



SINUE

Construcción de un Sistema NUEVO de Estadísticas y de Análisis de Información para la Tesorería General de la Seguridad Social basado en la tecnología de Data Warehousing

SEVILLA ESCRIBANO María Ángeles

Desde 1994 hasta 2000 ha desempeñado las funciones de Análisis, diseño y mantenimiento, así como Coordinación del grupo de trabajo del Sistema de Generación de Estadísticas sobre Información Laboral SILEST de la TGSS. En la actualidad desempeña las funciones de Coordinación del grupo de trabajo, así como del análisis y diseño del proyecto SINUE.

LOSCOS MUÑOZ, Carmen

Con funciones de Análisis y Diseño de nuevas Aplicaciones, en el entorno de AFILIACIÓN, así como mantenimiento de los procesos referentes a la gestión de Empresas. Actualmente incorporada al grupo de trabajo del proyecto SINUE.

IGLESIAS MENDEZ, José Luis

En el ámbito de Prestaciones desde el año 1983, dando servicio a las Entidades Gestoras INSS e ISM, desempeñando en este periodo de tiempo todo tipo de funciones, desde programación hasta dirección de proyectos. Actualmente en el grupo de trabajo del proyecto SINUE.



Ayuntamiento de A Coruña





INTRODUCCIÓN

En la actualidad, el objetivo de las organizaciones es convertir los “Datos” que posee en “Información”, para obtener CONOCIMIENTO en torno al negocio.

La Dirección, en nuestro caso, el órgano directivo de la Tesorería General de la Seguridad Social, solicita Conocimiento de la gestión de su organización para tomar decisiones de gran alcance sobre la Gestión, como establecer presupuestos, pautas de trabajo, etc., en los diferentes ámbitos de la Seguridad Social (Recaudación, Afiliación o Prestaciones). Esto implica que la información que se va a poner en conocimiento de los gestores debe ser ante todo fiable, es decir, debe reflejar la realidad de la gestión, así como debe hacerse en un breve periodo de tiempo, ya que como es bien sabido, la información es tanto más valiosa cuanto más rápida se obtenga, es decir, “cuando llega en el momento justo al lugar adecuado”.

En definitiva cualquier organización necesita tener un amplio Conocimiento para poder responder a preguntas del tipo:

- ¿Qué pasó?
- ¿Por qué paso?
- ¿Qué puede pasar?
- ¿Qué pasará?



La Subdirección General de Informática de la Tesorería General de la Seguridad Social (en adelante S.G.I. de la T.G.S.S.), viene facilitando información relevante de la “Gestión de la Seguridad Social”. En la mayoría de los casos, esto supone la realización de procedimientos que implican un gasto elevado de recursos tanto humanos, como de hardware, software (líneas de código) y tiempo de proceso. Para resolver consultas o peticiones de información, cuando estas se producen de forma puntual y urgente, se dispone de un Sistema de proceso masivamente paralelo donde los datos, réplica diaria de la información corporativa operacional que reside en plataforma OS/390, pueden ser analizados a tra-



vés de una aplicación desarrollada en la propia S.G.I., “Proyecto Q+”. Esta aplicación es un generador automático de consultas SQL (Structure Query Language). El hecho de que la información residente en este entorno se refresque diariamente, impide obtener series históricas para realizar estudios evolutivos.

De ahí la necesidad de disponer de un “Nuevo Sistema de Estadísticas y de Análisis de Información” que permita Extraer, Integrar y Modelar los datos del Negocio para ponerlos a disposición de los Responsables de la Toma de Decisiones en la Forma Óptima. Así los gestores podrán tener de forma ágil, sencilla, segura y con el menor coste posible, datos relevantes o indicativos, tanto históricos como actuales, de la “Organización” ‡ “La Seguridad Social”. La información se debe poder observar según distintas perspectivas, ya que los usuarios de este Sistema se pueden clasificar según diferentes niveles: Estratégicos, Analistas y Operativos. Esta clasificación responde al nivel de profundización, tipo de conocimiento que necesitan manejar según el perfil que posea dentro de la organización. En el caso de la “Seguridad Social”, por ejemplo en el ámbito de la gestión de la Recaudación Ejecutiva se pueden encontrar tres perfiles claramente diferenciados: usuarios de Servicios Centrales, usuarios de Direcciones Provinciales y Usuarios de Unidades de Recaudación Ejecutiva. Todos ellos pueden tener interés por el mismo tipo de información pero vista desde diferentes perspectivas y niveles de profundización.

Los datos que se toman como referencia para hacer un seguimiento del negocio dentro de la T.G.S.S., son los llamados “Cuadros de Mando”, así como otros informes estadísticos no encuadrados dentro de esta denominación, y que afectan a todas las Entidades Gestoras.

En la actualidad la tendencia, para responder a estas necesidades de información, es utilizar productos orientados a Soluciones Bussines Intelligent (Data Warehousing, Data Mart, Data Mining) de Análisis y Soporte a la Toma de Decisiones.

