mides
Ministerio de Desarrollo Social

SIIAS
SISTEMA DE INFORMACIÓN INTEGRADA DEL ÁREA SOCIAL

Sistema de análisis de información

Resumen de metodología técnica

Tabla de Contenidos

Acrónimos utilizados

Acrónimo	Significado
BI	Business Intelligence (Inteligencia de negocios)
DM	Data Marts (Repositorio de datos para análisis de un área de información específica)
DW	Data Warehouse (Repositorio de datos para análisis)
ETL	Extraction, Transformation and Loading (Procesos de extracción, transformación y carga)
ODS	Operational Data Store (Área de datos operacionales)
OLAP	On-Line Analytical Processing (Procesamiento de análisis de datos en línea)

1 Arquitectura general de una solución de BI y DW

La Figura 2.1 muestra un esquema de la arquitectura general de una solución de BI y DW.

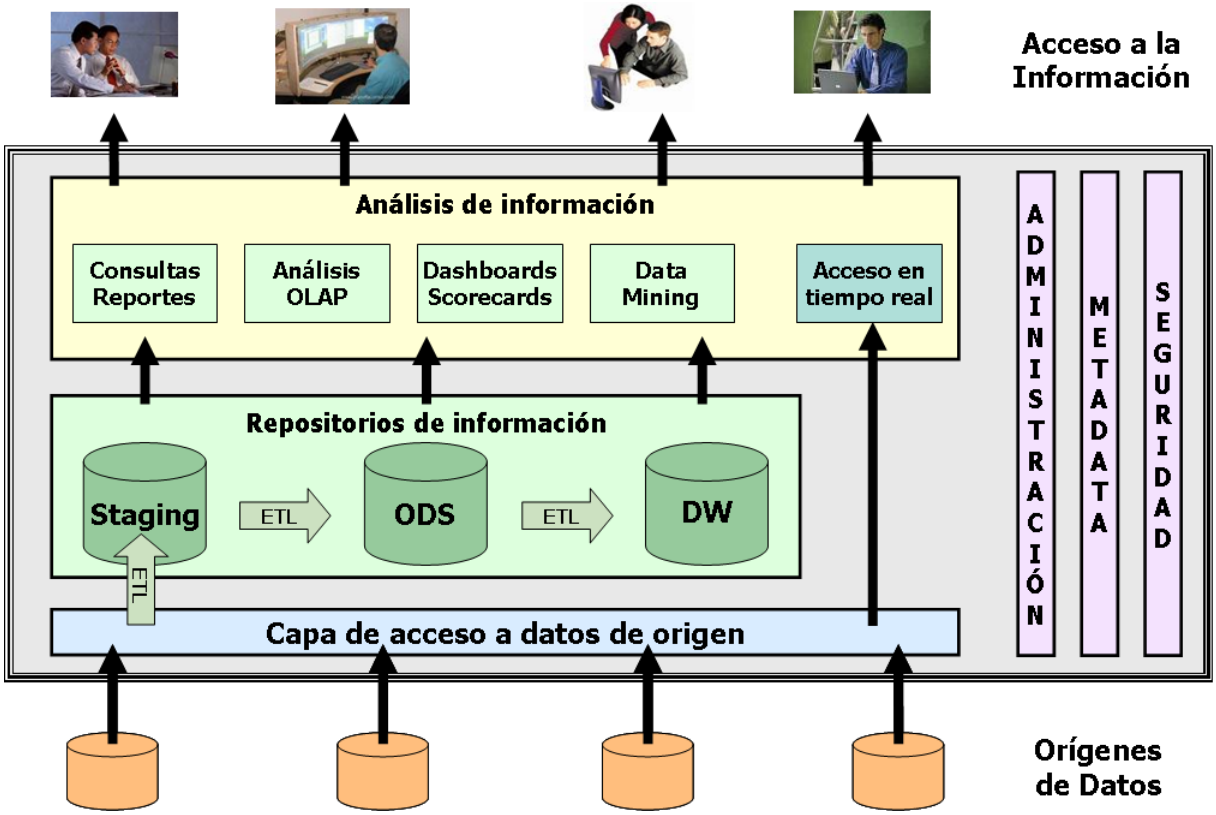

Figura 2.1: Arquitectura general de una solución de DW y BI

En los capítulos siguientes se describen y detallan estos componentes.

