1(III), 5-13, 2012

INGENIERÍA INFORMÁTICA

Propuesta de herramientas para la integración de datos

Debora Oliva Alfonso
Correo electrónico:doliva@udio.cujae.edu.cu

Thais Pineda Alfonso
Correo electrónico:tpineda@udio.cujae.edu.cu

Dalila Kindelán Castro
Correo electrónico:dkindelan@udio.cujae.edu.cu

Josue Carralero Iznaga
Correo electrónico:jcarralero@ceis.cujae.edu.cu
Complejo de Investigaciones Tecnológicas Integradas (CITI), La Habana, Cuba

Artículo Original

Resumen
Actualmente, en la mayoría de las empresas, para la realización de algunas operaciones de negocio y fundamentalmente para la toma de decisiones, se deben manipular grandes volúmenes de datos. Esta información reside en diversos repositorios de datos, de manera descentralizada, con errores y en ocasiones repetida. Debido a esto, el proceso de integración se torna complejo y poco eficiente, pues se consume mucho tiempo en realizar búsquedas directamente a la fuente de datos, requiriéndose conocer la estructura de cada una. Con mucha frecuencia el empleo efectivo de la información ha servido a las organizaciones para reducir sus costes, optimizar sus procesos, ofrecer nuevos productos y mejorar el servicio a sus clientes. Sin embargo, son muchos los obstáculos que se presentan para una gestión estratégica de la información, siendo probablemente los dos más citados la dispersión y la heterogeneidad. Esto también dificulta el proceso de integración de información, atentando contra su uso eficiente por parte de los directivos de las organizaciones. El presente trabajo está dirigido al estudio de los niveles de integración de información y técnicas para la integración de datos que permiten realizar un acercamiento al estado del arte de la integración de información. Además, se realiza un análisis sobre el empleo de herramientas de integración de datos en un entorno empresarial, llegando a propuestas concretas. El objetivo que se persigue es que las organizaciones que presentan problemas similares a los antes descritos, descubran las variantes de solución que existen y sepan utilizarlas en dependencia de sus necesidades.

Palabras clave: integración, información, datos, niveles, técnicas, herramientas

Recibido: 28 de octubre del 2011 Aprobado: 12 de diciembre del 2011

INTRODUCCIÓN

Hoy en día, para soportar todo tipo de procesos, las compañías disponen de variedad de sistemas de información desarrollados a lo largo del tiempo y que hacen uso de distintas fuentes de información. Habitualmente, estas fuentes de información presentan esquemas de consulta muy variados, están desarrolladas sobre tecnologías diferentes (bases de datos relacionales, ficheros planos, etc.) y de acuerdo con modelos de datos heterogéneos.

Para poder proporcionar información que sirva de soporte al análisis para la toma de decisiones claves sobre el negocio, es necesario integrarla bajo una estructura homogénea como paso previo a su empleo. En la actualidad existen técnicas de integración de datos, que no son más que enfoques conceptuales que definen las diversas formas de integrar la información. Estas técnicas tienen como base la construcción de modelos integrados que facilitan el acceso a las fuentes de datos, así como el trabajo de los analistas

de la información, eliminando los errores y la redundancia en los datos.

En este trabajo se introducen los niveles de integración de información, profundizando en el nivel de integración de datos a través de sus técnicas y tecnologías. Además, se define un conjunto de pautas a tener en cuenta para evaluar las herramientas y a partir de las cuales se realiza una propuesta de herramientas que se basan en la estrategia de consolidación y federación de datos. Dichas herramientas permiten, de una forma u otra, implantar una solución de integración a la medida de las necesidades de la empresa.

NIVELES DE INTEGRACIÓN DE INFORMACIÓN

La integración de información es la combinación de datos de diversos repositorios con diferentes representaciones conceptuales y contextuales; es un proceso complejo que puede ocurrir en cuatro niveles diferentes: datos, aplicaciones, procesos de negocio e interacción de usuario. En la actualidad, existe una tendencia a implantar soluciones que soporten múltiples niveles de integración y es, por consiguiente, importante diseñar una arquitectura de integración que pueda incorporar todos los niveles de integración de negocio de la empresa. En esta sección se describirá en qué consiste cada uno de los niveles de integración, se resaltan sus diferencias y los cambios que tendría en una organización asumir cada uno de ellos.

Integración de datos

El nivel de integración de datos proporciona una vista unificada de los datos que están dispersos por todas las fuentes de datos de la organización. Puede ser una vista física de datos que han sido recuperados de diferentes fuentes de información y consolidados en un repositorio, o puede ser una vista virtual que es construida dinámicamente, a la vez que se accede a los datos. Una tercera opción sería una vista de datos que han sido integrados producto de una propagación de datos de un repositorio a otro [1] [2]. El proceso de integrar a nivel de datos en las empresas es un proceso factible si se tiene en cuenta que solo requiere tener bien definidas y localizadas las fuentes de datos de las cuales se va a extraer la información que será integrada. Por otra parte, es necesario que existan especialistas en las empresas que conozcan estas fuentes de datos para poder realizar el proceso de integración con la mayor profundidad y calidad posible.

Integración de aplicaciones

El nivel de integración de aplicaciones provee una visión unificada de las aplicaciones que residen dentro o fuera de la organización. Esta visión unificada se logra a través de la gestión y coordinación del flujo de eventos (transacciones, mensajes o datos) entre aplicaciones [2]. Existen grandes diferencias entre este nivel y el anterior. El nivel de integración de datos posibilita crear un modelo de información unificado que permite proporcionar una visión a un nivel de abstracción superior. Este modelo puede ser accedido por usuarios y aplicaciones. Sin embargo, el nivel de integración de aplicaciones permite tener sincronizadas las aplicaciones a partir de eventos generados por las mismas.