La base fundamental de estos sistemas es la creación de un Almacén de Datos ‡ “Data Warehouse”. Bill Inmon, considerado padre del Data Warehousing, lo definió como

“Colección de datos no volátiles, orientados al usuario, integrados, variables en el tiempo, que se utilizan principalmente para ayudar a la toma de decisiones de la organización”.



Una definición menos formal sería, “Repositorio común centralizado, con información precisa, totalmente consolidada, consistente y limpia para el soporte a la toma de decisiones, facilitando la aplicación de mejores estrategias de negocio”.

Los datos que contendrá dicho Almacén procederán de los Sistemas Operacionales residentes en las bases de datos de negocio, también llamadas “Corporativas”, a los cuáles se les aplicarán Transformaciones (a través de reglas, ya sean simples o complejas), Clasificaciones, Sumarizaciones, etc., antes de su carga en el (Data Warehouse o Data Marts)

Un Data Mart se define como un subconjunto de información del Data Warehouse específica y orientada a una línea de negocio de la organización. Los data mart se puede clasificar dependiendo de la existencia previa o no de un Data Warehouse en: Dependientes o Independientes. Para este proyecto se ha optado por el modelo de construir un Data Mart independiente, ya que se pretende un desarrollo incremental.

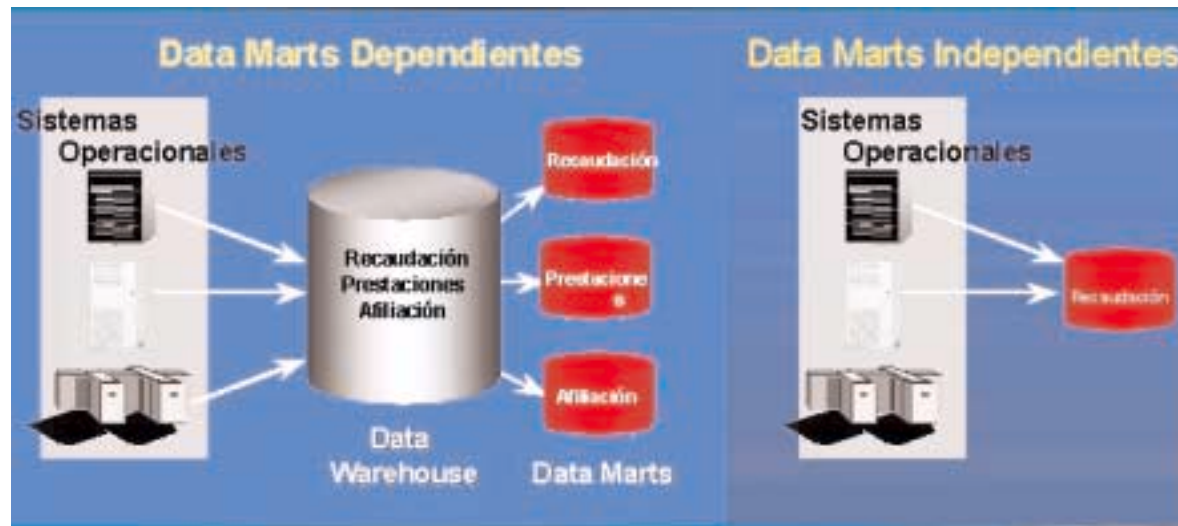


Figura 1.



Para que cualquier herramienta, ya sea de soporte a la toma de decisiones, o un aplicativo desarrollado a propósito para obtener una determinada estadística sea capaz de obtener información de estos almacenes, no es solo necesario sino fundamental la existencia de lo que se ha dado en llamar “Metadatos”, aquellos Datos que describen a los Datos.

Metadatos consiste en un Diccionario de “estructuras de datos” que se define para asistir en los procesos de consulta contra la base de datos del Almacén. Este diccionario contiene toda la información correspondiente a las características de los datos.

En definitiva lo importante es la búsqueda de los Datos Operacionales que van a ser de interés, los procesos de carga de estos datos en el Almacén, su Mantenimiento y las diferentes Actualizaciones.

Por tanto, y debido a las dificultades que se encuentran a la hora de obtener información a partir de los Sistemas Operacionales, nos proponemos desarrollar un sistema que permita obtener información de una forma ágil, sencilla, automática, y cercana al usuario (entorno amigable) que responda a las necesidades actuales de información así como a las futuras sin que afecte en profundidad, a la estructura del Sistema de Información.

Para abordar este proyecto se realizó una contratación mediante concurrencia de ofertas de una solución integrada para la construcción de un Sistema de Estadísticas y de Análisis de Información para la actual Subdirección de Informática de la Tesorería General de la Seguridad Social y que pudiera dar servicio a todas las Entidades Gestoras de la Seguridad Social, siendo la empresa adjudicataria Oracle Ibérica S.R.L.

