

Procesos ETL

La Base de la Inteligencia de Negocio

```
    // Implementación de la función de cálculo de la medida
    public double calculateMeasure(Measure measure, double value) {
        if (measure.getMetric() == null) {
            return value;
        }
        switch (measure.getMetric()) {
            case "AVERAGE": return calculateAverage(measure, value);
            case "SUM": return calculateSum(measure, value);
            case "COUNT": return calculateCount(measure, value);
            case "MIN": return calculateMin(measure, value);
            case "MAX": return calculateMax(measure, value);
            default: return value;
        }
    }

    // Implementación de la función de cálculo de la media
    private double calculateAverage(Measure measure, double value) {
        if (measure.getMetric() == null) {
            return value;
        }
        if (measure.getMetric().equals("AVERAGE")) {
            return calculateAverage(measure, value);
        }
        return value;
    }

    // Implementación de la función de cálculo de la suma
    private double calculateSum(Measure measure, double value) {
        if (measure.getMetric() == null) {
            return value;
        }
        if (measure.getMetric().equals("SUM")) {
            return calculateSum(measure, value);
        }
        return value;
    }

    // Implementación de la función de cálculo del conteo
    private double calculateCount(Measure measure, double value) {
        if (measure.getMetric() == null) {
            return value;
        }
        if (measure.getMetric().equals("COUNT")) {
            return calculateCount(measure, value);
        }
        return value;
    }

    // Implementación de la función de cálculo del mínimo
    private double calculateMin(Measure measure, double value) {
        if (measure.getMetric() == null) {
            return value;
        }
        if (measure.getMetric().equals("MIN")) {
            return calculateMin(measure, value);
        }
        return value;
    }

    // Implementación de la función de cálculo del máximo
    private double calculateMax(Measure measure, double value) {
        if (measure.getMetric() == null) {
            return value;
        }
        if (measure.getMetric().equals("MAX")) {
            return calculateMax(measure, value);
        }
        return value;
    }
```

Índice

Definición Proceso ETL	03
Proceso de extracción, proceso de transformación y proceso de carga	04
Procesamiento en herramientas ETL	07
Evolución de los procesos ETL	08
Porqué se usan herramientas ETL?	09
Desafíos para los procesos y herramientas de ETL	10
La integración inteligente de datos facilita la agilidad del negocio	12
Características herramientas ETL	14
Evaluar herramientas ETL	16
Pros y contras del desarrollo personalizado vs herramienta ETL	17
¿ETL O ELT?	19

Definición de Proceso ETL

El proceso ETL es una parte de todo proceso de integración de datos. Su función tiene gran relevancia, ya que completa el resultado del desarrollo de aplicaciones y sistemas imprimiendo la cohesión necesaria.

La palabra ETL corresponde a las siglas en inglés de:

- Extraer: extract.
- Transformar: transform.
- Y Cargar: load.

Aplicaciones de los procesos ETL

Gracias a los procesos ETL es posible que cualquier organización:

- Mueva datos desde una o múltiples fuentes.
- Reformatee esos datos y los limpie, cuando sea necesario.
- Proceda a su carga en otro lugar, como puede ser una base de datos, un data mart o un data warehouse.
- Analice esos datos una vez alojados en destino.
- O los emplee en otro sistema operacional para apoyar un proceso de negocio, cuando ya están cargados en su ubicación definitiva.

Otros usos de los procesos ETL

Los procesos ETL no sólo se utilizan cuando sobreviene la aparición de nuevas aplicaciones que se han de incorporar a las rutinas de la organización, sino que también es frecuente emplearlos para la integración con sistemas heredados.

Cuando se habla de sistemas heredados se está haciendo referencia a las aplicaciones antiguas que existen en el entorno de la empresa. Muchas veces, estos sistemas se deben integrar con nuevos aplicativos, por ejemplo con ERPs.

La principal dificultad que puede presentarse en este tipo de situaciones es que la tecnología utilizada en estas aplicaciones antiguas complique la integración con los nuevos programas.

Proceso de Extracción, Proceso de Transformación y Proceso de Carga

Cualquier proceso ETL consta de tres fases: extracción, transformación y carga. Es necesario conocer el funcionamiento y claves de cada una de estas etapas, sin embargo, aún es más decisivo comprender las medidas de seguridad y cautelas que se deben tener en cuenta a la hora de llevarlas a cabo. Conocer estas medidas es la única forma de prevenir situaciones cuyas consecuencias pudieran afectar al sistema y a su normal funcionamiento.