Los cambios de integrar a nivel de aplicaciones en las empresas son considerablemente grandes, variando en dependencia de la complejidad de cada empresa. Las implementaciones en este nivel requieren también capacidades y experiencia que, al tratarse de un concepto nuevo, no son abundantes. Las compañías más grandes tienen por lo general equipos de especialistas en arquitectura de sistemas y desarrolladores capaces de realizar la tarea. En dependencia de la tecnología de integración que se utilice, cada organización tendrá que asumir los retos necesarios para realizar una integración a este nivel. Esto conlleva que todas las aplicaciones se preparen para integrarse, ya sea a través de mensajes o a través de servicios, lo cual se puede tornar complejo en dependencia de las arquitecturas con la cual han sido concebidas dichas aplicaciones.

Integración de procesos de negocios

El nivel de integración de procesos de negocio proporciona una vista unificada de los procesos de una organización. Las herramientas de diseño de procesos de negocios permiten a los desarrolladores analizar, modelar, y simular los procesos, así como sus actividades subyacentes. Estas herramientas, por lo general, implementan y manejan estos procesos a través de tecnologías de integración de aplicaciones. Actualmente en la mayoría de las empresas se tiene una visión del funcionamiento de la misma, orientada a los sistemas informáticos que en ella se han desarrollado; este es un enfoque erróneo porque los procesos de negocio rara vez están regidos solo por un sistema o aplicación. La ejecución de muchos procesos de negocios puede implicar la combinación de varias aplicaciones, por lo que se evidencia una estrecha relación con el nivel anterior. Al tener integradas un conjunto de aplicaciones, estas pueden responder a determinados procesos de negocios. Es importante resaltar que la integración de procesos de negocio no implica solamente tener integradas las aplicaciones involucradas, sino también, que un proceso pueda colaborar con otro, es decir, lograr que los procesos de la empresa estén relacionados.

En el ámbito de la integración de información, integrar a nivel de procesos presupone conocer todos los procesos de negocio que se llevan a cabo en la empresa. Para lograr este fin se debe contar con un equipo compuesto fundamentalmente por conocedores del negocio que realicen un levantamiento de los procesos y sus actividades subyacentes y los modelen. Por otra parte, un grupo de desarrolladores conocedores de las aplicaciones que responden a estos procesos deben llevar a cabo el proceso de integración.

Interacción de usuarios

El último nivel es la interacción de usuarios que se basa en proveer a los usuarios una única interfaz personalizada donde se presenta el contenido empresarial (procesos de

negocio, aplicaciones y datos). Una vez que se haya efectuado el proceso de integración ya sea a nivel de datos, aplicaciones, procesos de negocios o una combinación de estos niveles, se debe exponer de manera homogénea la información integrada de forma que los usuarios puedan interactuar con ella, permitiendo su análisis y comprensión, para tomar decisiones más precisas. Esta interfaz debe permitir a los usuarios colaborar y compartir información. Un portal es la variante más empleada para exponer los resultados de la integración que se ha llevado a cabo en la empresa.

Implantación de los niveles de integración en la empresa

Una empresa puede implantar una solución que soporte el nivel de integración que más se ajuste para resolver sus necesidades, pasando por alto algún nivel. No obstante, la literatura recomienda que las empresas transiten por cada uno de estos niveles teniendo en cuenta los cambios que presupone su implantación: primero integrar a nivel de datos, luego a nivel de aplicaciones, posteriormente a nivel de procesos de negocio y finalmente exponer toda la información integrada de forma tal que pueda ser compartida, permitiendo así a las empresas tener un entorno de negocios totalmente integrado.

TÉCNICAS DE INTEGRACIÓN DE DATOS

Aunque son varios los niveles que existen para la integración de información, el presente trabajo se circunscribe en el nivel de integración de datos, que como la literatura recomienda, es el más sencillo de aplicar en una empresa. Existen tres técnicas principales para la integración de datos: *Consolidación de datos*, *Federación de datos* y *Propagación de datos*, las cuales se ilustran en la figura 1. Además, existe una técnica que complementa las dos primeras, y se denomina *captura de cambios en los datos*, de la cual también se hará un breve análisis en el presente epígrafe.

Consolidación de datos

La técnica *consolidación de datos* captura la información de las diferentes fuentes de datos y la almacena en un único repositorio central. Este repositorio puede ser utilizado para la generación de reportes, para los procesos de análisis de la información y toma de decisiones y también puede actuar como una fuente de datos para diversas aplicaciones. El uso de esta técnica implica que exista un retraso, o latencia, en cuanto a la actualización de la información, pues como el proceso de obtención y almacenamiento ocurre periódicamente, puede que cierta información no esté del todo actualizada en un momento determinado. En dependencia del tiempo de actualización que se defina, este período de latencia puede durar desde algunas horas hasta varios días. [3] Las ventajas que ofrece esta técnica es que permite que grandes volúmenes de datos sean transformados, reestructurados y limpiados cuando viajan hacia el repositorio central. Sin embargo, se requieren muchos recursos de cómputo para soportar el proceso de consolidación debido a los grandes volúmenes de información a almacenar. [3] Esta técnica es conveniente utilizarla cuando se requiere integrar grandes volúmenes de información, cuando los datos a integrar no son muy cambiantes y cuando es necesario realizar complejas transformaciones a los mismos.