Esta solución se estructuró en un conjunto de actividades y software a aplicar, de una forma iterativa, evolutiva y totalmente escalable. Se contrató además de la base tecnológica el conocimiento y la experiencia en el desarrollo e implantación de este tipo de soluciones. Para ello se comenzó con la elaboración e implementación de un primer incremento que corresponde a un Data Mart para Recaudación Ejecutiva. Con esta primera experiencia se probaron no sólo los productos sino también la metodología.



A continuación se muestra la planificación general de ejecución del proyecto:

Proyecto para la Definición de una Estrategia Global Data Warehouse

	Octubre					Noviembre				Diciembre				Enero				Febrero			Marzo						
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	
Trámite de Ejecución																											
Análisis																											
Reuniones para la toma de datos	█																										
Prototipado y pruebas de concepto	█																										
Documentación de requerimientos del TE	█																										
Final de la Toma de Requerimientos del TE	█																										
Diseño																											
Diseño y documentación	█					█																					
Estrategia (ORACLE) de un almacén del formato de los flujos de carga	█					█																					
Estrategia documento de Análisis y Diseño	█					█																					
Aprobación documento de Análisis y Diseño	█					█																					
Construcción																											
Construcción procesos de generación de los flujos de carga						█				█				█				█									
Estrategia (OGG) de los flujos de carga						█				█				█				█									
Construcción de módulos (ETT, BBDD, ACCESO, ...)						█				█				█				█									
Elaboración del Plan de Pruebas						█				█				█				█									
Aprobación del Plan de Pruebas						█				█				█				█									
Ejecución del Plan de Pruebas						█				█				█				█									
Sistema construido, probado y aceptado en Desarrollo						█				█				█				█									
Transición																											
Disponibilidad Servidor Producción (HW) y SW instalado y configurado																		█			█						
Migración del DW del entorno de desarrollo al entorno de producción																		█			█						
Sistema en Producción																		█			█						
Arranque																											
Actualizaciones del DW de TE en producción																		█			█						



CARACTERÍSTICAS TÉCNICAS



Arquitectura



Ayuntamiento de A Coruña



La arquitectura se basa en un entorno de tipo SMP (Symmetric Multiprocessing). Está soportado en una máquina PRIMEPOWER 1000 de FUJITSU SIEMENS, con Sistema Operativo Solaris 7. y la base de datos instalada es Oracle 8i.

Se han establecido dos entornos, uno de desarrollo y otro de producción. Ambos están soportados en la misma máqui-



na pero en particiones diferentes. El entorno actual de desarrollo dispone de 4 procesadores Sparc64 GP de 450 MHz, 8GB de memoria con 12GB de swap y páginas de 8 KB.

En cuanto al espacio en Disco, se ha dimensionado en 10 GB para instalación del software tanto en desarrollo como en producción, y 100GB para datos en desarrollo y 500GB para producción.

Diseño Conceptual

A continuación se muestra la topología del diseño conceptual de la solución.

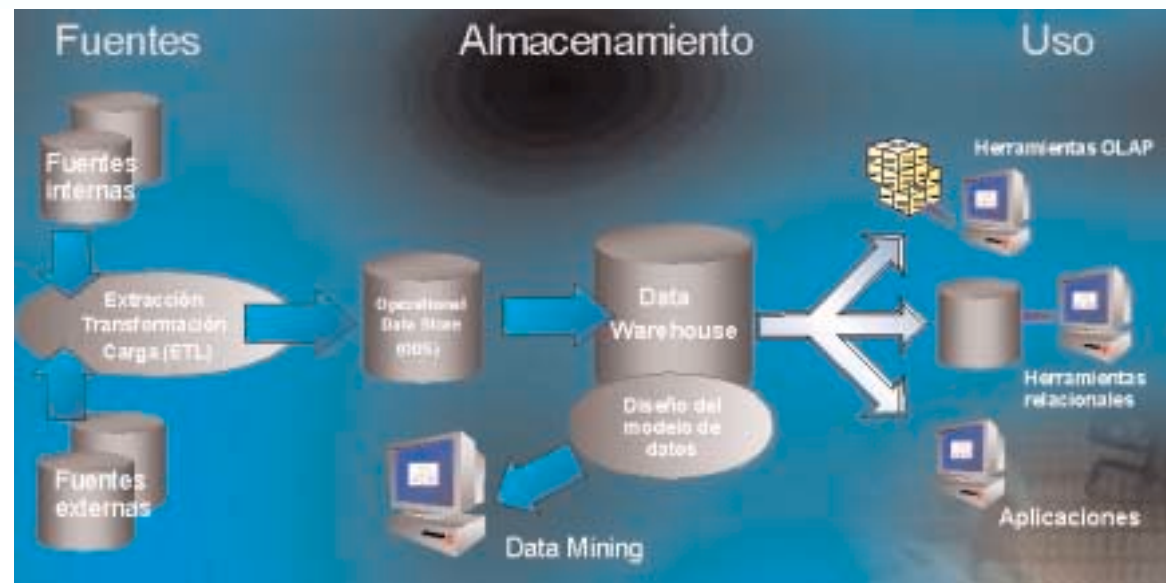
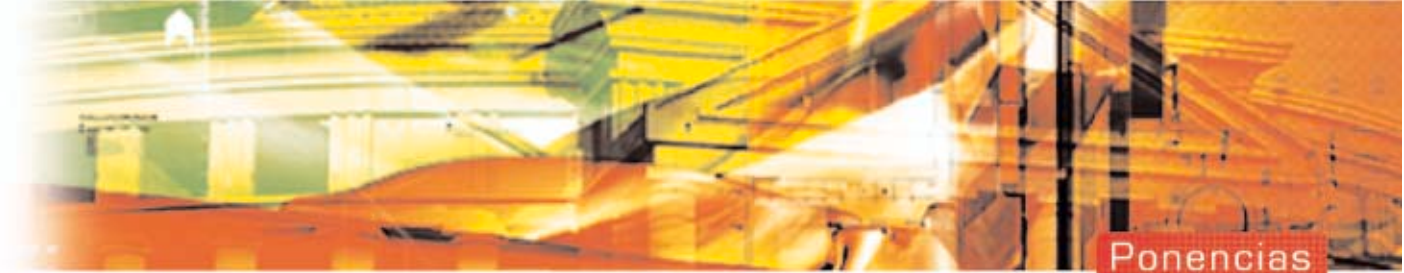


