TDWI RESEARCH

LISTA DE COMPROBACIÓN ELABORADA POR **TDWI**

Modernización de la integración de datos

para dar cabida a los nuevos requisitos de negocio y Big Data

Autor: Philip Russom

Patrocinado por

informatica sas

tdwi.org

tdwi
Advancing all things data.

DICIEMBRE DE 2015

LISTA DE COMPROBACIÓN ELABORADA POR TDWI

Modernización de la integración de datos

para dar cabida a los nuevos requisitos de negocio y Big Data

Autor: Philip Russom

tdwi
Advancing all things data.

555 S Renton Village Place, Ste. 700
Renton, WA 98057-3295 (EE. UU.)

Teléfono 425.277.9126
Fax 425.687.2842
Correo electrónico info@tdwi.org

tdwi.org

CONTENIDOS

PRÓLOGO

Todos cuantos nos dedicamos a la gestión de datos estamos atravesando un prolongado período de grandes cambios con la introducción del Big Data, otras categorías de datos nuevos y las nuevas plataformas de gestión de datos en nuestras organizaciones. En respuesta, la mayoría de las organizaciones de usuarios lucha con denuedo para aprender a usar las nuevas tecnologías y, aún más importante, a aprovechar las plataformas y los datos nuevos en beneficio del negocio. Como resultado, muchos profesionales de los datos se enfrentan ya tanto a los requisitos nuevos como a los futuros requisitos que surgirán a medida que aparezcan online nuevas fuentes de datos.

Los cambios que ya están en marcha llevan a numerosas organizaciones técnicas a replantearse su infraestructura, su equipo y sus conocimientos de gestión de datos con el fin de modernizarlos. De esas iniciativas, la *modernización de la integración de datos* es una de las más urgentes, puesto que desempeña un dilatado papel en la captura, el procesamiento y la transferencia de datos, tanto antiguos como nuevos. Sin las modernas soluciones de integración de datos, las organizaciones no pueden satisfacer ni los requisitos actuales ni los futuros de Big Data, análisis ni funcionamiento en tiempo real.

La modernización de la integración de datos adopta muchas formas según la situación de la infraestructura y las clases de plataformas y datos nuevos que se deban adoptar. En lugar de relacionarlas todas, vamos a ofrecer sendas recomendaciones para las siete tareas de modernización de la integración de datos más acuciantes. Estas recomendaciones sirven como guía a la hora de elegir los productos de los proveedores y actualizar el diseño de las soluciones para la modernización:

1. Complemente la elevada latencia de las antiguas prácticas de integración de datos con una gama más amplia de técnicas de incorporación de datos
2. Adopte las nuevas prácticas y herramientas de preparación de datos para ganar en agilidad, rapidez, sencillez y facilidad de uso
3. Integre los datos con métodos que franqueen a más usuarios el acceso de autoservicio a los datos nuevos y al Big Data
4. Modernice la infraestructura de integración de datos aprovechando los nuevos tipos de plataformas de datos
5. Continúe añadiendo más funciones oportunas a medida que modernice las soluciones de integración de datos
6. Modernice la funcionalidad de la integración de datos para obtener, por fin, valor analítico y de negocio de los datos multiestructurados y no tradicionales
7. Plantéese la modernización de su oferta de herramientas de integración de datos con una plataforma integrada de varias herramientas de gestión de datos

En este informe, se profundiza en cada una de esas siete recomendaciones y se explican muchos de los nuevos tipos de productos de los proveedores, su funcionalidad y las mejores prácticas del usuario que facilitan la modernización de la integración de datos. Asimismo, se presentan tanto el business case como los puntos fuertes tecnológicos aplicables a cada recomendación.

NÚMERO UNO

COMPLEMENTE LA ELEVADA LATENCIA DE LAS ANTIGUAS PRÁCTICAS DE INTEGRACIÓN DE DATOS CON UNA GAMA MÁS AMPLIA DE TÉCNICAS DE INCORPORACIÓN DE DATOS

Uno de los cambios más drásticos en la práctica de la integración de datos en los últimos años ha sido la modernización de la incorporación de datos. La incorporación consiste simplemente en cómo, dónde y con qué frecuencia se cargan o descargan los datos que penetran en un entorno en los destinos (por ejemplo, "staging areas" de datos, data warehouses o sistemas de archivos). Durante décadas, los procesos de incorporación de las prácticas de integración de datos de tipo ETL han sido "latentes" (esto es, lentos y ejecutados, a menudo, por la noche). Los datos del estudio de TDWI demuestran que la mayor parte de los datos de la mayoría de los data warehouses se actualiza en ciclos de 24 horas. Sin embargo, el porcentaje de datos que se recopila, prepara y entrega para la presentación con frecuencia o inmediatez no para de aumentar por distintos motivos:

