualización de estadísticas de los objetos de Oracle (índices y tablas) del esquema TFC:

- E:\TFC\4_Otros\stats\tfc.stats.sql
- E:\TFC\4_Otros\stats\tfc.stats.bat. Lanzador manual de actualización.

Otros procesos – Exportación completa del esquema del usuario

Proceso manual de backup de los datos del esquema TFC de Oracle para recuperación manual ante posibles desastres.

- E:\TFC\4_Otros\exp\tfc.exp.bat. Lanzador de proceso de exportación.
- E:\TFC\4_Otros\exp\tfc.exp.dmp. Fichero de backup
- E:\TFC\4_Otros\exp\tfc.exp.log. Fichero de registro.

Otros ficheros – Exportación de la Capa de Usuario Final (EUL)

Backup generado desde Oracle® BI Discoverer Administrator, para recuperación ante posibles desastres.

- E:\TFC\5_EUL\tfc.eul.eex
- E:\TFC\5_EUL\tfc.eul.txt
- E:\TFC\5_EUL\tfc.eul.log.txt

Otros ficheros – Sentencias SQL utilizadas en las carpetas personalizadas del Área de Negocio

Programación DML de las sentencias de selección utilizadas en el Área de Negocio TFC de Oracle® BI Discoverer Administrator:

- E:\TFC\5_EUL\CarpetasPersonalizadas\Pelicula_Pais.sql
- E:\TFC\5_EUL\CarpetasPersonalizadas\PremiosEntreCategorias.sql

Informes de Oracle® BI Discoverer Desktop

Informes propietarios de Oracle® BI Discoverer Desktop con extensión DIS:

- Informe 1: Premios por actor o actriz, año y festival
 - o E:\TFC\6_Informes\1_PremiosActores_AñoFestival (listado).DIS
 - o E:\TFC\6_Informes\1_PremiosActores_AñoFestival.DIS
- Informe 2: Premios por actor o actriz, año, festival y categoría
 - o E:\TFC\6_Informes\2_PremiosActores_AñoFestivalCategoría (listado).DIS
 - o E:\TFC\6_Informes\2_PremiosActores_AñoFestivalCategoría.DIS
- Informe 3: Premios por director, año, festival y película.
 - o E:\TFC\6_Informes\3_PremiosDirectores_AñoFestivalPelicula (listado).DIS
 - o E:\TFC\6_Informes\3_PremiosDirectores_AñoFestivalPelicula.DIS
- Informe 4: Premios por año, festival, categoría y película.
 - o E:\TFC\6_Informes\4_Premios_AñoFestivalCategoríaPelicula (listado).DIS
- Informe 5: Número de ediciones del festival, total de actores nominados, total de actrices nominadas y total de películas nominadas por tipo de festival, dada una fecha a seleccionar.
 - o E:\TFC\6_Informes\5_Ediciones.DIS
- Informe 6: Premios por año, película y festival.
 - o E:\TFC\6_Informes\6_Premios_AñoPeliculaFestival (listado).DIS
- Informe 7: Definición de "película exitosa" y "película fracasada", por año.
 - o E:\TFC\6_Informes\7_PeliculaExistosaFracasada.DIS
- Informe 8: País con más premios ganados. Número de premios por país y categoría.

- E:\TFC\6_Informes\8_PremiosPaises.DIS
- Informe 9: Relación entre premios en distintos festivales.
 - E:\TFC\6_Informes\9_PremiosEntreCategoria.DIS

8.4 Scripts de procesos ETL

8.4.1 Proceso de Extracción

```
sqlldr <usuario>/<clave>@XE -control=loader_Academy.ctl -log=sqlloader_Academy.log errors=0
sqlldr <usuario>/<clave>@XE -control=loader_Berlinale.ctl -log=sqlloader_Berlinale.log errors=0
sqlldr <usuario>/<clave>@XE -control=loader_Cannes.ctl -log=sqlloader_Cannes.log errors=0
sqlldr <usuario>/<clave>@XE -control=loader_Cesar.ctl -log=sqlloader_Cesar.log errors=0
sqlldr <usuario>/<clave>@XE -control=loader_MTV.ctl -log=sqlloader_MTV.log errors=0
```

Ejemplo de uno de los ficheros de control:

```
OPTIONS (SKIP=1)
load data
infile '<fichero_datos_csv>'
append
into table <usuario>.temp
fields terminated by ";" optionally enclosed by ""
(original, english, sort, year, country, awards, category, winner, nominee, academy, item, idrow
sequence(MAX,1))
```

8.4.2 Script de Transformación

Desarrollo de la programación PL/SQL del script de transformación:

```
/* CAMPO ORIGINAL */
update temp set original = trim(original);
update temp set original = upper(original);
update temp set original = upper(english) where original is null and not english is null;
update temp set original = '[NO SPECIFIC FILM TITLE]' where original IS NULL;
update temp set original = replace(original,' : ',' ');
update temp set original = replace(original,'&', 'AND');
update temp set original = replace(original,' AND', ' AND');
/* FIN CAMPO ORIGINAL */

/* CAMPO YEAR */
update temp set year = substr(year,instr(year,'-')+1, LENGTH(year));
/* FIN CAMPO YEAR */

/* CAMPO COUNTRY */
update temp set country = upper(country);
update temp set country = '[NO COUNTRY]' where country IS NULL;
update temp set country = replace(country,',','/');
```

```
/* FIN CAMPO COUNTRY */

/* CAMPO AWARDS */
update temp set awards = upper(awards);
/* FIN CAMPO AWARDS */

/* CAMPO CATEGORY */
update temp set category = upper(category);
update temp set category = trim(category);
/* FIN CAMPO CATEGORY */

/* CAMPO WINNER */
update temp set winner = DECODE(winner,'X','S','0','N','N');
/* FIN CAMPO WINNER */

/* CAMPO NOMINEE */
update temp set nominee = trim(nominee);
update temp set nominee = upper(nominee);
update temp set nominee = replace(nominee,' ','');
update temp set nominee = '[NO NOMINEE]' WHERE nominee is null;
/* FIN CAMPO NOMINEE */

/* CAMPO ACADEMY */
update temp set academy = trim(academy);
update temp set academy = upper(academy);
update temp set academy = replace(academy,'»RE NUIT DES C...SARS','');
update temp set academy = replace(academy,'»ME NUIT DES C...SARS','');
update temp set academy = replace(academy,'ST','');
update temp set academy = replace(academy,'ND','');
update temp set academy = replace(academy,'RD','');
update temp set academy = replace(academy,'TH','');
/* FIN CAMPO ACADEMY */
```