Figura 2.

Previo a cualquiera de las fases que se muestran en la figura 2, hay una primera fase estratégica que corresponde a la "Toma de requerimientos".



En esta fase mediante entrevistas con el usuario, se establecen de forma clara y precisa cuáles son las necesidades de información del usuario, así como los diferentes perfiles de conocimiento que hay que satisfacer. En definitiva se definen, clarifican y priorizan los propósitos y objetivos del Data Warehouse o del Data Mart , así como su alcance.

En la figura se puede observar que la construcción y explotación del Data Warehouse (DWH), comienza por una fase de "Identificación y Definición de la fuentes de información".

Esta fase está muy ligada a la Toma de Requerimientos, ya que en ella se establecerán tanto las fuentes externas como internas que alimentarán al Data Warehouse.

- Fuentes Externas, información que no controla la propia organización, como puede ser datos de población procedentes del Instituto Nacional de Estadística..
- Fuentes Internas, corresponden a los datos operacionales provenientes del sistema transaccional (ficheros Adabas, ficheros secuenciales planos, tablas Oracle...)

Dentro de esta fase está el diseño del modelo de datos. Se basa en un modelo multidimensional (en forma de estrella) que presenta una visión sencilla y orientada al negocio de datos organizados mediante diferentes perspectivas o "dimensiones" de análisis, organizadas jerárquicamente.

Este esquema se basa en grandes tablas llamadas "tablas de hechos" (la central en la figura anterior Gestión P.A.), que contienen los valores a consultar, por ejemplo Número de Títulos Ejecutivos Cobrados. Se completa el diseño con "dimensiones" por las cuales se





desea analizar la información, corresponde a las tablas periféricas (URE, Geográfica, Actividad....). Este diseño permite realizar consultas a millones de registros con un tiempo de respuesta óptimo.

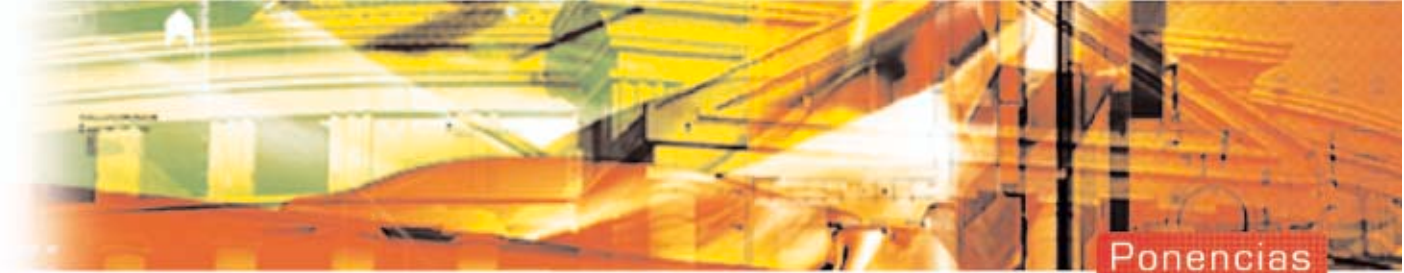
Por ejemplo, un modelo podría corresponder dentro del ámbito de la Recaudación en Vía Ejecutiva a la Gestión de los Títulos Ejecutivos según la dimensiones de Tiempo (fundamental y que siempre existirá en cualquier estrella), Régimen, Geografía, URE (Unidad de Recaudación Ejecutiva), Tipo de documento, etc., de tal manera que cada usuario puede centrar su visión de negocio en la perspectiva que sea de más interés, por ejemplo un usuario de una Dirección Provincial centrará su visión en el ámbito que corresponda a su dirección provincial, mientras que un usuario de Servicios Centrales, puede necesitar verla con una perspectiva global de comparativa de la gestión sin fijarse en una unidad geográfica en concreto. Debido a la jerarquización de la información a través de las dimensiones es posible dado un informe profundizar en algunos de sus ejes de análisis con gran sencillez y agilidad, así como realizar agregaciones. Es lo que se conoce como drill down/drill up.