Algunas fuentes nuevas generan datos con frecuencia. Una categoría importante del Big Data son los datos automáticos. Hay sensores integrados (o, si no, se añaden) en una creciente lista de máquinas, incluidos vehículos, dispositivos móviles y robots de líneas de fabricación. Aparte de los de las máquinas, proliferan los sensores de GPS en los palés de transporte y en otros bienes móviles. Algunos sensores generan y emiten datos en transmisiones continuas de eventos, mientras que otros solo emiten cuando se hace ping (como ocurre con los chips de RFID) o cuando la máquina donde se encuentran efectúa alguna acción (por ejemplo, cuando un robot instala un widget). El caso es que muchas organizaciones desean aprovechar las nuevas fuentes de transmisión con intención de mejorar la logística, la supervisión de opiniones, los acuerdos de nivel de servicio, el cumplimiento de cuotas, la vigilancia de las instalaciones, el análisis operacional y la supervisión de actividades de empresa.

Siguen aumentando las prácticas de negocio que exigen datos muy recientes. Los usuarios llevan tiempo empleando un business intelligence operacional, el cual actualiza con frecuencia los cuadros de mando de gestión y otros informes operacionales con datos cuya novedad varía entre minutos y horas. Al igual que el business intelligence operacional, las prácticas en los ámbitos de la gestión del rendimiento, la elaboración de informes, OLAP y el análisis avanzado exigen datos cada vez más actualizados. Los datos recientes aumentan la ventaja de los negocios frente a la competencia, mejoran las relaciones con los clientes, refuerzan la excelencia operativa y facilitan la toma ágil, aunque documentada, de decisiones tácticas.

Las prácticas de incorporación de datos deben dar cabida a datos de rapidez y frecuencia variables. No olvide que, pese a todo, los procedimientos de ETL y ELT latentes siguen siendo imprescindibles tanto para mantener la precisión como para las uniones de extremos, las transformaciones y la pista de auditoría propias de los datos destinados a data warehousing, la mayoría de los informes estándar y numerosas soluciones de OLAP. El reto consiste en diseñar nuevas soluciones de integración de datos (o bien en ajustar las existentes) que capturen e incorporen los datos nuevos con mayor frecuencia y rapidez. Es lo que, en ocasiones, se denomina incorporación temprana o continua, mucho más rápida y frecuente en comparación con las cargas en batch durante la noche. Como contrapartida, la incorporación temprana apenas realiza transformación o agregación alguna de los datos (si es que la hace) antes de la carga, ya que ralentizaría la incorporación. Un beneficio es que los datos se capturan en su estado original y, por consiguiente, se pueden reutilizar con fines distintos a medida que surjan otros requisitos de análisis o elaboración de informes. El mayor beneficio es que los datos se encuentran a disposición de los informes, los análisis y las operaciones lo antes posible.

Otro pequeño reto en este sentido es que la reutilización de los datos se produce, cada vez más, sobre la marcha en tiempo de ejecución, en lugar de antes de la carga, por ejemplo, mientras el analista de datos o el especialista en datos explora los datos y desarrolla conjuntos de datos nuevos para el análisis, cuando una rutina de integración de datos analiza los datos cambiados (recién incorporados) a fin de actualizar los análisis o los informes operacionales del día corriente, o cuando el director de ventas actualiza un cubo para comprobar las ventas diarias hasta ese momento.

Puesto que el hardware y el software modernos de hoy en día son rápidos y escalables, ya sí es factible aprovechar el rendimiento y el procesamiento en tiempo de ejecución de la incorporación continua. Además, la funcionalidad de las herramientas para el procesamiento de datos sujetos a plazos ha alcanzado una gran madurez, tal y como se observa en algunas funciones explicadas en el presente informe, por ejemplo, el procesamiento de transmisiones de eventos, la federación y la preparación de datos o el acceso de autoservicio a los datos.

