VENTAJAS DEL USO DE HERRAMIENTAS DE ETL SOBRE ANSI SQL

LIC. DIEGO KRAUTHAMER

PROFESO R ADJUNTO INTERINO

UNIVERSIDAD ABIERTA INTERMERICANA (UAI)

SEDE BUENOS AIRES

COMISION DE INVESTIGACION

Abstract

El presente trabajo tiene por objetivo ayudar al lector, consultor, o académico en la decisión de si debe seleccionar una herramienta o software de ETL, o realizar un desarrollo propio basado en ANSI-SQL para llevar a cabo el proceso de Extracción, Transformación y carga de un Proyecto de Inteligencia de Negocios o Business Intelligence. Para abordar está temática partiremos desde el concepto de Inteligencia de Negocios, para posteriormente profundizar sobre las características de cada una de las arquitecturas que por las cuáles podemos optar, para finalmente plasmar en las conclusiones.

Introducción

Tomando como punto de partida que "Business Intelligence es la habilidad para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios. Asociándolo directamente con las tecnologías de la información, podemos definir Business Intelligence como el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada (interna y externa a la organización) en información estructurada, para su explotación directa (reporting, análisis OLTP / OLAP, alertas…) o para su análisis y conversión en conocimiento, dando así soporte a la toma de decisiones sobre el negocio[1]".

El presente trabajo versa sobre las ventajas de la utilización herramientas ETL (Extracción, Transformación y Carga), respecto del uso de ANSI SQL.

Un proyecto de "Inteligencia de Negocios" o "Business Intelligence" posee una serie de etapas bien definidas, empezando por la identificación de los distintos orígenes de datos, para luego pasar a la etapa de extracción, transformación y carga en el almacén de datos o Warehouse, para finalizar con la explotación de la información mediante de una herramienta OLAP (Análisis o Reportes).

Es importante remarcar que la etapa de ETL no es visible para los usuarios finales de BI; en otras palabras nos referimos a que no es visible porque en primer lugar las tareas de transformación y limpieza de datos requieren esfuerzo en la implementación de este tipo de soluciones no se verán en el Sistema de BI Implementado .En palabras de Ralph Kimball "La etapa de ETL consume el 70% de las necesidades de recursos para el desarrollo y mantenimiento de un sistema DW. Además, estos procesos no son solamente un mero traspaso de información de un sistema o otro. Son mucho más, pues pueden dar un valor significativo a los datos[2]".

Asimismo se presentaran las dos arquitecturas posibles en función de la solución de ETL adoptada, destacando las características que posee cada una de ellas, como así también las ventajas y desventajas de cada una de ellas. Posteriormente se analizarán algunas de las herramientas ETL del mercado, realizando una breve descripción de cada una de ellas.

Finalmente se expondrán las conclusiones obtenidas a partir del análisis de las distintas arquitecturas de ETL presentadas durante el presente ensayo.

Arquitecturas

En cualquier organización que tomemos como caso de estudio para la puesta en funcionamiento, es decir para la implementación de una solución de Inteligencia de Negocios o Business Intelligence, veremos que son dos las arquitecturas posibles a utilizar. La primera de ellas

[1] The Datawarehouse ETL Toolkit Ralh Kimball Wiley Publishers 2004

[2] The Datawarehouse ETL Toolkit Ralh Kimball Wiley Publishers 2004

es realizar el ETL a través de un desarrollo basado en una solución ANSI-SQL, y la segunda arquitectura se basa en la utilización de los denominados softwares o herramientas de ETL.

A continuación trataremos con mayor profundidad cada una de las alternativas mencionadas previamente destacando las mayores ventajas de cada una de elllas.

Arquitectura de una solución basada en una herramienta ETL

Como se mencionó en la introducción el uso de una arquitectura basada en una herramienta de ETL consiste en realizar el proceso mediante un software desarrollado por un tercero para realizar el proceso mencionado.

La adopción o no de una herramienta o software de ETL no posee una respuesta sencilla; sin embargo, mencionaremos sus ventajas y desventajas por sobre la alternativa planteada en el apartado anterior.