2 Orígenes y extracción de datos

Un sistema de DW puede recibir e integrar datos de distintos orígenes y formatos. En cada caso es necesario analizar y diseñar una solución acorde a la realidad, en función de distintos factores y características de los sistemas fuentes. Más allá de esta consideración, como principio general, la extracción de los datos debe mantener una cierta independencia del funcionamiento operativo de los sistemas de origen, de manera de no entorpecer el normal funcionamiento de los mismos.

Es por ello que una buena práctica es crear vistas y procesos que extraigan los datos desde las bases de datos de origen, a archivos de texto, los cuales serán transferidos luego al ambiente de la solución de DW.

En este capítulo se detallan algunas características de los procesos de extracción y los archivos de datos generados.

2.1 Procesos de extracción

Los procesos de extracción incluyen la creación de vistas y procedimientos almacenados, a través de los cuales se realizan las consultas sobre las bases de datos operacionales donde se encuentran los datos a extraer.

El código de las vistas y procedimientos debe ser lo más estándar e independiente de herramientas posible, de manera de garantizar su usabilidad, cualquiera sea la plataforma utilizada. Esto es una buena práctica de diseño e implementación, pensando en posibles migraciones, cambios y actualizaciones que puedan sufrir los sistemas de origen con el paso del tiempo.

Por otro lado, los procesos de extracción deben incidir lo menos posible en la operativa y el normal funcionamiento de los sistemas de producción, por lo que el análisis y diseño debe tener en cuenta, entre otros, los siguientes factores fundamentales:

- Performance de las operaciones de extracción
- Ventanas de tiempo en las cuales el uso de los sistemas de producción es menor (o nulo)

Estos procesos deben ser desarrollados por los administradores de las bases de datos de origen, o en su defecto, por personal técnico que tenga buen conocimiento de las mismas. Se tomará como entrada la especificación de los datos requeridos para el Sistema de Análisis de Información.

2.2 Archivos de datos de extracción

En la Tabla 3.1 se especifican las características recomendadas para los archivos de datos que son generados en las extracciones desde los sistemas de origen.

Tipo de archivo	Texto ANSI (.txt)
Nombres de campos en primera fila	Sí
Delimitador de filas	{CR}{LF} (retorno de carro y salto de línea)
Delimitador de campos	Tabulador
Calificador de texto	Ninguno

Tabla 3.1: Formato de los archivos de datos de extracción

3 Repositorios de información

En una solución de BI y DW tenemos varios repositorios de información:

- Área de volcado inicial de datos ("Staging")
- Operational Data Store (ODS)
- Data Warehouse y/o Data Marts (DW, DM)
- Bases de datos multidimensionales ("cubos" OLAP)
- Repositorio de administración y operación

3.1 Bases de datos de Staging

Los datos son volcados inicialmente en este repositorio, desde los archivos de texto que contienen los datos exportados por los sistemas de origen.

El objetivo del volcado a Staging es copiar los datos tal cual vienen de origen, sin ningún control de calidad, ni transformación, ni formateo. Incluso es recomendable que los campos sean todos de tipo alfanumérico, para que no haya problema con los formatos, los cuáles se verificarán en el paso siguiente (carga al ODS).

En el área de Staging, se recomienda tener un esquema de base de datos por cada sistema origen, nombrándolos "*Staging_X*", donde "*X*" es un nombre que referencia al sistema origen en cuestión.

Pueden existir datos creados específicamente para el Sistema de Análisis, y que no estén almacenados en ninguno de los sistemas fuentes. Se recomienda que estos datos se carguen en esquemas distintos a los correspondientes a los sistemas de origen, nombrándolos de forma de poder identificar que son creados para el Sistema de Análisis. Estos datos también serán cargados en las tablas a partir de archivos de texto (creados manualmente o a partir de alguna aplicación desarrollada a efectos de realizar el mantenimiento de los mismos).

El área de Staging es "volátil", es decir, en cada carga se elimina toda información almacenada en cargas anteriores.

3.2 Operational Data Store (ODS)

El ODS es un área de datos en la cual se almacena la información operacional de los distintos sistemas, con el mismo nivel de detalle que en las aplicaciones originales, sin agregaciones ni sumarizaciones.