A continuación se resumen los aspectos más importantes de cada uno de estos procesos.

Proceso de extracción

Para llevar a cabo de manera correcta el proceso de extracción, primera fase de ETL, hay que seguir los siguientes pasos:

Extraer los datos desde los sistemas de origen.

- Analizar los datos extraídos obteniendo un chequeo.
- Interpretar este chequeo para verificar que los datos extraídos cumplen la pauta o estructura que se espabia. Si no fuese así, los datos deberían ser rechazados.
- Convertir los datos a un formato preparado para iniciar el proceso de transformación.

Qué hay que tener en cuenta durante el proceso de extracción

Es necesario extremar la cautela en esta fase del proceso de ETL que es la extracción, por lo que se debe tener en cuenta que:

- En el momento de la extracción, análisis e interpretación: los formatos en que se presenten los datos o los modos como éstos estén organizados pueden ser distintos en cada sistema separado, ya que la mayoría de los proyectos de almacenamiento de datos fusionan datos provenientes de diferentes sistemas de origen.
- En el momento de la conversión de datos: conviene recordar que los formatos de las fuentes normalmente se encuentran en bases de datos relacionales o ficheros planos, pero pueden incluir bases de datos no relacionales u otras estructuras diferentes.

Sin embargo, la medida más importante a considerar sería el exigir siempre que la tarea de extracción cause un impacto mínimo en el sistema de origen. Este requisito se basa en la práctica ya que, si los datos a extraer son muchos, el sistema de origen se podría ralentizar e incluso colapsar, provocando que no pudiera volver a ser utilizado con normalidad para su uso cotidiano.

Para evitar este impacto y sus consecuencias, en sistemas grandes, las operaciones de extracción suelen programarse en horarios o días donde la interferencia con el sistema y su uso sea nula o mínima.

Proceso de transformación

La fase de transformación de un proceso de ETL aplica una serie de reglas de negocio o funciones sobre los datos extraídos para convertirlos en datos que serán cargados. Estas directrices pueden ser declarativas, pueden basarse en excepciones o restricciones, pero, para potenciar su pragmatismo y eficacia, hay que asegurarse de que sean:

- Declarativas.
- Independientes.
- Claras.
- Inteligibles.
- Con una finalidad útil para el negocio.

El lado más práctico del proceso de transformación

En ocasiones será necesario realizar alguna pequeña manipulación de los datos, sin embargo, y dependiendo siempre de las fuentes de datos, a veces lo que hará falta será aplicar algunas de las siguientes transformaciones:

- Seleccionar sólo ciertas columnas para su carga (estableciendo que, por ejemplo, las columnas con valores nulos no se carguen).
- Traducir códigos (puede suceder que la fuente almacene una "H" para hombre y "M" para mujer pero el destino tenga que guardar los registros como "1" para hombre y "2" para mujer).
- Codificar valores libres (en la práctica y siguiendo el caso anterior, consistiría en convertir "Hombre" en "H" o "Sr" en "1").
- Obtener nuevos valores calculados (por ejemplo, total_venta = cantidad * precio).
- Unir datos de múltiples fuentes (que pueden ser búsquedas, combinaciones, etc.).
- Calcular totales de múltiples filas de datos (como las ventas totales de cada región).
- Generar campos clave en el destino.
- Transponer o pivotar (girando múltiples columnas en filas o viceversa).
- Dividir una columna en varias (esta acción permitiría transformar la columna "Nombre: García, Miguel" en dos columnas "Nombre: Miguel" y "Apellido: García").
- Aplicar para formas simples o complejas, la acción que en cada caso se requiera, como por ejemplo:
 - Datos OK: entregar datos a la siguiente etapa (fase de carga).
 - Datos erróneos: ejecutar políticas de tratamiento de excepciones.

Proceso de carga

En esta fase, los datos procedentes de la fase anterior (fase de transformación) son cargados en el sistema de destino. Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones diferentes. Por ejemplo, en algunas bases de datos será necesario sobreescibir la información antigua con nuevos datos mientras que en otras, bastará con resumir las transacciones y almacenar un promedio de la magnitud considerada.

Los data warehouse mantienen un historial de los registros, de manera que es posible en todo momento hacer una auditoría de los mismos. Esto permite disponer de un rastro de toda la historia de un valor a lo largo del tiempo.