La siguiente fase es la de “Extracción, Depuración, Transformación y Carga”, de los datos, desde las fuentes de datos operacionales al Data Warehouse.

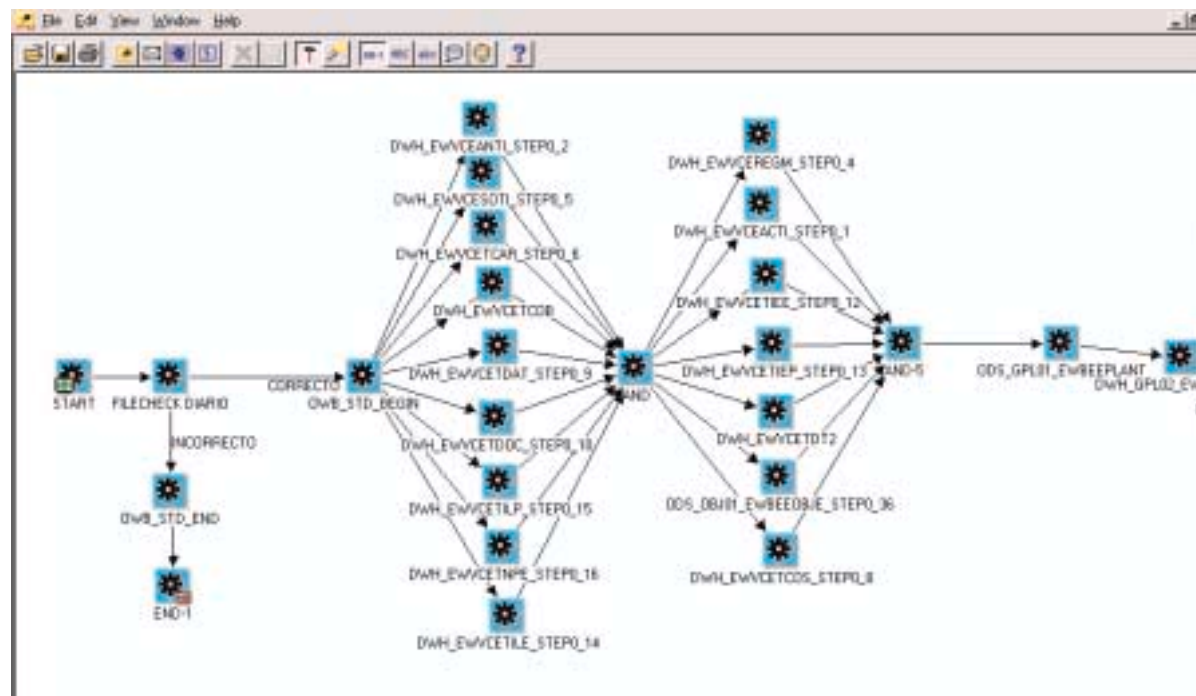
Antes de cargar la información con la estructura propia del Data Warehouse, se establece un Área de Almacenamiento Temporal , ODS (Operational Data Store o Staging Area), en la que se almacena la información unificada de las distintas fuentes de datos, al máximo nivel de detalle. Los procesos de sumarización y de consolidación de información se llevan a cabo dentro de este área a fin de preparar los datos para ser incorporados al Data Warehouse.