Existen varias tecnologías que implementan esta técnica. Una de ellas es la ETL (siglas en inglés de Extraer, Transformar y Cargar) que consiste en que la información es extraída desde un origen de datos y pasa por una secuencia de transformaciones antes de ser cargado en un repositorio central. El conjunto de fuentes de datos puede ser muy diverso y las transformaciones complejas son implementadas en entornos de desarrollo o dentro de las bases de datos. [4] La tecnología MDM (siglas en inglés de Gestión de Datos Maestros) es muy similar a la tecnología ETL en cuanto a la concepción de integración de información.

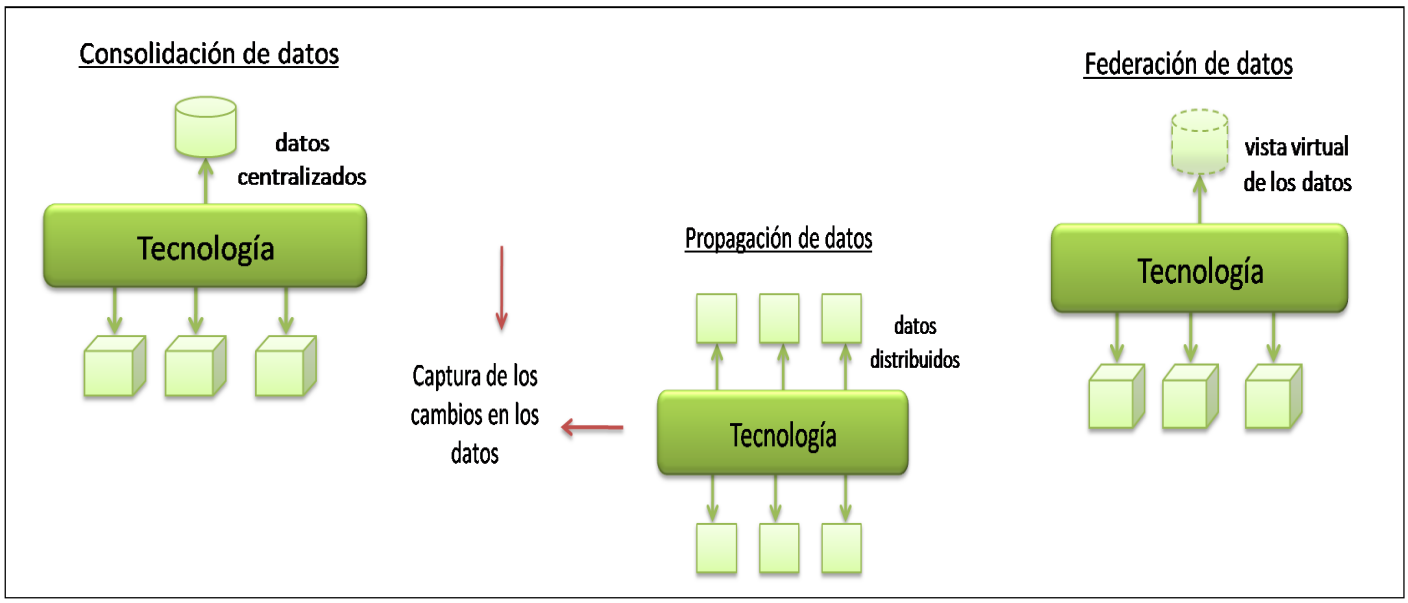

Fig. 1. Técnicas fundamentales de integración de datos.

La principal diferencia es que MDM se emplea para integrar los datos maestros de la empresa, es decir, la información básica, más valiosa sobre el negocio. Se encuentran muchas herramientas en el mercado que implementan esta tecnología y traen incorporados perfiles específicos del negocio, como productos, clientes, que facilitan el proceso de integración. [5] La tecnología ECM (siglas en inglés de Gestión de Contenido Empresarial) permite integrar información no estructurada basándose en la técnica de consolidación. Los productos ECM se concentran en el intercambio y la gestión de grandes cantidades de datos no estructurados como documentos, información de la Web y multimedia, relacionados con los procesos de la organización. [6]

Federación de datos

La técnica *Federación de datos* provee una vista virtual única de la información recuperada de diferentes fuentes de datos. Uno de los elementos claves de un sistema de federación son los metadatos que son usados por el motor de federación para acceder a los datos, pues estos contienen información acerca de la localización real de los mismos. Esta información adicional puede ayudar a la solución federada a optimizar el acceso a los repositorios de datos. [7] Es importante resaltar que el acceso a los datos se realiza en tiempo real y que los mismos se mantienen en su lugar de origen; esta es su principal diferencia con la técnica de consolidación. Cuando una aplicación solicita información, el motor de federación obtiene los datos directamente de las fuentes, los une en una vista virtual y los resultados de esta unión son devueltos a la aplicación. Las ventajas que brinda esta técnica es que siempre la información se encuentra actualizada, debido a que la integración se realiza en tiempo real. Presenta como inconveniente que no permite recuperar grandes volúmenes de datos y realizar grandes transformaciones a los mismos. Es conveniente utilizarla cuando se necesiten pequeños volúmenes de información para pequeñas consultas y cuando los datos a integrar sean muy cambiantes. [3]

La tecnología más comúnmente usada en el mundo para implementar la federación de datos es EII (siglas en inglés de Integración de Información Empresarial) la cual consiste en construir vistas virtuales de la información que se encuentra dispersa en diferentes fuentes de datos. La tecnología EII facilita establecer una capa intermedia de servicios de datos, que lanzan consultas a los diferentes sistemas de información [8]. Los servicios web son una tecnología que no surgió explícitamente para la integración de datos, pero por su característica de ser interoperables, son ampliamente utilizados para federar datos. De hecho, muchos productos EII exponen la integración de sus datos como servicios web accesibles para terceras aplicaciones que necesiten contar con datos integrados en tiempo real.