Los datos incorporados con rapidez también se pueden procesar con los métodos tradicionales. La incorporación continua pone los datos nuevos a disposición de las tecnologías y los usuarios que los necesitan sin dilación, en tanto que la subsiguiente aplicación práctica de los datos capturados recurre a las mejores prácticas en calidad, modelado y agregación de datos establecidas. Por ejemplo, el procesamiento online de transmisiones de datos provenientes de robots de fabricación puede revelar la existencia de lotes dañados u otros problemas relevantes que se deben atender de inmediato; al estudiar offline esos mismos datos, también se ponen de manifiesto valiosas tendencias a largo plazo en la relación entre el rendimiento de los proveedores y la calidad de los productos.

NÚMERO DOS

ADOPTE LAS NUEVAS PRÁCTICAS Y HERRAMIENTAS DE PREPARACIÓN DE DATOS PARA GANAR EN AGILIDAD, RAPIDEZ, SENCILLEZ Y FACILIDAD DE USO

Tal y como decíamos, el procesamiento de datos sobre la marcha se ha manifestado como una práctica de integración de datos diferenciada que adopta nombres diversos, como manipulación o combinación de datos. Hay quien califica esta integración de datos de "descafeinada" porque, por lo general, sus implementaciones incluyen un subconjunto reducido de funciones de integración, recortadas por motivos de utilidad y rendimiento. En TDWI, en cambio, se le suele asignar el nombre de *preparación de datos*. Existen distintos tipos de herramientas que facilitan alguna forma de preparación de datos, por ejemplo, las de integración de datos, perfilado de datos, calidad de datos, exploración de datos, análisis y visualización de datos.

En concreto, la preparación de datos ya se efectúa de manera habitual para muchas clases de análisis. Permite a los analistas de datos, los especialistas en datos o usuarios parecidos trabajar con datos brutos muy detallados, sin la merma que se sigue de la aplicación de estandarizaciones o modelos de datos existentes. Al fin y al cabo, esta clase de análisis suele ser una misión de detección y la preparación rigurosa de los datos (como la de ETL para data warehousing) puede eliminar las valiosas pepitas que el analista trata de hallar, por ejemplo, valores atípicos que sugieran la existencia de un segmento nuevo de clientes o datos que no sean estándar e indiquen la posibilidad de fraude o acceso sin autorización.

Aparte del análisis, la preparación y la exploración de los datos suelen ir de la mano, como cuando el usuario examina recopilaciones voluminosas de datos, por lo general, las gestionadas en data lakes, depósitos de datos, hubs de datos empresariales y algunos data warehouses. A medida que avanza en la exploración, el usuario crea un conjunto de datos que se utiliza, a continuación, para el análisis o la visualización de los datos. Un ejemplo relacionado es que la preparación de datos se suele combinar con funciones de autoservicio para acceder a los datos y para crear o analizar informes, tal y como se explica en el siguiente apartado del presente informe.

Tenga en cuenta que la preparación de datos es una actividad complementaria a las prácticas tradicionales de gestión de datos. Las dos se aplican a usuarios diferentes, aplicaciones distintas y contextos diversos. En general, la nueva preparación de datos se aplica al análisis y a la exploración de los datos, no a diseños permanentes, ni informes de máxima precisión. Ambas pueden funcionar en colaboración: los conjuntos de datos creados en primer lugar mediante la preparación de datos (como ayuda para analizar y examinar los datos) pueden convertirse en prototipos permanentes una vez que se pongan en práctica los resultados de la exploración o del análisis.

Durante la aplicación práctica, los resultados de la preparación de datos se mejoran en gran medida usando funciones de calidad, transformación, modelado y agregación de datos. Por eso, una oferta moderna de integración de datos ha de incluir herramientas válidas para ambas prácticas.

A propósito de las herramientas, la preparación de datos saca partido con frecuencia de las funciones de federación y virtualización de datos, ya que resultan idóneas para las uniones de tablas, las transformaciones sencillas y el acceso a varias plataformas de datos que suele requerir. La federación y la virtualización de datos crean vistas dinámicas e integradas de datos dispares que posibilitan el funcionamiento virtual de la preparación de datos.