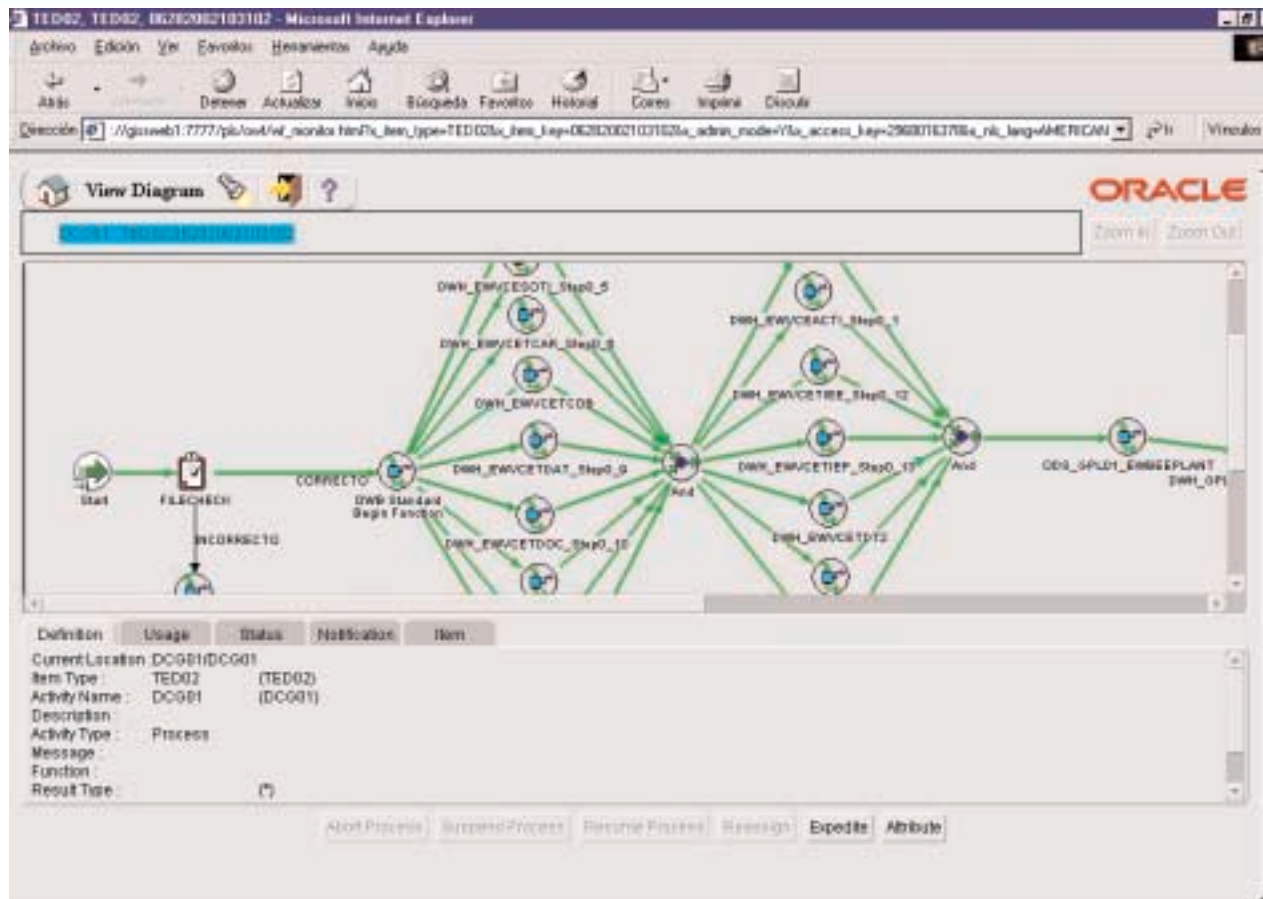
Estos procesos se realizan mediante las llamadas herramientas de ETL (Extraction, Transformation & Loading).

La herramienta utilizada tanto para la fase de Diseño como para la fase de ETL, es Oracle Warehouse Builder (OWB). Esta herramienta permite mediante una interfaz gráfica de usuario(GUI), el desarrollo y mantenimiento de las estructuras definidas tanto para las fuentes primarias, como para el ODS, así como para las definitivas del Data Warehouse. OWB, está implementada en entorno Cliente-Servidor.

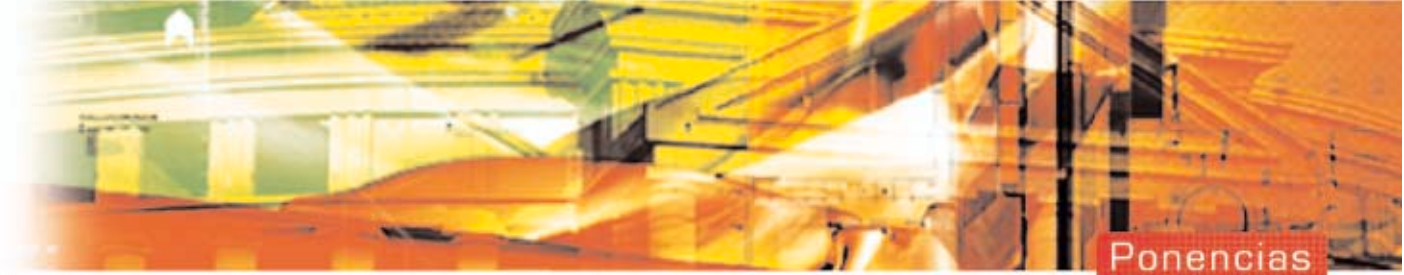


Complementaria a esta herramienta se utiliza Oracle Workflow Builder, que permite definir y mantener los flujos de las cargas de información. Los flujos se definen en Cliente-Servidor, pero el seguimiento de los mismos se realiza con Oracle Workflow Monitor mediante una interfaz gráfica en entorno Web.