Desarrollo del proceso de carga de datos

Existen dos formas básicas de desarrollar el proceso de carga:

- Acumulación simple: esta manera de cargar los datos consiste en realizar un resumen de todas las transacciones comprendidas en el período de tiempo seleccionado y transportar el resultado como una única transacción hacia el data warehouse, almacenando un valor calculado que consiste típicamente en un sumatorio o un promedio de la magnitud considerada. Es la forma más sencilla y común de llevar a cabo el proceso de carga.

- Rolling: este proceso sería el más recomendable en los casos en que se busque mantener varios niveles de granularidad. Para ello se almacena información resumida a distintos niveles, correspondientes a distintas agrupaciones de la unidad de tiempo o diferentes niveles jerárquicos en alguna o varias de las dimensiones de la magnitud almacenada (por ejemplo, totales diarios, totales semanales, totales mensuales, etc.).

Sea cual sea la manera elegida para desarrollar el proceso, hay que tener en cuenta que esta fase interactúa directamente con la base de datos de destino y, por eso, al realizar esta operación se aplicarán todas las restricciones que se hayan establecido. Si están bien definidas, la calidad de los datos en el proceso ETL quedará garantizada.

Ejemplos de estas restricciones pueden ser:

- Valores únicos.
- Integridad referencial.
- Campos obligatorios.
- Rangos de valores.

Procesamiento en Herramientas ETL

Un desarrollo reciente en el software ETL es la aplicación de procesamiento paralelo. Este avance ha permitido desarrollar una serie de métodos que mejoran el rendimiento general de los procesos ETL cuando se trata de grandes volúmenes de datos.

Existen principalmente tres tipos de paralelismo que se pueden implementar en las aplicaciones ETL. No sólo no son excluyentes, sino que además pueden combinarse para llevar a cabo una misma operación ETL:

- a. Paralelismo de datos: consiste en dividir un único archivo secuencial en pequeños archivos de datos para proporcionar acceso paralelo.
- b. Paralelismo de segmentación (pipeline): se basa en permitir el funcionamiento simultáneo de varios componentes en el mismo flujo de datos. Un ejemplo de ello sería buscar un valor en el registro número 1 a la vez que se suman dos campos en el registro número 2.
- c. Paralelismo de componente: este tipo de procesamiento consiste en el funcionamiento simultáneo de múltiples procesos en diferentes flujos de datos para el mismo puesto de trabajo.

Dificultades en el procesamiento en herramientas ETL

Actualización y sincronización son los caballos de batalla de esta fase del proceso. La convivencia de distintos tipos de datos que provienen de orígenes diferentes plantea esta dificultad y para superarla es necesario:

Que los datos que se carguen sean relativamente consistentes, o lo que es lo mismo:

- Que tengan sentido.
- Que su contenido esté acorde a las reglas de negocio.
- Que estén actualizados.

Que las fuentes estén sincronizadas, por lo que hay que tener en cuenta los ciclos de actualización de las bases de datos de origen, para lo cual puede ser necesario:

- Detener ciertos datos momentáneamente.
- Establecer puntos de sincronización y de actualización, cuando un almacén de datos necesite actualizarse con los contenidos en un sistema de origen.

Evolución de los procesos ETL

Hasta hace unos años, al hablar de ETL sólo se hacía referencia a lo siguiente:

- Procesos de extracción de datos.
- Procesos de transformación de datos.
- Procesos de carga de datos.
- Gestión de metadatos.
- Servicios de administración y operacionales.

Actualmente, es necesario hablar de integración de datos (Data Integration) como evolución de los procesos ETL. Aspectos tan importantes y decisivos para un buen resultado a nivel de sistema como la calidad o el perfil del dato, se han incorporado a la definición de ETL y por eso hoy día en ella se encuentran incluidos todos los siguientes puntos:

- Servicios de acceso a datos.
- Data profiling.
- Data quality.
- Procesado de datos operacionales.
- Servicios de transformación: CDC, SCD, validación, agregación.
- Acceso en tiempo real.
- ETL
- Transporte de datos.
- Gestión de metadatos.
- Servicios de entrega.

¿Por qué se usan herramientas ETL?