Propagación de datos

La técnica *Propagación de datos* consiste en la distribución de datos desde una fuente de información hacia otra, lo que posibilita que la información de ambas fuentes se encuentre siempre sincronizada. Las fuentes de datos deben ser constantemente actualizadas y este proceso consiste en mover grandes volúmenes de datos de un sistema a otro. Es común, para el movimiento de estos grandes volúmenes de datos, que sea realizado en lotes dentro de un período breve de tiempo. Cuanto mayor sea el volumen de datos a mover, más difícil y complejo se volverá este proceso. Es necesario entonces, encontrar la manera de moverlos con mayor rapidez y de identificar y mover solo los datos que han sido modificados desde la última actualización. Esta técnica se relaciona en algunas ocasiones con la técnica *Captura de cambios* en los datos, para capturar y propagar solo los cambios ocurridos. La gran ventaja que presenta esta técnica es que puede ser utilizada en tiempo real o en un tiempo cerca de lo real. [3]

EDR (siglas en inglés de Replicación de Datos Empresariales) es una tecnología que se basa en la técnica de propagación de datos y frecuentemente emplea la técnica de *Captura de cambios en los datos* para replicar solo la información que ha cambiado en el origen de datos.

Captura de cambios en los datos

Es una estrategia de integración de datos para determinar los datos que han sido cambiados, la cual se basa en la identificación, captura y entrega de los cambios (nuevas inserciones, eliminaciones, modificaciones) realizados a las fuentes de datos. Consiste en una variedad de métodos que optimizan la transferencia de datos al mover solo los que han cambiado desde la última transferencia [9]. Sin la utilización de esta técnica, todos los datos deben ser accedidos en lotes y movidos a la fuente que los necesite. Con esta técnica solo un pequeño número de cambios es procesado y enviado a la fuente destino. [10]

Existen varios métodos que son utilizados por esta técnica para localizar y capturar los cambios ocurridos en los datos. El método *Time-stamp* (marca de tiempo) muestra la fecha del último cambio en cualquier fila de cualquier tabla que tenga una marca de tiempo asociada; la columna que tenga la fecha de cambio más reciente, fue la que cambió. La enumeración de versiones es otro método muy utilizado donde se considera que han cambiado todos los datos con números de versiones más recientes. También se emplean *triggers* para almacenar una copia de los datos modificados e indicadores de estado. La mayoría de los sistemas de gestión de base de datos administran un registro de transacciones que registra todo cambio realizado al contenido y los metadatos de la base de datos. [3] Las marcas de tiempo y los números de versiones son comúnmente utilizados en las aplicaciones de datos no estructurados. Cuando un documento es creado o modificado, los metadatos del documento son actualizados para reflejar la fecha y hora de la actualización. Muchos sistemas de datos no estructurados también crean una nueva versión del documento cada vez que se modifica.

Enfoque híbrido para la integración de datos

Existe una tendencia a implantar un enfoque híbrido que incorpore la combinación de varias técnicas. Este enfoque es muy apropiado ya que se pudiera tener la información poco variable en el tiempo en un repositorio físico de datos utilizando la técnica *Consolidación de datos,* mientras que por otra parte se tendría la información, que cambia con frecuencia, de manera virtual mediante la técnica de *Federación de datos*. Igualmente se puede propagar la información, cuando se efectúe un cambio en algún sistema usando la técnica de *Propagación de datos*. [3]

HERRAMIENTAS PARA LA INTEGRACIÓN DE DATOS

Lo más recomendable para cualquier organización es comenzar por el nivel de integración de datos, pues solo es necesario ubicar los repositorios de datos y construir vistas de la información que puedan ser accedidas posteriormente desde aplicaciones externas. Por esta razón la presente investigación aborda herramientas basadas en las técnicas de federación y consolidación de datos. Las herramientas de consolidación evaluadas tienen implementadas además las técnicas de captura de los cambios en los datos y la propagación. No se tuvo en cuenta la tecnología ECM porque es necesario tener amplios conocimientos de cómo explotar la información textual antes de comenzar a evaluar herramientas de este tipo. Por otra parte tampoco se tuvieron en cuenta herramientas MDM puesto que son una variante específica de la técnica de consolidación para gestionar los datos maestros de una empresa.

Arquitectura genérica de las herramientas de integración de datos

Las herramientas para la integración de datos, tanto de consolidación como de federación, se caracterizan por brindar conectores o adaptadores para un amplio conjunto de fuentes de datos estructuradas, desde diferentes tipos de bases de datos hasta servicios web. Cuando se desea desarrollar un modelo de integración, el primer paso es identificar las fuentes de datos que van a brindar la información que necesita, obteniendo de esta forma el modelo físico de los datos. Luego, se aplican reglas de transformación a los datos que permiten lograr un modelo lógico de datos que exprese las necesidades del usuario. Este modelo una vez confeccionado puede ser accedido desde aplicaciones externas, lo cual facilita la toma de decisiones a los clientes, al contar con la información integrada. La figura 2 muestra una vista de la arquitectura de estas herramientas basándose en lo explicado anteriormente.