NÚMERO TRES

INTEGRE LOS DATOS CON MÉTODOS QUE FRANQUEEN A MÁS USUARIOS EL ACCESO DE AUTOSERVICIO A LOS DATOS NUEVOS Y AL BIG DATA

Cada vez crece más el grupo de usuarios finales que emplea más a menudo funciones de autoservicio en una amplia gama de plataformas y herramientas de software, incluidas herramientas para el análisis, la elaboración de informes y la integración de datos. Las funciones de autoservicio están diseminadas por herramientas diversas, ya que los propios usuarios se caracterizan por su diversidad: desde personal con gran experiencia técnica (analistas de datos, especialistas en datos y otros profesionales de la gestión de datos) hasta usuarios de negocio con ligeros conocimientos técnicos (administradores de datos, analistas de negocio y otros usuarios avanzados). Debido a esta diversidad, el autoservicio adopta múltiples formas, entre otras, el acceso a los datos, la preparación de datos, la elaboración de informes, la visualización y el análisis mediante autoservicio.

Las funciones de datos de autoservicio son importantes. Permiten a los usuarios trabajar con los datos con espontaneidad, rapidez y agilidad porque no tienen que esperar a que el equipo de TI o de gestión de datos cree un análisis, un informe o conjunto de datos exclusivo para ellos. A su vez, los equipos de TI o de otra materia se descargan de trabajo, ya que, si los datos de autoservicio están configurados, los usuarios crean tanto sus propios conjuntos de datos como los informes y los análisis basados en ellos. Según un informe reciente de TDWI, las cuatro tareas en las que más interesa a los usuarios de business intelligence usar el autoservicio son, por orden de prioridad, la detección de datos, la visualización, la creación de cuadros de mando y la preparación de datos. No se trata de meros deseos expresados: en ese mismo informe, se constata que la mitad de los usuarios ya emplea el autoservicio basado en datos sin problemas.[1]

La modernización de la integración de datos debe mejorar los datos de autoservicio. Las empresas muestran un enorme interés en aumentar y mejorar tanto el acceso de autoservicio a los datos como la preparación de los mismos, así como en facilitar que los usuarios se encarguen del análisis, la exploración de los datos y la elaboración de informes. Esto se logra de distintas formas:

- **Integración de datos específica para el autoservicio.** Durante años, los data warehouses y data marts cubrieron esta necesidad. Sin embargo, constan, sobre todo, de valores calculados y totalizados. Aunque siguen siendo pertinentes, la tendencia apunta a datos brutos detallados que se recopilan en bases de datos y sistemas de archivos. Las prácticas que van ganando adeptos, como la exploración de datos o el análisis avanzado, funcionan a la perfección con datos sin procesar ni retocar. Por eso, hoy en día, las soluciones modernas de integración de datos suministran los datos nuevos y el Big Data a data lakes, depósitos de datos y hubs de datos empresariales que se pueden almacenar en ecosistemas de Hadoop, bases de datos relacionales o sistemas de archivos. Recuerde que estos complementan (pero no sustituyen) los data warehouses, los data marts y los cubos tradicionales.
- **Dependencia de funciones de herramientas de autoservicio.** Las herramientas de integración de datos cuentan con funciones que no solo facilitan en gran medida el uso y ofrecen vistas de datos de fácil interpretación para los usuarios de negocio, sino que posibilitan la preparación de datos y el acceso de autoservicio a los datos. Aunque dichas funciones están diseñadas para los usuarios con menos conocimientos técnicos, TDWI ha observado que también los usuarios muy técnicos se sirven de ellas, puesto que todo el mundo se beneficia de su agilidad y su autonomía. No obstante, tenga en cuenta que, a pesar de su enorme facilidad de uso, los usuarios con menos conocimientos técnicos siguen necesitando formación tanto en la herramienta como en las mejores prácticas de la gestión de datos.

[1] Consulte la figura 9 del informe de 2015 "TDWI Best Practices Report *Emerging Technologies For Business Intelligence, Analytics, and Data Warehousing*", disponible para descarga gratuita en www.tdwi.org/bpreports.

NÚMERO CUATRO

MODERNICE LA INFRAESTRUCTURA DE INTEGRACIÓN DE DATOS APROVECHANDO LOS NUEVOS TIPOS DE PLATAFORMAS DE DATOS

Una de las novedades más apasionantes para los profesionales de datos de los últimos años ha sido la aparición de varias plataformas de datos nuevas, como la familia Hadoop de productos de código abierto y los nuevos sistemas de gestión de bases de datos (DBMS columnados, dispositivos, bases de datos gráficas y NoSQL). La mayoría se encuentra disponible para el entorno local o en el cloud, lo cual demuestra que el cloud y SaaS ya son componentes importantes de la infraestructura para la integración de datos, los DBMS y otras plataformas de datos. Pese a ser DBMS u otras clases de plataformas de datos (recuerde que Hadoop no es un DBMS), todas poseen ramificaciones positivas para modernizar la infraestructura de integración de datos.