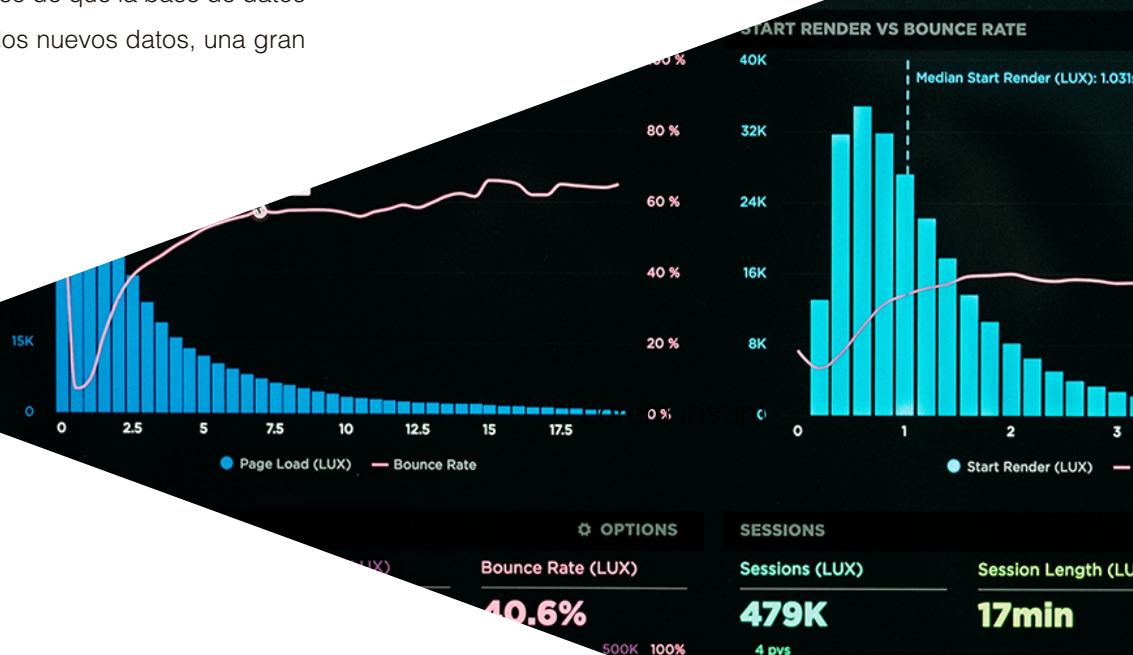
El uso de herramientas ETL responde a criterios de sincronización, conectividad, actualización, idoneidad y detalle. Sin embargo, puede que el motivo más importante que impulsa a una organización a optar por esta alternativa sea el económico.

Para ahorrar tiempo y dinero en el desarrollo de un data warehouse, la mejor solución es eliminar la necesidad de codificación manual. Las razones de esta decisión son múltiples, pero pueden resumirse en tres circunstancias (bastante frecuentes en la práctica) y que se describen a continuación:

- Dificultades a la hora de conectar sistemas: es muy difícil para los administradores de bases de datos la conexión entre diferentes sistemas de bases de datos sin el uso de una herramienta externa.
- Actualizaciones: en el caso de que la base de datos se altere o sean integrados nuevos datos, una gran

cantidad de código manual tiene que ser completamente rehecho.

- Rendimiento: la codificación manual sigue siendo la forma más común de la integración de los datos de hoy, a pesar de que requiere horas de desarrollo y experiencia para crear un sistema realmente eficiente.



Desafíos para los procesos y herramientas ETL

Los procesos ETL pueden ser muy complejos. Un sistema ETL mal diseñado puede causar importantes problemas operativos. Puede suceder que, en un sistema operacional, el rango de valores de los datos o la calidad de éstos no coincidan con las expectativas de los diseñadores a la hora de especificarse las reglas de validación o transformación.

Para evitar este tipo de situaciones, es recomendable realizar durante el análisis un examen completo de la validez de los datos (Data Profiling) del sistema de origen, para identificar las condiciones necesarias para que los datos puedan ser tratados adecuadamente por las reglas de transformación especificadas. Esto conducirá a una modificación de las reglas de validación implementadas en el proceso ETL.

Normalmente los Data Warehouse son alimentados de manera asíncrona desde distintas fuentes, que obedecen a propósitos muy diferentes. El proceso ETL es clave para lograr que los datos extraídos asíncronamente de orígenes heterogéneos se integren finalmente en un entorno homogéneo, todo ello sin perder la fluidez y agilidad que se desea.

La escalabilidad de los sistemas y los procesos ETL

La escalabilidad de un sistema de ETL durante su vida útil tiene que ser establecida durante el análisis. En concreto, el término escalabilidad hace referencia a la capacidad del sistema para reaccionar y adaptarse, para crecer y para manejar con fluidez el crecimiento continuo de trabajo sin que ello suponga un menoscabo en su calidad. Estas capacidades incluyen la comprensión de los volúmenes de datos, que tendrán que ser procesados según los acuerdos de nivel de servicio (SLA: Service Level Agreement).