Características

Las siguientes características[3] son deseables en una herramienta de ETL:

- *Conectividad*: se refiere a los distintos orígenes de datos que da soporte la herramienta, desde base de datos relaciones y no relacionales, distintos formatos de archivos, XML , aplicaciones empresariales (ERP, CRM y SCM), emails, herramientas de ofimática etc.
- *Capacidades de entrega de datos*: es la posibilidad que posee el software de brindar datos a otros softwares, ya sea mediante proceso batch, o en tiempo real.
- *Capacidades de transformación de datos*: como lo dice el título, valga la redundancia, es transformar los datos, desde lo más simple como conversión del tipo de datos, o cálculos simples para mencionar algunos ejemplos, hasta transformaciones más complejas como agregaciones, sumarizaciones o lookups.
- *Capacidades de metadatos y modelado de datos*: es la recuperación de los modelos datos desde sus respectivos orígenes de datos o aplicaciones, mapeo del modelo físico al lógico, documentación, y sincronización de los cambios en los distintos componentes de la herramienta.
- *Capacidades de diseño y entorno de desarrollo*: comprende la representación gráfica de los objetos del repositorio, modelos de datos, y flujos de datos, soporte para el test y debugging, también incluye a las capacidades para trabajo en equipo.

Ventajas[4]

- *Visualización de flujos y autodocumentación*: podemos ver gráficamente cuál es el recorrido de los flujos de datos, desde que los mismos son leídos desde un origen de datos, por ejemplo una base de datos relacional, hasta que son escritos en un tabla de hechos. Autodocumentación significa que como consecuencia de lo primero, a medida que se diseñan gráficamente los ETL's, se esta documentando el desarrollo; además las herramientas poseen "plug-ins" como cuadros de texto que permiten documentar por ejemplo las tareas realizadas en un ETL.
- *Diseño estructurado*: si bien fueron diseñadas para resolver el problema de la carga del Datawarehouse, estas herramientas proveen una metodología que fomenta las buenas prácticas y que beneficiará a aquellos consultores de TI que se encuentran desarrollando sus primeros proyectos de Business Intelligence.

[3] Procesos ETL. Escenarios para el diseño de los procesos.http://churriwifi.wordpress.com/2010/05/01/16-procesos-etl-escenarios-para-el-diseno-de-los-procesos/

[4] Should You Use an ETL Tool. John Mundy. Kimball University. 2008 http://www.kimballgroup.com/html/articles_search/articles2008/0804E.html

- *Facilidad de operación:* Muchas herramientas de este tipo poseen las funcionalidades que permiten el monitoreo y seguimiento de sus operaciones en el ambiente de producción.
- *Limpieza de datos:* posee funcionalidades avanzadas, e incluso superiores a las que posee el Lenguaje ANSI-SQL, que permiten "limpiar" los datos de origen, evitando de esta manera, cargar información errónea o incompleta en el Almacén de Datos.

Desventajas

- *Costo de las licencias:* la organización deberá desembolsar en el inicio el proyecto de Inteligencia de negocios el costo correspondiente. En la mayoría de los casos son software costosos; pero si optará por el uso de una herramienta o software de ETL open source o de código abierto, podría la organización podría ahorrarse el costo de las licencias.
- *Flexibilidad*: nos referimos a que es posible que nos encontremos limitados a las capacidades de "scripting" de la herramienta; en otras palabras podemos entrar ciertas carencias en la herramienta, que normalmente no encontraríamos en un desarrollo de ETL basado en ANSI-SQL
- *Incertidumbre*: muchos equipos de desarrollo carecen del "Know- how" suficiente respecto de las funcionalidades y capacidades de la herramienta de ETL.

Principales Herramientas ETL

En este apartado mencionaremos algunas de las principales herramientas de ETL del mercado:

- *IBM Infosphere Datastage*[5]: es una de los softwares de integración de datos propietario más potentes del mercado y que fue adquirida por IBM hace dos años.
- *Informatica Powercenter*[6]: otra de las herramientas ETL propietarias que ha logrado gran adopción en el mercado durante los últimos años.
- *Pentaho Kettle*[7]: una alternativa open source o de código abierto que forma parte de la suite de BI open source Pentaho.

Seguramente podríamos haber mencionado más herramientas, o haber puesto mayor énfasis en las características de algunos de los productos mencionados previamente; sin embargo no es parte del presente ensayo detallar las características en particular de alguna herramienta de ETL.