Pautas para evaluar las herramientas de integración de datos

Con el incremento de la demanda de soluciones de integración por parte de las empresas, los requisitos hacia las herramientas que ofrecen este tipo de soluciones han aumentado. Las compañías que proveen estas herramientas han pasado de brindar un conjunto de productos, a ofrecer plataformas de integración [11]. Ante esta situación se hace necesario contar con un conjunto de pautas que ayuden a decidir cuál es la solución que más se adecua a las necesidades particulares de cada entorno. A continuación se exponen un conjunto de pautas que los autores consideran deben tomarse en cuenta siempre.

Interfaz amigable e intuitiva: Es importante tener en cuenta este aspecto pues los usuarios de las herramientas no son necesariamente informáticos. Es conveniente que brinde la facilidad de mostrar la información contenida en las fuentes de datos, por ejemplo, las tablas de las bases de datos conectadas y que haga uso de asistentes que ayuden a los usuarios en la construcción de los modelos de manera sencilla.

Soporte para múltiples fuentes de datos: Todas las herramientas existentes tienen un conjunto de conectores o adaptadores para diversas fuentes de datos, pero es limitado; estas herramientas van a ser utilizadas en los más disímiles entornos.

Extensibilidad en cuanto a fuentes de datos: Es beneficioso que las herramientas tengan la posibilidad de incluir nuevas fuentes de datos que no estén contenidas en el producto, mediante mecanismos de programación.

Transformaciones a los datos, extensibilidad en cuanto a reglas: Deben brindar reglas de transformación de los datos, desde las más sencillas hasta las más complejas. Además, deben permitir desarrollar reglas personalizadas y extender las ya empaquetadas.

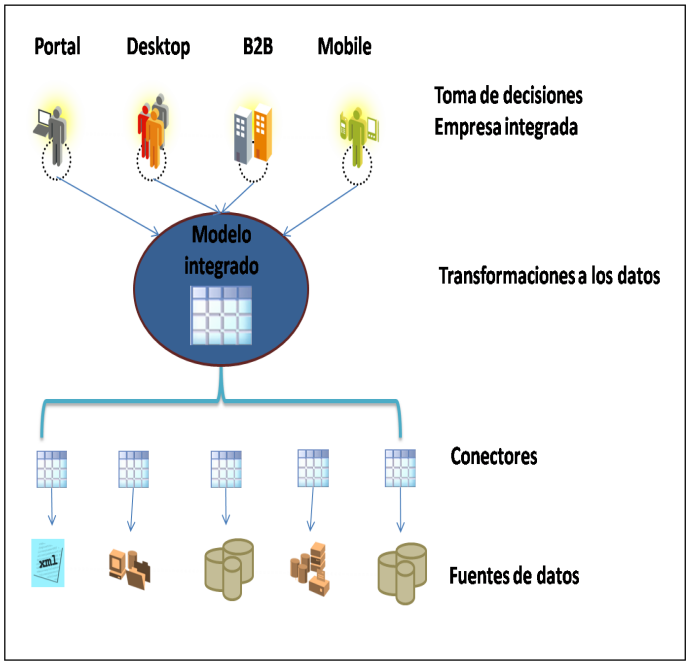

Fig. 2. Arquitectura de las herramientas de integración de datos.

Como la información que se encuentra en las fuentes de datos en ocasiones posee errores, es necesario que las herramientas permitan realizar procesos de limpieza a los datos garantizando cierto nivel de calidad en la información integrada.

Acceso a la información: Las herramientas van a constituir una plataforma base para muchos sistemas, debido a las potencialidades en cuanto a integración que van a poseer, por lo que uno de los requisitos más importantes que deben cumplir es ser interoperables. Para lograr esto, el acceso a la información modelada debe ser mediante el uso de estándares. En tal sentido, como cada día gana más aceptación el concepto de servicio de datos, las herramientas de integración igualmente deben exponer características orientadas a servicios.

Técnicas de optimización de comsulta, rendimiento: No se tiene conocimiento de cuántos usuarios van a estar realizando peticiones concurrentemente sobre un mismo modelo. Por tal motivo el rendimiento es un aspecto clave a analizar. En el caso de las herramientas de federación de datos el rendimiento se ve más afectado pues cada vez que se realiza una petición hay que buscar la información directamente en las fuentes de datos y realizar la integración *on the fly*. Es por esta razón que las herramientas han incorporado técnicas de optimización para mejorar sus tiempos de respuesta. Mientras más mecanismos de optimización incorporen las herramientas de federación, más probabilidades hay de que tengan buen rendimiento.

Variantes para sincronizar la información: En el caso de las herramientas de consolidación de información es importante tener en cuenta las variantes para sincronizar la información que estas proveen. Estas variantes deben ser lo menos intrusivas posible en las fuentes de datos y deben ser diversas. Hay que tener en cuenta que generalmente los que integran información no son dueños de las fuentes de datos que intervienen en el proceso, por lo que solo pueden realizar operaciones de lectura sobre sus datos.

Herramientas para la consolidación de los datos

Después de realizar un estudio de la gama de herramientas de integración de datos disponibles que siguen la estrategia de consolidación de datos, se seleccionaron las herramientas Novell Identity Manager y Microsoft Forefront Identity Manager 2010 para pasar a la fase de evaluación. Otras como Sun One Metadirectory Server, Informatica Power Center, SAP Data Integrator y Critical Path XML no se pudieron evaluar pues las compañías dueñas de las herramientas no ofrecían ninguna versión de prueba de su producto.