El tiempo disponible para realizar la extracción de los sistemas de origen podría cambiar, lo que implicaría que la misma cantidad de datos tendría que ser procesada en menos tiempo. Algunos sistemas ETL son escalados para procesar varios terabytes de dato, siendo capaces de actualizar un Data Warehouse que puede contener decenas de terabytes de datos.

El aumento de los volúmenes de datos que pueden requerir estos sistemas pueden hacer que los lotes que se procesaban a diario pasen a

procesarse en micro-lotes (varios al día) o incluso a la integración con colas de mensajes o a la captura de datos modificados (CDC: Change Data Capture) en tiempo real para una transformación y actualización continua.

La funcionalidad de las herramientas ETL

Las herramientas ETL no tienen por qué utilizarse sólo en entornos de Data Warehousing o construcción de un Data Warehouse, sino que pueden ser útiles para multitud de propósitos, como por ejemplo:

- Tareas de Bases de datos: que también se utilizan para consolidar, migrar y sincronizar bases de datos operativas.
- Migración de datos entre diferentes aplicaciones por cambios de versión o cambio de aplicativos.
- Sincronización entre diferentes sistemas operacionales (por ejemplo, entre nuestro entorno ERP y la web de ventas).
- Consolidación de datos: sistemas con grandes volúmenes de datos que son consolidados en sistemas paralelos, ya sea para mantener históricos o para llevar a cabo procesos de borrado en los sistemas originales.
- Interfases de datos con sistemas externos: como el envío de información a clientes o proveedores. También servirían para la recepción, proceso e integración de la información recibida.
- Interfases con sistemas Frontoffice: serían interfases de subida/bajada con sistemas de venta. Otros cometidos: como la actualización de usuarios a sistemas paralelos o la preparación de procesos masivos (tipo mailings o newsletter).

En referencia a este tema, el informe de Gartner hace una comparativa de los productos más importantes del mercado, posicionándolos en su cuadrante según diferentes criterios, y exponiendo las ventajas y factores de riesgo de cada fabricante; por lo que resulta muy útil a la hora de tener acceso a las herramientas ETL más importantes.

La integración inteligente de datos facilita la agilidad del negocio.

Todos los integrantes de una organización necesitan datos para desarrollar su trabajo. Brindar los datos correctos a las personas adecuadas en el instante en que lo necesitan supone satisfacer una necesidad básica en la empresa. Pero, en esta búsqueda por la precisión, la agilidad es una variable importante a tener en cuenta, como así lo son los costes. La integración inteligente de datos es la respuesta a todas estas cuestiones.

La inteligencia empresarial se apoya en datos. Pero no son un elemento aislado, ya que en su camino confluyen también:

- Las relaciones con los clientes y su administración (CRM).
- El cumplimiento de la normativa vigente.
- La eficiencia a la hora de ser capaz de combinar todos los datos participantes en esta dinámica.

No es sencillo alcanzar la excelencia en la integración porque hoy día, las empresas se mueven en un entorno complejo y cambiante, donde la flexibilidad, un valor en alza, también puede jugar en contra. Concretamente, en lo referente a herramientas.

Los problemas de la integración y cómo solucionarlos mediante herramientas ETL

Muchas veces se busca que la integración cumpla un propósito específico y se crean herramientas ad hoc. Hasta aquí ningún problema. Pero los inconvenientes comienzan a la hora de flexibilizar, cuando se intenta dar un uso más amplio a esas herramientas, ya que esta decisión suele generar resultados poco satisfactorios debido a las siguientes razones:

- La coexistencia de herramientas disímiles.
- La compleja infraestructura en la que se apoyan.
- La fragilidad de la combinación de factores que resulta de la flexibilización (improvisación, en algunos casos).
- La falta de agilidad del sistema.
- Los elevados costes.

Las empresas alcanzan un punto crítico cuando se dan cuenta de que, cambiar la infraestructura de datos lo suficientemente rápido como para mantenerse a la par de las actividades del sector, es prácticamente misión imposible.