En el ODS, se recomienda tener un esquema de base de datos por cada esquema de Staging (relativos a los sistemas de origen, más aquellos correspondientes a los datos creados especialmente para el Sistema de Análisis), además de uno por cada integración de sistemas que se realice. Los nombres pueden ser "*ODS_XX*", donde "*XX*" sea un nombre que haga referencia a los datos correspondientes allí almacenados.

3.3 Data Warehouse y Data Marts

El DW es un repositorio de información orientada al análisis y a la toma de decisiones, que se basa en cuatro principios fundamentales:

- **Integración** – debe garantizar la correcta integración de los datos provenientes de distintas fuentes, resolviendo temas de unicidad, repetición, inconsistencia y otros aspectos que hacen a la calidad de la información almacenada

- **Orientado a temas o entidades de negocio** – los objetos y componentes del DW responden a los elementos conceptuales de la realidad a analizar, y no a la operativa transaccional del sistema

- **Historización de la información** – poder analizar los distintos estados de la información a través del tiempo es fundamental en un DW, aunque a nivel transaccional no sea de relevancia o no esté resuelto. Un concepto importante a tener presente en relación a esta característica es lo que se denomina "*slowly changing dimensions*"

- **Permanencia en el tiempo** – la información del DW debe permanecer con el tiempo, y no ser volátil, ya que los análisis no siempre son sobre información reciente, sino que frecuentemente interesa realizar análisis y comparaciones contra información histórica

Los esquemas de datos de un DW corresponden a **modelos multidimensionales**, ya sea de tipo "star schema" ("esquema estrella") o "snowflake" ("copo de nieve").

En estos esquemas, las tablas principales ("**fact tables**") reflejan los "hechos" relevantes de la realidad estudiada. A estos hechos se asocian:

- **Variables** – elementos o entidades que permiten analizar la información por distintos criterios o ejes; se almacenan en las llamadas tablas de dimensiones ("**dimension tables**")

- **Medidas** – valores que cuantifican los hechos

A su vez, las dimensiones o entidades de análisis pueden ser organizadas en distintos niveles de agrupamientos o agregaciones, a los que llamamos **jerarquías**.

3.4 Bases de datos multidimensionales

Son almacenamientos de datos especiales, diseñados y optimizados para el análisis de información multidimensional, también denominados "cubos" OLAP (On-Line Analytical Processing).

Gran parte de los indicadores e información de un sistema de DW y BI se obtienen de estas bases de datos o cubos multidimensionales, que garantizan mejores tiempos de respuesta en las consultas.

Cada base de datos multidimensional será creada con un nombre que la identifique y la represente.

3.4.1 Orígenes de datos

Los orígenes de datos de las bases multidimensionales son las bases de datos relacionales del Data Warehouse y/o los Data Marts.

3.4.2 Dimensiones

Las dimensiones se cargan a partir de las vistas de publicación que acceden a los datos de las tablas dimensionales del DW y/o DM.

La dimensiones se nombran en relación a lo que representan, pero nombrando con el mismo prefijo a las jerarquías alternativas de una misma dimensión (Ej: "Niño.Edad" y "Niño.Pais").

3.4.3 Medidas

Las medidas se calculan a partir de los campos de las tablas de hechos que almacenan los valores a "medir" en la realidad analizada.

Pueden ser medidas simples, calculadas directamente a partir de los campos, o medidas calculadas en función de otras medidas.

Algunas de las propiedades a definir sobre las medidas son:
- Formato de los campos de origen

- Funciones de sumarización para los niveles superiores (funciones de agregación o "rollup")
- Formato de visualización
- Si son visibles o no al usuario final (pueden ser medidas auxiliares para el cálculo de otras medidas)

3.4.4 Cubos

Los cubos son conjuntos de dimensiones y medidas, y sus nombres se elijen según la realidad que representan.

Pueden ser cubos físicos o virtuales, pero esto depende de las herramientas y tecnologías utilizadas para su implementación.

3.5 Repositorio de administración y operación

En este repositorio (esquema único, independiente de los otros) se almacenan los objetos relativos a la administración y operación del Sistema de Análisis de Información.