Arquitectura de una solución basada en ANSI SQL

Un desarrollo de ETL basado en ANSI-SQL implica, realizar todos los procesos vinculados las Extracción, Transformación y Carga del Warehouse o Almacén de Datos, mediante la codificación de instrucciones en Lenguaje ANSI-SQL.

La elección de un desarrollo de estas características involucra a priori la adopción de un estándar de desarrollo que sea acordado para posteriormente atenuar los problemas de mantenimiento que adolecen este tipo de desarrollos.

Ventajas

- *Herramientas de Debug*: es posible "debugear" los errores mediante la utilización de los disntintos front-ends basados en ANSI-SQL.
- *Reutilización de código*: permite reutilizar gran parte del código fuente ANSI-SQL de otros desarrollos.

---

[5] Sitio Oficial de IBM Infosphere Datastage. http://www-01.ibm.com/software/data/infosphere/datastage/

[6] Sitio Oficial de Informatica Powercenter
http://www.informatica.com/products_services/powercenter/Pages/index.aspx

[7] Sitio Oficial de Pentaho Kettle http://kettle.pentaho.com/

- *Recursos Humanos:* se refiere a la utilización de recursos internos de la organización especializados en Lenguaje SQL, evitando de este modo la contratación de consultores la capacitación de los distintos con el consiguiente ahorro de costos o capacitar
- *Flexibilidad ilimitada:* para desarrollar los procesos vinculados al ETL y a las tareas vinculados a dichos procesos. Un claro ejemplo de esto es que existe mayor facilidad para el manejo de la metadata.

Desventajas

- *Mantenimiento*: sin lugar a dudas los desarrollos que con el correr tiempo se convierten en una gran cantidad de líneas de código, tarde o temprano se vuelven dificultosos de mantener, y este no es el caso de los desarrollos de ETL basados en ANSI-SQL.
- *Documentación*: es un punto muy débil de esta arquitectura, recordamos que las herramientas de ETL permiten la autodocumentación.

Conclusiones

A esta altura deberíamos estar en condiciones de responder a la pregunta del millón, ¿Deberíamos usar un software de ETL?. La respuesta no es sencilla ya que "una de las decisiones más tempranas e importantes que se deberá tomar, es si se optará por un desarrollo ETL a medida, o se utilizará un paquete ETL de software[8]"; lo que se infiere a partir de la frase anterior, es que más allá de la arquitectura ETL que se vaya a seleccionar, no debe tomarse a la ligera la decisión, entonces podríamos, y porque no, tener en cuenta los siguientes "tips" o "sugerencias" a la hora de elegir alguna de las alternativas planteadas durante la presente discusión.

- *Siempre pensar a largo plazo*: debemos tener en cuenta como horizonte, el largo plazo, y por ende, las decisiones que tomemos, deben apuntar con ese horizonte de planeamiento; por ello si tenemos programadores acostumbrados a escribir código, deberíamos tener en cuenta que si adoptamos una herramienta de ETL, no se adaptarán fácilmente a la misma. "Por este motivo muchas organizaciones creen que realizar un desarrollo ETL a medida es una solución razonable[9]"

- *No esperar resultados en forma inmediata*: Los resultados mágicos en el área de Sistemas de Información no existen. No podemos desear que en primera instancia la adopción de una herramienta de ETL, resuelva automáticamente todos los problemas aparejados de la integración de datos en una primera iteración. Los beneficios se verán concretamente en una segunda o tercera iteración de la solución de BI.

Finalmente y a manera de cierre nos gustaría decir que en sistemas de información no existen soluciones universales, es decir soluciones que siempre van a dar buenos resultados en el ciento por ciento de los casos. Por esta razón durante el presente desarrollo se abordó la temática desde este punto de vista, planteando en cada caso cuáles eran los pros y los contras de la elección de un desarrollo ETL basado en ANSI SQL, o la adopción de un software o herramienta de ETL.

---

[8] Six Key Decisions for ETL Architectures. Bob Becker. Kimbal University. http://www.informationweek.com/news/software/info_management/220600174?queryText=%22kimball+university%22

[9] Six Key Decisions for ETL Architectures. Bob Becker. Kimbal University. http://www.informationweek.com/news/software/info_management/220600174?queryText=%22kimball+university%22