Beneficios de las herramientas ETL

La solución más inteligente y eficiente para resolver este complicado problema de diversidad de requisitos para la integración de datos de las empresas, si se quiere lograr la agilidad del negocio pasa por usar alguna de las herramientas ETL. Éstas ofrecen a sus clientes:

- Técnicas de integración de datos en una solución única y lista para usar.
- La posibilidad de elegir la técnica que más convenga (ETL, replicación, federación, búsqueda o integración basada en eventos)
- La seguridad de que se lograrán elaborar flujos de datos flexibles y heterogéneos.
- Unos costes operativos mucho más reducidos que con cualquier otra solución.
- La disminución también de la complejidad, al contar con un marco uniforme para todas las técnicas y al brindar compatibilidad con una gran variedad de fuentes de datos.



Características de las herramientas ETL

En un proceso ETL no todo vale, por eso hay que saber elegir. Para hacerlo con conocimiento de causa es necesario conocer las características más importantes que debe incluir un software ETL. Según Gartner, son las siguientes:

- Conectividad / capacidades de adaptación (con soporte a orígenes y destinos de datos): se refiere a la habilidad para conectar con un amplio rango de tipos de estructura de datos, entre los que podrían incluirse: bases de datos relacionales y no relacionales, variados formatos de ficheros, XML, aplicaciones ERP (sistema de planificación de recursos empresariales), CRM (sistema de gestión de clientes) o SCM (supply chain management - gestión de proveedores), formatos de mensajes estándar (EDI, SWIFT o HL7), colas de mensajes, emails, websites, repositorios de contenido o herramientas de ofimática.
- Capacidades de entrega de datos: suponen la habilidad para proporcionar datos a otras aplicaciones, procesos o bases de datos en varias formas, con capacidades para programación de procesos batch, en - tiempo real o mediante lanzamiento de eventos.
- Capacidades de transformación de datos: habilidad para la transformación de los datos, desde transformaciones básicas (conversión de tipos, manipu-

lación de cadenas o cálculos simples) o transformaciones intermedias (agregaciones, summarizaciones, lookups) hasta transformaciones complejas, como análisis de texto en formato libre o texto enriquecido.

- Capacidades de Metadatos y Modelado de Datos: recuperación de los modelos de datos desde los orígenes de datos o aplicaciones, creación y mantenimiento de modelos de datos, mapeo de modelo físico a lógico, repositorio de metadatos abierto (con posibilidad de interactuar con otras herramientas), sincronización de los cambios en los metadatos en los distintos componentes de la herramienta, documentación, etc.
- Capacidades de diseño y entorno de desarrollo: representación gráfica de los objetos del repositorio, modelos de datos y flujos de datos, soporte para test y, capacidades para trabajo en equipo, gestión de workflows de los procesos de desarrollo, etc.
- Capacidades de gestión de datos (calidad de datos, perfiles y minería)
- Adaptación a las diferentes plataformas hardware y sistemas operativos existentes: mainframes (IBM Z/OS), AS/400, HP Tandem, Unix, Wintel, Linux, Servidores Virtualizados, etc.
- Operaciones y capacidades de administración: habilidades para gestión, monitorización y control de los procesos de integración de datos, como

gestión de errores, recolección de estadísticas de ejecución, controles de seguridad, etc.

- Arquitectura e integración: grado de compactación, consistencia e interoperabilidad de los diferentes componentes que forman la herramienta de integración de datos (con un deseable mínimo número de productos, un único repositorio, un entorno de desarrollo común, interoperabilidad con otras herramientas o vía API), etc.
- Capacidades SOA.

Está claro que la elección de un software de ETL puede ser una tarea compleja y, debido a la repercusión que ello tendrá en el desarrollo posterior de un proyecto, es una decisión que no puede tomarse a la ligera. Una elección correcta garantiza:

- Un aumento en la velocidad de desarrollo del proceso.
- La descentralización del control de la ejecución y de todos los procesos.
- Una mayor agilidad en el negocio.

Evaluar herramientas ETL

Una vez se tienen claras las características que debe reunir la herramienta ETL que se desea adquirir y se ha descartado ya la posibilidad de emplear algún otro proceso de gestión de datos, es el momento de evaluar las distintas opciones disponibles en el mercado, para poder tomar la decisión de compra. Hay que tener en cuenta que esta elección:

- Tiene que centrarse en una herramienta ser capaz de adaptarse a las necesidades de desarrollo actuales y futuras de los usuarios de negocio, independientemente del índice de crecimiento.
- Tiene que ser tomada (y utilizada, una vez adquirida la herramienta) de manera eficiente para ganar en eficacia y ser capaz, incluso, de compensar la necesidad de plantilla adicional.