En cuanto a la conexión a las fuentes de datos Novell y Forefront no tienen muchas diferencias. Su principal diferencia radica en la definición de reglas para transformar los datos, el acceso a la información desde aplicaciones externas y las variantes de sincronización que ofrecen.

Novell contiene una serie de políticas con grandes potencialidades y de fácil uso para la transformación de los datos; todas las reglas y eventos son manipuladas mediante documentos en formato XML, pero la interfaz de usuario es poco intuitiva. Forefront ofrece un asistente que guía al desarrollador en todos los pasos a realizar, aunque no tiene la variedad de reglas de Novell pero estas se pueden personalizar de manera ágil a través de una API que brinda la herramienta.

El acceso a la información integrada es muy superior en Forefront, pues brinda unas API en .NET que permiten acceder a servicios web, mientras que con Novell hay que tener un amplio conocimiento sobre LDAP para saber cómo comunicarse con el modelo a través de las API que proporciona.

Las variantes de sincronización son un aspecto determinante en las herramientas de consolidación de datos. Novell obliga al usuario a crear una tabla en la fuente de datos a la que se va a tener acceso y debe crear *triggers* para llenarlas; esto es imposible de acometer cuando el usuario no es el dueño de la fuente de datos o no tiene permisos para realizar cambios en la misma. Forefront además de esta variante intrusiva permite sincronizar cada cierto tiempo toda la información modelada.

A manera de resumen, en la tabla 1 se muestra una comparación entre estas dos herramientas teniendo en cuenta los aspectos definidos anteriormente.

Tabla 1
Comparación de herramientas de consolidación de datos

Parámetros	Novell Identity Manager	Forefont Identity Manager
Interfaz amigable e intuitiva	No	Sí
Soporte para múltiples fuentes de datos	Sí	Sí
Extensibilidad en cuanto a fuentes de datos	No	Sí
Transformaciones a los datos	Muchas	Pocas
Extensibilidad en cuanto a reglas	No	Sí
Acceso a la información	No estandarizado	Estandarizado
Variantes para sincronizar la información	Intrusivas	Intrusivas y no intrusivas

Como se puede observar Forefront está mucho más acorde que Novell a los parámetros definidos para la evaluación, por lo que se propone Forefront Identity Manager como herramienta para la consolidación de datos.

Herramientas para la federación de datos

Después de realizar un estudio de las herramientas de integración de datos disponibles que siguen la estrategia de federación, se seleccionaron Symlabs VDS (siglas en inglés

de Servidor de Directorio Virtual), RadiantOne VDS y Oracle Data Service Integrator (ODSI) para pasar a la fase de evaluación. Otras como SAP Data Federator, Composite Software Information Server o iWay's Enterprise Information Management Suite no se pudieron evaluar pues las compañías dueñas de las herramientas no ofrecían ninguna versión de prueba de su producto.

Aunque las herramientas seleccionadas cumplen en alguna medida con los aspectos definidos para evaluarlas, hay ciertos puntos que marcan la diferencia entre unas y otras. Symlabs VDS contiene una serie de *plugins* que permiten el acceso y el manejo de la información, pero el proceso de integración de los datos se torna lento y complejo pues todas las configuraciones hay que realizarlas de manera manual con una interfaz poco amigable. Además, es necesario tener profundo conocimiento sobre el protocolo LDAP (esquemas, atributos, entradas) tanto para el desarrollo del modelo como para poder accederlo desde una aplicación externa, lo cual complejiza el proceso de despliegue de la solución y el tiempo de aprendizaje por parte de los usuarios. Aunque la herramienta permite desarrollar reglas y conectores personalizados a las fuentes de datos, lo hace a través de un lenguaje desarrollado por la compañía llamado Directory Script con una sintaxis difícil de asimilar. Esta situación se ve agravada pues Symlabs VDS trae pocas reglas para transformar los datos, y por tanto casi todo el modelo hay que implementarlo con Directory Script. Esto último trae como consecuencia que el tiempo de desarrollo por parte de los usuarios sea muy elevado, lo cual es un inconveniente teniendo en cuenta la premura con que se necesita tener los datos integrados. La estrategia que utiliza Symlabs VDS para el rendimiento es mediante la memoria cache; para esto es necesario configurar varios plugins que retardan igualmente el tiempo de desarrollo de los modelos de integración.

Por su parte RadiantOne VDS brinda un amplio conjunto de conectores a las fuentes de datos, tiene una interfaz muy intuitiva que facilita el proceso de integración de datos pues visualiza la información referente a los esquemas, tablas y columnas de las bases de datos sin necesidad de realizar procesos tediosos de configuraciones. En cuanto a las reglas, la herramienta no brinda grandes posibilidades de transformación a los datos, pero tiene un entorno de desarrollo en java (lenguaje muy conocido y empleado por los programadores) que facilita la personalización de las reglas y el desarrollo de conectores a las fuentes de datos que la herramienta no incluya.

La única vía que proporciona RadiantOne VDS para que las aplicaciones clientes se conecten al modelo es mediante LDAP. Aunque LDAP es un estándar, todos los entornos de desarrollo no incluyen API para comunicarse con servidores de este tipo. Además serían aplicaciones que habría que perfeccionar constantemente en la medida que el modelo va creciendo en cuanto a fuentes de datos a integrar. Una idea válida para solventar esta limitación sería desarrollar una capa de servicios web dinámica para un servidor LDAP. De esta manera las aplicaciones clientes consumirían los datos del modelo integrado a través de la capa de servicios, abstrayéndolos de la estructura de protocolo LDAP. Esto serviría de igual manera para Symlabs VDS.