Algunos de los elementos que contiene este repositorio son:
- Tablas y vistas de parámetros y configuraciones del sistema (por ejemplo, el período de fechas a incluir en determinada carga de datos, el último período cargado, etc.)
- Tablas y vistas de logs con las trazas de las ejecuciones de los procesos de carga
- Tablas y vistas de auditoría del sistema
- Procedimientos almacenados para realizar el mantenimiento y actualización de estas tablas y vistas

4 Procesos de transformación y carga

Mediante estos procesos se realiza la transformación y carga de información en los repositorios del Sistema de Análisis, a partir de los datos de los sistemas de origen.

Los distintos procesos de carga se resumen en los siguientes pasos:

- Carga de Staging
- Carga de los ODS
- Carga del DW y los DM
- Carga de las bases de datos multidimensionales (cubos)

Las características de cada uno de estos repositorios están descriptas en secciones anteriores.

Estos procesos incluyen varias etapas o subprocesos, asociados a la transferencia desde las fuentes, la validación y formateo de datos, las transformaciones y carga de información en los repositorios del DW, y la generación de las bases de datos multidimensionales (cubos).

El ciclo completo de carga debe estar integrado y controlado por un único proceso principal, que gestione el flujo de ejecución, disparando los distintos subprocesos en el orden correspondiente.

Los procesos deben incluir los controles y transformaciones necesarios que garanticen la calidad, veracidad y confiabilidad de la información final.

A su vez, teniendo en cuenta los diversos orígenes de datos del sistema, los procesos de carga también deben incluir los mecanismos de integración necesarios, con reglas de validación que permitan llevar la información a un único repositorio que refleje la realidad vista de forma integral.

Este conjunto de tareas es parte del proceso de **gestión de la calidad de datos** ("data quality management"), y es una componente muy importante en los procesos de carga.

Otra tarea fundamental de los procesos de carga son las **transformaciones** a aplicar sobre los datos de origen, para llevarlos al formato y estructura de la información final. Estas transformaciones implican resolver temas de formato, nomenclatura, unicidad, mapeo entre los orígenes y destinos, inconsistencias, datos inexistentes, etc.

Los procesos de carga deben estar diseñados e implementados con buenas prácticas que garanticen la **optimización** de los tiempos de ejecución ("performance"), tanto en las extracciones desde las fuentes, en las transformaciones y controles de calidad, las cargas a los repositorios del DW, y la generación de los cubos .

Todos los procesos de carga deben mantener un registro ("log") de lo que se va ejecutando, estableciendo el estado final de cada paso (si terminó bien o falló), y en caso de fallo, los posibles errores detectados.

5 Análisis de la información

Una solución de BI debe brindar la posibilidad de acceder a la información mediante el uso de distintas tecnologías y herramientas de análisis, con distintas opciones y formatos de visualización.

5.1 Tecnologías de acceso a la información

Algunas de las formas principales de acceso a la información son:

- **Tableros de mando** ("Dashboards" o "Scorecards") – permiten organizar y analizar indicadores de alto nivel de agregación, que dan una visión global de las áreas y temáticas de estudio

- **Análisis OLAP** – facilita la navegación por la información, mediante las operaciones típicas de selección de variables, filtrado, paginación, cambio de nivel de detalle ("drilling"), ajuste por parámetros, y distintas opciones y funciones de visualización de la información (apariencia, sumarización y totalización, "rankings", funciones estadísticas, etc.)

- **Reportes detallados** – análisis de la información a un bajo nivel de detalle, permitiendo la profundización y análisis causales específicos, o listados con fines informativos

- **Data Mining** – permite realizar búsquedas de patrones, tendencias y comportamientos en grandes volúmenes de información

Todas estas formas de acceso a la información se realizan sobre las bases de datos del sistema de DW (relacionales y/o multidimensionales), pero también debe existir la posibilidad de consultar la información actual (acceso en "**tiempo real**"), para poder tener la visión exacta al momento de realizar determinada consulta. Esta funcionalidad, si bien es tecnológicamente posible, tiene que considerar los tiempos de respuesta, por lo que debe estar restringida a determinadas consultas puntuales.

5.2 Opciones de visualización

Un sistema de BI debe permitir visualizar la información en distintos formatos, entre los cuales se encuentran:

- Cuadros o tablas
- Gráficas
- Mapas (cuando esté incluida alguna variable geográfica que permita georreferenciación)

El usuario debe tener la posibilidad de imprimir los resultados de las consultas, y también exportar a formatos conocidos (pdf, xls, txt, xml, etc.).