Pasos a seguir para comprar una herramienta ETL

Conviene considerar algunos pasos que es recomendable seguir antes de comprar herramientas ETL. Los más importantes son los siguientes:

- Preguntarse por la cantidad de datos (en GB) para un cierto período de procesamiento de la herramienta.
- Comprobar la conectividad con el tipo de fuentes de datos en la herramienta ETL.

- Comprobar el formato de datos que solicitado, ya sea texto o CSV, XML, las bases de datos (como Oracle, MySQL, SQL Suite, DB2, Sybase), EDI, HIPAA, dseACORD AL3, cualquier longitud fija formato o cualquier otro formato.
- Hacer cumplir las reglas de validación de datos mediante la especificación de las mismas en los procesos.
- Documentar y formalizar los flujos de datos y reglas de asignación.
- Preguntarse cuáles son las funciones de registro y control de las cargas y la forma de manejar condiciones de error.

Una vez concluidas estas reflexiones y para proceder con la compra, como cliente, es importante:

- Centrarse en las necesidades técnicas clave, para asegurarse de que la referencia que se obtiene es apropiada y útil para el proceso de evaluación.
- Llevar a cabo un proceso de prueba de concepto, que es el punto de partida de una evaluación general a cada uno de los proveedores elegidos.
- A la vista de los datos obtenidos, comparar los productos en el propio entorno, con los propios datos y aplicados al negocio.
- Tomar la decisión final en base al hallazgo de acuerdo con el ejercicio, tras seleccionar el producto que mejor se ajusta a las necesidades de la organización.

Pros y contras del desarrollo personalizado vs herramienta ETL

La cuestión de la elección de un código personalizado (también conocido como código custom o código manual) en comparación con el uso de una herramienta ETL es a la que hay que enfrentarse cada vez que es necesario crear un Data Warehouse (DWH). Cada uno tiene sus ventajas y desventajas.

El código manual en la creación de un data Warehouse
La alternativa del código personalizado es una buena solución a la hora de crear un Data Warehouse, ya que esta opción proporciona a las organizaciones la capacidad de codificar exactamente lo que quieren, expresado en el modo cómo les gustaría que sus programas de transformación quedasen estructurados.

Entre las principales ventajas de la utilización de código personalizado, se encuentran las siguientes:

- Bajo coste (por lo general usan el lenguaje de la casa, por ejemplo: C ++, PL / SQL, Java).
- Idoneidad, ya que el código está construido sólo para sus necesidades.
- Optimización de los programas.
- Disponibilidad, porque esta opción permite construir lo que se quiera, en el momento en que sea necesario.

Sin embargo, este sistema no está exento de desventajas. Además de ser necesaria una amplia base de conocimientos de los programadores, optar por la codificación manual puede presentar los siguientes inconvenientes:

- Es difícil alcanzar la eficacia en cuanto a la gestión y mantenimiento de los programas.
- Si se produce algún cambio en el Data Warehouse, muchos programas podrían verse afectados.
- Esta opción trabaja sin repositorio centralizado de código.
- La codificación manual implica que las capacidades de metadatos sean limitadas.
- Su ciclo de desarrollo es más largo.
- La depuración es más difícil.
- La capacidad de auditoría queda limitada.

Creación de un Data Warehouse mediante una herramienta ETL

El uso de una herramienta ETL representa el otro lado de la ecuación de carga. Una herramienta ETL, por lo general, proporciona una interfaz agradable para los usuarios a la hora de crear y gestionar las transformaciones.

Al igual que la solución anterior, este método cuenta con sus pros y sus contras, aunque las ventajas superan a los inconvenientes. Las más importantes serían:

- La obtención de una interfaz visual agradable para crear y mantener programas.
- El almacenamiento centralizado de los programas.
- El control de versiones de los programas.
- La simplicidad relativa de la personalización de transformaciones.
- El adecuado soporte de metadatos que proporciona.
- La rápida implementación de transformaciones.
- El sistema de depuración integrado en la mayoría de los productos.
- La posibilidad de transformar la programación.
- La posibilidad de transformar la auditoría.

Frente a los numerosos beneficios de esta opción, aparecen algunas desventajas como por ejemplo:

- El alto coste inicial que implica.
- El conocimiento de usuario limitado de la mayoría de los productos.
- La optimización, que a veces está limitada debido a los métodos de programación genéricos.