La estrategia que utiliza RadiantOne VDS para el rendimiento es mediante la memoria cache; esta puede ser configurada para cualquier rama en el directorio virtual. Este modelo de cache presenta dos tipos de memoria: principal y virtual. [12] La memoria principal es la memoria real donde un cierto número de la mayoría de entradas recientemente usadas permanecerán. La memoria virtual es la memoria en disco y es donde residirán todas las entradas que excedan la cantidad permitida en la memoria principal. El intercambio de entradas desde la memoria virtual hacia la principal (y viceversa) es gestionado por el servidor virtual. [13]

ODSI es una solución completamente basada en estándares y declarativa, que permite rehusar los servicios de datos. Tiene los más variados conectores a fuentes de datos y posibilita el desarrollo de adaptadores personalizados. Es importante resaltar que esta herramienta convierte todas las fuentes de datos conectadas en esquemas XML abstrayéndose de la especificidad de cada repositorio de datos, lo que facilita el uso de XQuery como lenguaje de consulta a ficheros XML. El mapeo y las transformaciones son diseñadas en un componente fácil de usar con más de 200 funciones XQuery; además, brinda un editor de código fuente para el código XQuery donde se puede incorporar alguna transformación compleja no contemplada en las funciones. Al exponer los datos como servicios web de manera semiautomática, la información es accesible a una amplia variedad de tipos de clientes.

ODSI logra un rendimiento óptimo para las consultas mediante la realización de SQL pushdown. Pushdown es una técnica de optimización que alivia la carga de procesamiento del motor de XQuery mediante el envío de consultas nativas SQL hacia las fuentes de datos. [14] Mediante el almacenamiento en la cache se pueden mejorar los tiempos de respuesta para los clientes y reducir la carga de procesamiento en las fuentes de datos subyacentes. [15]

La tabla 2 muestra, a manera de resumen, los aspectos claves evaluados en cada herramienta de federación de datos. El rendimiento no se pudo evaluar en ninguna de las tres herramientas porque no se tuvo un entorno real donde se pudiera probar la concurrencia y la capacidad de respuesta de cada una al integrar un conjunto determinado de datos, por eso no se incluyó en la tabla comparativa, sin embargo, se reflejó como elemento importante a tener en cuenta para evaluar cualquier herramienta de federación de datos.

A manera de conclusión parcial es evidente la superioridad de ODSI respecto a RadiantOne VDS y Symlabs VDS. Aunque ODSI tiene una arquitectura muy robusta se propone que sea evaluado en un entorno de prueba para validar las técnicas de optimización que ofrece y medir aspectos claves como el rendimiento en una estrategia de federación de datos.

De los resultados de la evaluación, es posible llegar a la conclusión de que las herramientas Symlabs VDS y RadiantOne VDS resuelven parcialmente el problema de la integración de datos, pero carecen de un grado aceptable de estandarización, lo que entorpece en gran medida la compatibilidad con otros sistemas y la reutilización de los modelos integrados. La escasa implementación de técnicas de optimización en las citadas herramientas conlleva a que existan problemas de latencia y poca fiabilidad en el acceso a los datos integrados, dado que el tiempo de respuesta se vería afectado y los usuarios y/o aplicaciones tendrían que esperar espacios de tiempo prolongados, comprometiendo de esta manera la gestión de los mismos.

que integra datos mediante la técnica de consolidación con un grado de calidad elevado.

En general con la correcta elección de la técnica, tecnología y herramienta para la integración de datos en su empresa, se logrará que la información sea confiable y exacta. Existirá mayor rapidez en la toma de decisiones de la organización y un aumento de la capacidad de respuesta operativa de la entidad. Se reducirá el tiempo de búsqueda, acceso y uso de la información global de la empresa, dado que habrá un único punto de acceso a los datos, permitiendo una visión unificada, homogénea y en un único formato.

Tabla 2
Comparación de herramientas de federación de datos

Parámetros	Symlabs VDS	RadiantOne VDS	ODSI
Interfaz amigable e intuitiva	No	Sí	Sí
Soporte para múltiples fuentes de datos	Sí	Sí	Sí
Extensibilidad en cuanto a fuentes de datos	No	Sí	Sí
Transformaciones a los datos	Pocas	Pocas	Muchas
Extensibilidad en cuanto a reglas	No	Sí	Sí
Acceso a la información	No Estandarizado	No Estandarizado	Estandarizado
Técnicas de optimización de consultas	Pocas	Pocas	Variadas

CONCLUSIONES

La tarea fundamental para la implantación de una herramienta de integración de datos es definir en qué entorno y estado se encuentran los mismos, para realizar la elección de la técnica y tecnología que más se ajusta a la situación empresarial en cuestión. Si es necesario consultar datos en tiempo real y estos no son de gran volumen, se debe optar por la técnica de federación, y por consecuente la tecnología EII, y se propone la utilización de la herramienta ODSI, que posee grandes potencialidades en la implementación de la técnica y tecnología mencionadas. Por otra parte, si se desean integrar grandes volúmenes de datos y estos no van a ser muy cambiantes en el tiempo, la consolidación es la opción correcta. En cuanto a las tecnologías, si lo que se quiere realizar son transformaciones a los datos, para que luego sean almacenados en un repositorio determinado, la tecnología a utilizar es ETL, si lo que se necesita es integrar los datos más importantes de la empresa para utilizarlos en los procesos claves del negocio, se recomienda implementar la MDM, y si lo que busca es gestionar datos no estructurados como imágenes, correos o documentos, entonces la más adecuada es ECM. En este caso se sugiere la utilización de la herramienta Forefont Identity Manager