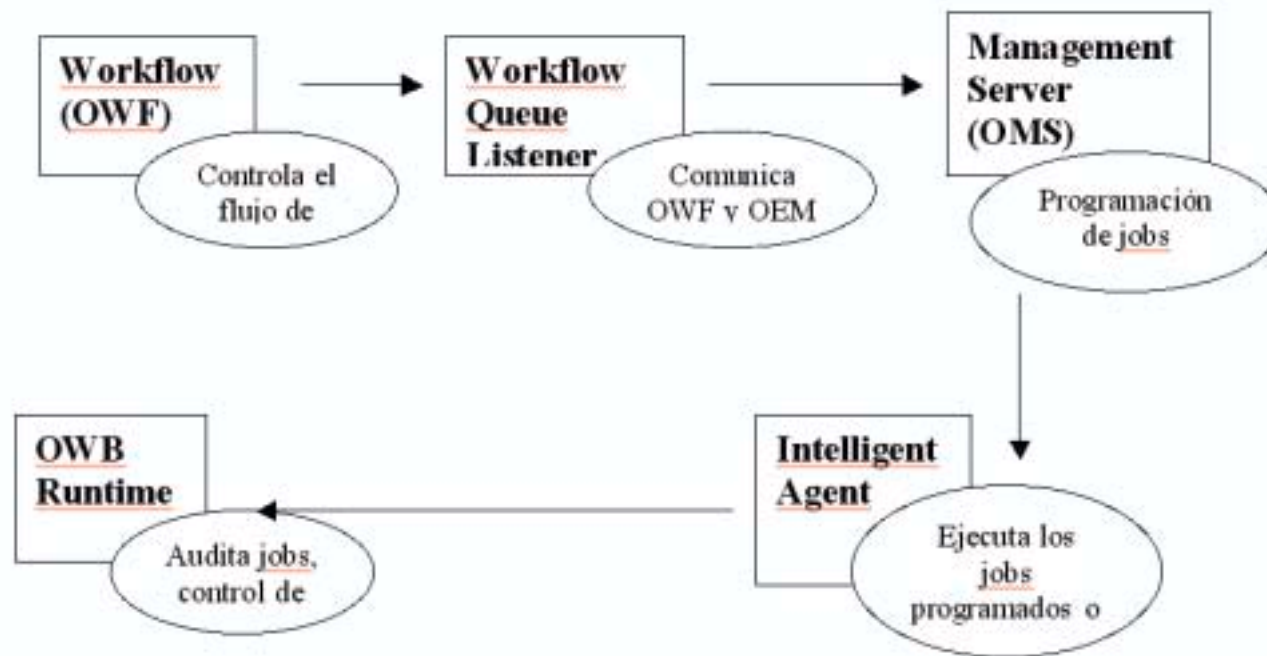




OWF Monitor



Además de las herramientas citadas anteriormente, existen otros componentes que son fundamentales para la ejecución de las cargas: Oracle Enterprise Manager, Oracle Intelligent Agent. A continuación se muestra el flujo de comunicaciones entre los diferentes componentes:



Workflow controla el flujo del proceso de carga, solicitando la ejecución de los jobs necesarios en cada momento, y deteniendo el proceso en caso de error. Para hacer esto, se comunica con el OMS (Oracle Management Server, parte de Oracle Enterprise Manager que gestiona la ejecución de jobs), a través del Workflow Queue Listener. El OMS remite el job al Intelligent Agent, que se encarga de ejecutarlo en el entorno del Oracle Warehouse Builder Runtime. Las operaciones de carga quedarán monitorizadas por este último en todo detalle. A su vez OEM y OWF almacenan también información sobre las tareas efectuadas.



La siguiente fase se corresponde con la Explotación de los datos del Data Warehouse y/o de los Data Mart.

El Sistema se basa en la utilización de la Intranet de la Seguridad Social para permitir el acceso a la información desde cualquier puesto de trabajo, sin necesidad de instalar un cliente específico. Esto convierte a cualquier herramienta de navegación (explorador) en la herramienta de trabajo básica para la explotación del sistema por parte del usuario final.

Para acceder desde cualquier navegador a los datos almacenados en el Data Warehouse o en el Data Mart, es necesario interponer entre ambos extremos una herramienta de acceso a los datos \ddagger el listener de Oracle 9iAS quien recogerá las peticiones de información del usuario y se las transmitirá a la herramienta de acceso a datos \ddagger Oracle Discoverer.

Oracle Discoverer, es una herramienta de usuario final, fácil de usar, para realizar análisis, informes y consultas ad-hoc. Permite a los usuarios navegar por los datos almacenados en el data warehouse, así como realizar exportaciones a otras aplicaciones como Excel, Lotus, a páginas web y formato Word. La herramienta está construida sobre una capa de usuario final, que organiza la información en "Áreas de Negocio" que independizan al usuario final de la estructura y complejidad de la base de datos, de tal forma que puedan ver los datos en términos conocidos por él. Además provee de una librería de gráficos que permite visualmente mostrar tendencias.