A la vista de las características principales de ambas opciones y de las ventajas que conllevan, y teniendo también en cuenta sus aspectos negativos, parece obvio que la balanza se inclina hacia las herramientas

ETL que, más que una alternativa, son en realidad una necesidad en cualquier organización, si no siempre, al menos en algún momento del ciclo de vida de su sistema de almacenamiento de datos.

¿ETL o ELT?

Entendemos ETL como el proceso extracción, transformación y carga de los datos, que es parte del ciclo de vida de una implementación de Business Intelligence. Partiendo de esta premisa, nos damos cuenta que existen ciertas variaciones conceptuales relativas al mismo proceso de ETL, de las que dependerá el rendimiento de los procesos de manejo de los datos. Por ello es necesario considerar las tecnologías aplicadas en cada parte del proceso, de principio a fin.

A modo de resumen, podría decirse que un proceso cualquiera daría comienzo en el origen de los datos (Base de datos, archivos de texto, etc.), continuaría con la intervención de la herramienta de ETL, para concluir en el destino de los Datos (Base de datos) que se disponga.

La herramienta de ETL permitiría:

- Conectarse a la fuente de los datos.
- Hacer la transformación dentro de la misma herramienta.
- Cargar los datos a la base de datos destino.

Entendiendo el concepto E-LT

E-LT podría definirse siguiendo el orden de las iniciales que lo denominan. Así se puede decir que consiste en la extracción, carga y transformación de datos, y se resume en los siguientes tres pasos:

1. Primero: habrá que extraer y cargar los datos de manera "BULK" directamente a una Base de Datos o a unas tablas especialmente creadas para los datos de paso (conocidas también como staging). Esto supone que este medio servirá solo temporalmente, por lo que podrá ser limpiado en cada proceso de carga. Por ello se recomienda hacer transformaciones simples y emplea básica de información.
2. Segundo: cuando la información se halla contenida en staging habrá que proseguir con la elaboración del proceso de transformación de los datos, que posteriormente pasará a la base de datos del Data Warehouse. Esta transformación se hará con el lenguaje propio de la base de datos, por ejemplo T-SQL, PL/SQL.

3. Tercero: una vez que se tienen los datos transformados en los procesos propios de la base de datos, se insertarían en el Data Warehouse. Terminada esta acción, se pueden limpiar los datos de paso, si se cree conveniente.

De esta manera el proceso de transformación queda integrado en el motor de la Base de Datos.

Ventajas de E-LT sobre ETL

Aunque ambos conceptos, E-LT y ETL conducen a un mismo resultado, la diferencia está en el rendimiento y la velocidad de proceso del proceso de carga en cada caso. Las principales ventajas de usar E-LT en vez de ETL serían las relativas a:

- Velocidad de proceso y transformación. La principal ventaja de E-LT es la forma en que trabaja cada herramienta implicada. En el caso de ETL las herramientas de transformación evalúan registro por registro, mientras que en E-LT la transformación se hace en la base de datos que evalúa los registros en lotes.
- Uso de recursos. Otra ventaja de E-LT, es que una base de datos está preparada para la optimización de recursos ya sean de disco, memoria o proceso y esto hace que el rendimiento del proceso sea

administrado por la configuración de la base de datos. Sin embargo, las herramientas de ETL no toman ventaja de la configuración del disco (RAID) ni de la distribución de la memoria y procesador, ya que hacen transformaciones temporales y en muchos casos redundantes.

Cada herramienta nos provee de unas ventajas diferentes. Algunas nos dan mayor facilidad para desarrollar una transformación, aunque no el mejor rendimiento; mientras que en ocasiones sucede al contrario. En la práctica, puede suceder que un cliente que tiene una herramienta E-LT utilice ETL al no saber usar sus ventajas. Por eso, es importante estar informado y conocer el alcance de los recursos de que se dispone, para poder tomar decisiones correctas, obteniendo el mejor rendimiento.



www.powerdata.es

Arenales 1678 Piso 3
Tel (+54 11) 5277-4650
Buenos Aires- Argentina

Padre Mariano N°82 oficina 602 Las Condes
Tel (+56) 2 29363 100
Santiago de Chile

Carrera 16 # 93 A - 16 Oficina 504
Tel (+57 1) 616 7796
Bogotá- Colombia

Av República de Panamá 3545, oficina 1102 San Isidro
Tel (+51) 1 634 4900
Lima- Perú

HAZ CLICK AQUÍ PARA AGENDAR
UNA CONSULTA GRATUITA CON UNO
DE NUESTROS EXPERTOS

linkedin.com/company/powerdata
twitter.com/powerdatagruo