Por ejemplo, un clúster de Hadoop desempeña varias funciones en la modernización de la integración de datos, en especial cuando esta sustenta un entorno de data warehouse multiplataforma, como en los ejemplos siguientes:

- **Hadoop es una eficaz área de descarga de datos para infinidad de velocidades de suministro y tipos de datos.** Hadoop Distributed File System (HDFS) es un sistema apto para batch de alta latencia, microbatch de baja latencia, captura de transmisiones e incorporación continua. Es más, HDFS basado en archivos puede capturar, gestionar y procesar cualquier dato susceptible de almacenamiento en un archivo.
- **Hadoop es un "staging area" de datos eficaz y escalable.** HDFS goza de fama por su escalabilidad lineal con terabytes y petabytes de datos. También es una eficaz plataforma de procesamiento en paralelo que se puede aplicar al análisis, la fusión, la transformación y la preparación de conjuntos de datos masivos.
- **Hadoop también sirve para el archivo de datos.** En muchos entornos de data warehouse, el "staging area" de datos se duplica en forma de archivo de los datos brutos detallados; en muchos casos, el volumen de datos de este archivo supera el del data warehouse real. Hadoop constituye una buena opción si debe archivar grandes cantidades de datos sin procesar, tal como hacen numerosas organizaciones para el análisis.
- **Hadoop se escala con el procesamiento mediante lógica pushdown.** La práctica de ELT habitual consiste en insertar el procesamiento de datos en una base de datos relacional de destino, como la incluida en un data warehouse o un almacén de datos operacionales. Aun así, muchos tipos de procesamiento mediante lógica pushdown también funcionan (a escala masiva) con Hadoop.
- **Hadoop descarga de trabajo el hub o la plataforma de integración de datos.** Hadoop libera capacidad del hub, la cual se puede destinar a otras rutinas de integración de datos o a soluciones nuevas; de ese modo, se facilita la escala vertical del hub.[2]

Aparte de Hadoop, existen otras plataformas de datos relativamente nuevas que facilitan el procesamiento de datos en el contexto de la integración de datos. Por ejemplo, gran parte del procesamiento mediante lógica pushdown también es relacional de forma inherente, algo en lo que Hadoop no destaca en la actualidad. Sin embargo, casi todas las plataformas basadas en columnas y dispositivos son relacionales y vienen optimizadas de serie para la lógica pushdown. Otro ejemplo lo constituyen las primeras empresas en adoptar la tecnología, que utilizan bases de datos NoSQL para procesar datos estructurados sin esquema y de forma impredecible (algo habitual en fuentes nuevas de datos como los sensores, las aplicaciones web y las redes sociales).

[2] Para obtener más información sobre los usos de Hadoop en distintas arquitecturas de datos, consulte el informe de 2014 "TDWI Best Practices Report *Evolving Data Warehouse Architectures*" y el informe de 2015 "TDWI Best Practices Report *Hadoop for the Enterprise*", ambos disponibles para descarga gratuita en www.tdwi.org/bpreports.

NÚMERO CINCO

CONTINÚE AÑADIENDO MÁS FUNCIONES OPORTUNAS A MEDIDA QUE MODERNICE LAS SOLUCIONES DE INTEGRACIÓN DE DATOS

Términos como *análisis en tiempo real*, *cuadros de mando en tiempo casi real* o *informes oportunos* llevan a confusión. La mayoría de las veces, ni los análisis, ni los cuadros de mando, ni los informes son en tiempo real, en tiempo casi real, ni oportunos. Por lo general, son la infraestructura de integración de datos y sus interfaces especializadas las que transfieren los datos de forma rápida y frecuente. De igual modo, ciertas metodologías de negocio como el business intelligence operacional, la empresa con latencia cero y la gestión del rendimiento de negocio dependen en gran medida de las funciones en tiempo real de la integración de datos. En aras de estas usuales e importantes prácticas técnicas y de negocio, los usuarios de muchos contextos siguen modernizando la integración de datos, además de sus plataformas de análisis, elaboración de informes y data warehousing, con objeto de imbuirlas de más funcionalidad en tiempo real.