8.4.3 Script de Carga

Desarrollo de la programación PL/SQL del script de carga:

```
/* TABLA PELICULA */
insert into pelicula (select SEQ_IDPELICULA.Nextval, t.original from
    (select distinct original from temp where not exists (SELECT 1 from pelicula where pelicula.original =
temp.original) order by temp.original) t);
/* FIN TABLA PELICULA */

/* TABLA PAIS */
call tfc.misfunciones.cargaPaises();
/* FIN TABLA PAIS */

/* TABLA PELICULAPAIS */
```

```
call tfc.misfunciones.cargaPeliculaPais();
/* FIN TABLA PELICULAPAIS */

/* TABLA NOMINADO */
insert into nominado (select SEQ_IDNOMINADO.Nextval, t.nominee from
    (select distinct temp.nominee from temp where not exists (SELECT 1 from nominado where
nominado.nombre = temp.nominee) order by temp.nominee) t);
/* FIN TABLA NOMINADO */

/* TABLA FESTIVAL */
insert into festival (select SEQ_IDFESTIVAL.Nextval, t.awards from
    (select distinct temp.awards from temp where not exists (SELECT 1 from festival where festival.nombre =
temp.awards) order by temp.awards) t);
/* FIN TABLA FESTIVAL */

/* TABLA EDICION */
insert into Edicion (idEdicion, Ano, Idfestival, Edicion)
    (SELECT SEQ_IDEDICION.Nextval, t.year, t.idfestival, t.Academy
    FROM (select distinct festival.idfestival, year, academy
        from temp inner join festival on temp.awards = festival.nombre order by year, festival.idfestival,
academy) t
    WHERE NOT EXISTS (SELECT 1 FROM EDICION WHERE Edicion.Ano = t.year AND Edicion.Idfestival =
t.idfestival AND Edicion.Edicion = t.academy));
/* FIN TABLA EDICION */

/* TABLA CATEGORIA */
insert into Categoria (idcategoria, Nombre) (select SEQ_IDCATEGORIA.Nextval, t.category from
    (select distinct temp.category from temp where not exists (SELECT 1 from categoria where
categoria.nombre = temp.category) order by temp.category) t);
/* FIN TABLA CATEGORIA */

/* TABLA DE HECHOS */
call tfc.misfunciones.cargaPremios();
/* FIN TABLA DE HECHOS */
```

8.4.4 Script de actualización de estadísticas

```
execute dbms_stats.gather_schema_stats(ownname=> '<PROPIETARIO>' , cascade=> TRUE);
```

8.5 Carpetas personalizadas del Área de Negocio

8.5.1 Script Pelicula_Pais

```
select pelicula.original as NombrePelicula, pais.nombre as NombrePais, pelicula.idpelicula, pais.idpais
from pelicula, peliculapais, pais
where pelicula.idpelicula = peliculapais.idpelicula
and peliculapais.idpais = pais.idpais
```

8.5.2 Script PremiosEntreCategorias

```
select categoriatipos.idtipo, categoriatipos.nombre, fo.idfestival idFestivalOrigen, fo.nombre festivalorigen,
fd.idfestival idFestivalDestino, fd.nombre festivaldestino, t.cuantos, contador.cuantos p_categoria,
(t.cuantos/contador.cuantos) porcentaje
from (select origen.idfestival as festivalorigen, destino.idfestival as festivaldestino, origen.idtipo, count(*) as
cuantos
from (select premios.idpelicula, premios.idedicion, premios.idcategoria, premios.idnominado,
edicion.idfestival, categoria.idtipo from premios
inner join edicion on premios.idedicion = edicion.idedicion
inner join categoria on premios.idcategoria = categoria.idcategoria
where premios.esganador='S' and categoria.idtipo <> 0) origen
inner join (select premios.idpelicula, premios.idedicion, premios.idcategoria, premios.idnominado,
edicion.idfestival, categoria.idtipo from premios
inner join edicion on premios.idedicion = edicion.idedicion
inner join categoria on premios.idcategoria = categoria.idcategoria
where premios.esganador='S' and categoria.idtipo <> 0) destino
on origen.idpelicula = destino.idpelicula and
origen.idnominado = destino.idnominado and
origen.idtipo = destino.idtipo
and origen.idfestival <> destino.idfestival
group by origen.idfestival, destino.idfestival, origen.idtipo) t
inner join (select categoriatipos.idtipo, count(*) cuantos from premios
inner join categoria on premios.idcategoria = categoria.idcategoria
inner join categoriatipos on categoria.idtipo = categoriatipos.idtipo
where premios.esganador = 'S' group by categoriatipos.idtipo) contador on t.idtipo = contador.idtipo
inner join categoriatipos on t.idtipo = categoriatipos.idtipo
inner join festival fo on fo.idfestival = t.festivalorigen
inner join festival fd on fd.idfestival = t.festivaldestino
order by categoriatipos.nombre, fo.nombre, fd.nombre
```