REFERENCIAS

1. *Data Integration: Using ETL, EAI, and EII Tools to Create an Integrated Enterprise*. Colin White, B.R. 2005.
2. EAI: *Integración de Aplicaciones de Empresa de algunos conceptos básicos* [en línea]. IDG COMMUNICATIONS, S.A.U. Computer World [ref. de 2006]. Disponible en Web: http://www.idg.es/computerworld/EAI:-Integracion-de-Aplicaciones-de-Empresa.Alguno/seccion-tec/articulo-110938.
3. *A Roadmap to Enterprise*. [en línea] White, Colin [ref. de 2006]. Disponible en Web: ftp://service.boulder.ibm.com/ps/software/emea/de/db2/RoadmapToEDI_WP_021306.pdf
4. *ETL(Extracción Transformación y Carga)* [en línea]. Technologies, O.D. [ref. de 2006]. Disponible en Web: http://www.ondtech.com.
5. *Enterprise Master Data Management an SOA Approach to Managing Core Information* [en línea]. Allen Dreibelbis, E.H., Ivan Milman, Martin Oberhofer, Paul van Run and Dan Wolfson [ref. de 2008]. Disponible en Web: http://cdn.ttgtmedia.com/searchDataManagement/downloads/MasterDataManagementSOA1.pdf

6. *What is ECM? AIIM* [en línea]. Association, A.-T.E.C.M. [ref de 2005]. Disponible en Web: http://www.aiim.org/What-is-ECM-Enterprise-Content-Management.aspx.

7. *What is Data Federation Technology* [en línea]. TechTarget [ref. de 2009]. Disponible en Web: http://searchdatamanagement.techtarget.com/sDefinition/0,,sid91_gci1376262,00.html.

8. *Enterprise Information Integration: Successes, Challenges and Controversies* [en línea]. Halevy, A.Y., Naveen Ashish, Dina Bitton, Michael Carey [ref de 2005]. Disponible en Web: http://www.cs.washington.edu/homes/alon/files/eiisigmod05.pdf

9. PAWEL PLASZCZAK, T.M. *Real-time Data Integration Us*ing Change Data Capture [en línea]. BigDataMatters [ref. de 2009]. Disponible en Web: http://bigdatamatters.com/bigdatamatters/2009/08/real-time-data-integration-using-change-data-capture.html.

10. *Change Data Capture: Driving Results with Event Driven Data* [en línea]. Informática . [ref de 2005]. Disponible en Web: http://www.information-age.com/article_assets/articledir_1962/981362/6816_wp_PECDC_web.pdf

11. *Unified Data Management: A Collaboration of Data Disciplines and Business Strategies*. P., R. 2010.

12. *RadiantOne VDS Context Edition Architect's Guide* [en línea]. Radiant Logic, I. [ref de 2009]. Disponible en Web: http://www.radiantlogic.com/products/radiantone-vds-context-edition/

13. *RadiantOne VDS Context Edition Performance and Scalability* [en línea]. Radiant Logic, I.l [ref de 2009] . Disponible en Web: http://www.radiantlogic.com/main/products_vcs_performance.html.

14. *XQuery and XQSE Developer's Guide. XQuery Engine and SQL* [en línea]. Oracle. [ref de 2008]. Disponible en Web: http://docs.oracle.com/cd/E13167_01/aldsp/docs30/pdf/xquery.pdf

15. *Administration Guide* [en línea]. Oracle [ref de 2008]. Disponible en Web: http://download.oracle.com/docs/cd/E13190_01/liquiddata/docs85/pdf/admin.pdf.

AUTORES

Debora Oliva Alfonso
Ingeniera Informática, Máster en Informática, Instructora, Complejo de Investigaciones Tecnológicas Integradas (CITI), La Habana, Cuba

Thais Pineda Alfonso
Ingeniera Informática, CITI, La Habana, Cuba

Dalila Kindelán Castro
Ingeniera Informática, CITI, La Habana, Cuba

Josue Carralero Iznaga
Ingeniero Informático, Máster en Informática, Asistente, CITI, La Habana, Cuba

Proposed Tools for Data Integration

Abstract

Currently, in most of the companies, for performing certain business operations and primarily to make decisions, large volumes of data must be handled. This information is located in various data repositories with the result that it is decentralized with errors and often repeated in different sources. Because of this, the integration process becomes complex and it takes time to search data source directly, requiring to know the structure of each one. Very often the effective use of information has helped organizations to reduce costs, optimize processes, offer new products and improve service to its customers. However, many obstacles arise for strategic management of information, being probably the two most cited the dispersion and heterogeneity. This also makes it difficult the information integration process in a way that it can be accessed by external applications, which hampers its efficient use by organizations' managers. This investigation aims to study the information integration levels and data integration techniques to perform an approach to the state of the art of the information integration. This research also does an analysis on the use of data integration tools in an enterprise environment, in order to arrive at concrete proposals. The objective pursued is that organizations that have similar problems to those described above, can discover the existing solutions and know how to use them depending on their needs.

Key words: integration, data, information, levels, techniques, tools