El término *oportuno* da por sentado la necesidad de varias velocidades y frecuencias porque cada paso de un proceso de negocio cualquiera (o cada dato de una base de datos) puede tener su propio grado de urgencia o un plazo específico de actualización. Por ese motivo, la modernización de la integración de datos con vistas a la oportunidad implica la inclusión de varias funcionalidades técnicas, entre otras, gran rendimiento (para consultas, actualización de cuadros de mando o carga de data warehouses), microbatch (con ejecución frecuente diurna para complementar el procesamiento de batch nocturno) y federación de datos (para recuperar cantidades reducidas de datos destinadas a métricas sujetas a plazos). Muchas funciones son flexibles, así que es posible configurar su ejecución a varias velocidades oportunas, por ejemplo, replicación de datos, sincronización de datos o captura de cambios de datos. Si el procesamiento de batch es la categoría inferior de la oportunidad, el otro extremo implica el "auténtico tiempo real" (respuestas en milisegundos), el cual facilitan las herramientas para el procesamiento de eventos, el procesamiento de eventos complejos, la inteligencia operativa y el procesamiento de transmisiones.[3]

Eso son una gran cantidad de opciones y funciones oportunas. Por suerte, las plataformas modernas de integración de datos admiten multitud de funcionalidades y tipos de herramientas en un entorno de desarrollo y de desarrollo integrado. Los usuarios de estos entornos integrados con multiplicidad de herramientas tienen a su disposición numerosas opciones para que sus soluciones manejen los datos con la velocidad o la frecuencia apropiada.

Es probable que se haya dado cuenta de que muchas de las prácticas modernas en la integración de datos que ya se han mencionado a lo largo de este informe tienen algún requisito de oportunidad:

- **Incorporación de datos**: depende de numerosos índices de oportunidad, desde el tradicional procesamiento de batch nocturno hasta la necesidad de incorporación continua del procesamiento de transmisiones, con distintas gradaciones en medio.
- **Preparación de datos**: en teoría, puede aprovechar todo tipo de función de integración de datos (además de las de calidad de datos), pero presenta una tendencia a las técnicas en tiempo casi real como microbatch o federación de datos.
- **Exploración de datos** (como otras variantes del acceso de autoservicio a los datos): se presupone la respuesta inmediata del usuario, que se suele ejecutar por medio de consultas de gran rendimiento.

[3] Para obtener una explicación más detallada, lea el informe de 2014 "TDWI Best Practices Report *Real-Time Data, BI, and Analytics*", disponible para descarga gratuita en www.tdwi.org/bpreports.

NÚMERO SEIS

MODERNICE LA FUNCIONALIDAD DE LA INTEGRACIÓN DE DATOS PARA OBTENER, POR FIN, VALOR ANALÍTICO Y DE NEGOCIO DE LOS DATOS MULTIESTRUCTURADOS Y NO TRADICIONALES

Todos llevamos años alegando de boquilla la certeza absoluta de la existencia de información valiosa en los tipos de datos que no tienen los habituales formatos estructurados o relacionales. Sin embargo, pocas organizaciones han tomado medidas al respecto, ni mucho menos han utilizado esos formatos en producción. Los usuarios a quienes entrevista TDWI suelen coincidir en reivindicar sus maduros conjuntos de competencias y sus ofertas de herramientas para datos relacionales y algunos otros tipos de datos estructurados, así como las interfaces asociadas a ellos. El problema es que no son aplicables directamente, tal cual están, a los "datos novedosos" o distintos de los tradicionales, es decir, a los datos que se hallan fuera del paradigma estructurado establecido.

Como punto de partida, conviene modernizar los conocimientos y las herramientas de integración de datos a fin de habilitar la funcionalidad clave para exprimir todo el valor de negocio del nuevo Big Data y de otros datos exóticos:

Captura: las transmisiones de datos constituyen el caso extremo. Las fuentes de las transmisiones (casi siempre, máquinas de tipos distintos) *insertan* los datos en el entorno de integración de datos, que está más atrasado que el paradigma habitual de *extracción* que siguen las soluciones de integración de datos. Por eso, la plataforma de integración de datos precisa de interfaces que capturen la ingente cantidad de mensajes breves que genera la mayoría de las transmisiones y, luego, los almacenen o los procesen de la forma adecuada. Esto resulta fundamental para obtener valor de negocio del cada vez mayor número de sensores integrados o incorporados en casi cualquier componente del "Internet de las cosas", entre otros, pozos petrolíferos, camiones, vagones, palés de transporte, instalaciones físicas, intersecciones de tráfico o dispositivos móviles.