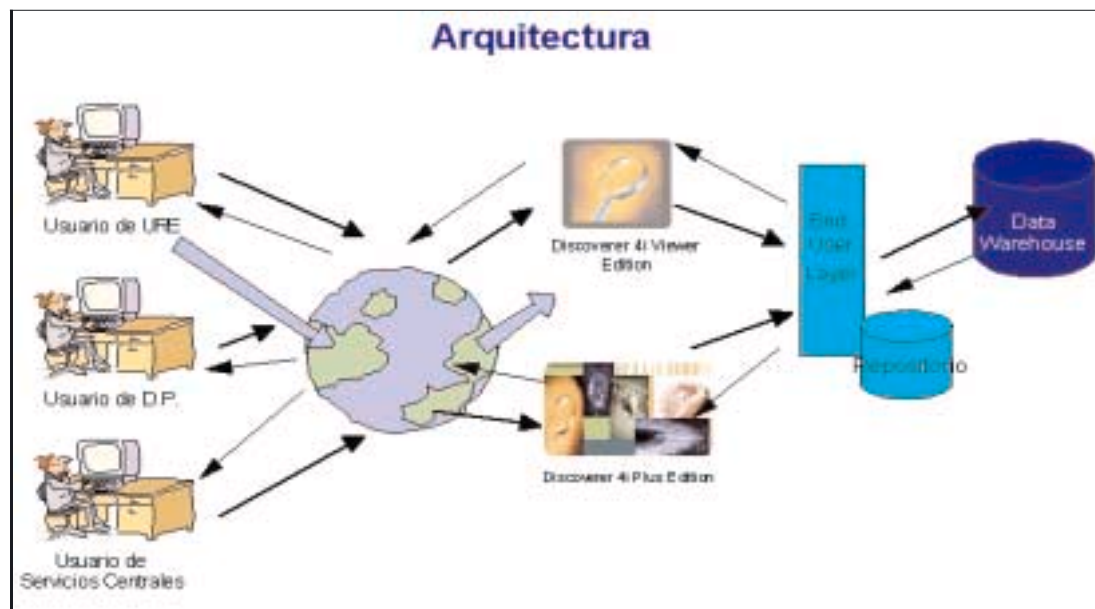
Discoverer mantiene una estadística de los tiempos de ejecución de cada una de las consultas realizadas, esto lo hace para poder dar al usuario la información del tiempo estimado de ejecución de la consulta lanzada, y con ella el usuario decida ejecutarla o no. Dispone de dos ediciones:



Oracle Discoverer Viewer, con la que el usuario puede ejecutar informes predefinidos, mientras que con la edición Oracle Discoverer Plus, son los propios usuarios los que pueden construir sus informes.



El acceso a Oracle Discoverer quedará representado por el usuario en una dirección web o URL determinada. Los niveles de seguridad implantados para acceder a los datos siguen la normativa corporativa de SILCON. En la actualidad se han definido tres niveles de confidencialidad de datos, Perfil de URE, perfil de Dirección Provincial y perfil de Servicios Centrales, de tal manera que los usuarios sólo tendrán acceso a los datos que le correspondan según el perfil que se define cuando se dé de alta en la aplicación.



Arquitectura global de acceso a la información

La última fase es la de Presentación de la Información al Usuario. Como ya se ha citado anteriormente, el usuario accede a las aplicaciones de consulta al Data Warehouse a través del portal de la Intranet de la Seguridad Social, identificándose una vez cuando entra en la Intranet, y si este está dado de alta en la aplicación SINUE, accederá a la misma



desde la pantalla de aplicaciones. De ahí se llama a una URL que corresponde a un portal diseñado con la herramienta Oracle Portal, y que tiene el acceso a las diferentes herramientas de consulta del Data Warehouse (Oracle Discoverer ediciones Viewer y Plus), así como a los cuadros de mando, que corresponden a informes estáticos colgados en el portal de acceso y que van dirigidos principalmente a la Dirección.



imagen anterior corresponde a un informe realizado con Discoverer Plus.



CONCLUSIONES:

Para conseguir que un proyecto de esta envergadura tenga éxito hay que tener en cuenta una serie de factores fundamentales

- Visión del negocio proporcionada por el usuario ≠ Toma de requerimientos en términos cercanos al usuario no a la tecnología.
- Calidad y fiabilidad de la información
- Ofrecer resultados a corto plazo. Lo ideal es facilitar al usuario información después de 6 meses.
- Desarrollo incremental aprendiendo de la experiencia.

Los “Beneficios” que se pueden obtener son entre otros:

- Visión Espacial
 - Se pueden implementar con facilidad representaciones de la información con herramientas GIS (Sistemas de Información Geográfica)
 - La información se puede distribuir por áreas de negocio, lo que facilita su acceso.
- Visión Temporal
 - Se consigue obtener conocimiento histórico a través de comparativas y evolutivos con el mismo esfuerzo que si la información fuera de hoy. Visiones predictivas, simulaciones de escenarios, etc.
 - Obtener conocimiento puntual y oportuno (debido a los tiempos cortos de respuesta) del Negocio.