Otros ámbitos de captura son más parecidos a los casos tradicionales de integración de datos. Durante décadas, las soluciones de integración de datos han recogido y procesado archivos planos con datos algo estructurados, por lo general, archivos que contenían el volcado de una tabla, el registro de una aplicación, registros de datos modificados o un documento de intercambio de datos. En la actualidad, proliferan los datos basados en archivos por el aumento en el uso de formatos de archivo estandarizados (como XML o JSON) y registros de aplicaciones tanto web como empresariales. Las organizaciones llevan mucho tiempo adquiriendo datos de terceros para obtener información demográfica sobre los consumidores, pero, ahora, muchas también adquieren datos de redes sociales, que tienen sus propios formatos. Hacen falta plataformas modernas de integración de datos que capturen y gestionen de forma nativa los formatos basados en archivos antiguos, nuevos y en evolución y, además, permitan a los desarrolladores diseñar un modelo de compatibilidad propio con los formatos que no sean estándar.

Almacenamiento: si todos los datos que penetran en el entorno de integración de datos son casi relacionales o lo son por completo, tiene sentido almacenarlos en un DBMS relacional. Sin embargo, muchos usuarios han fracasado en su intento de transformar estructuras de datos únicas de modo que se ajusten al modelo relacional. Algunos casos fallidos son la conversión de estructuras jerárquicas planas en tabulares o el almacenamiento de cantidades ingentes de texto en lenguaje humano en forma de objetos grandes binarios. Con dichas prácticas, se falsean los datos originales, se limita la viabilidad de las consultas y las búsquedas y se entorpecen la auditoría y el linaje de datos. Un problema semejante se presenta con los formatos de archivos planos algo estructurados: si bien la mayoría se transforma con facilidad y precisión en tablas relacionales, no siempre es algo bueno, ya que la transformación conlleva una sobrecarga y el almacenamiento relacional tiene un coste relativamente elevado.

La tendencia en la descarga y el escalonado basados en integración de datos consiste en almacenar los datos en su formato original, siempre que sea posible, para que estos se procesen y transformen con métodos nuevos cuando las aplicaciones exijan otros requisitos. De este modo, los datos son aplicables a más situaciones, en lugar de verse limitados a formatos de almacenamiento que los falsean e impiden tanto la exploración como el análisis de detección.

Procesamiento: tal como apuntábamos antes, esos problemas son algunos de los motivos por los que los usuarios están implantando una gama más amplia de plataformas de datos. La diversificación de las plataformas de datos permite satisfacer los distintos requisitos de almacenamiento y procesamiento en la plataforma de los nuevos datos multiestructurados y no tradicionales de hoy en día. Debido a que estos problemas afectan tanto al data warehousing como a la integración de datos, sus arquitecturas superpuestas comparten, cada vez más, plataformas nuevas de datos, desde dispositivos hasta Hadoop.

El almacenamiento de datos en formato nativo en plataformas nuevas es incluso más viable con las nuevas plataformas de datos (basadas en columnas, dispositivos, Hadoop o NoSQL), las cuales pueden procesar conjuntos de datos masivos *in situ* sin apenas procesamiento previo ni transferencia de datos o sin nada en absoluto. Una de las mayores tendencias de modernización (que afecta a la integración de datos, el análisis, el data warehousing, etc.) consiste en llevar los algoritmos y otra lógica de procesamiento a los datos, en lugar de recurrir a la antigua costumbre de transferir los datos a la herramienta de procesamiento. Las nuevas plataformas se crearon con este fin y las marcas relacionales de DBMS más antiguas se han adaptado con análisis en la base de datos y otros tipos de procesamiento *in situ*.

Estructura: pese a la tendencia al procesamiento en el lugar, aún quedan muchos ámbitos donde alguna herramienta independiente debe acceder a los datos y procesarlos en diversas plataformas. Un caso especial lo constituyen las herramientas de minería de textos, análisis de textos y otras formas de procesamiento del lenguaje natural. Las herramientas de esta clase están optimizadas para los datos basados en archivos y casi todas ellas ofrecen una estrecha integración con Hadoop. Como los datos que se manejan poseen una estructura gramatical, que no relacional, es habitual configurar y programar esas herramientas de modo que analicen el lenguaje humano y generen estructuras de datos legibles con otras herramientas, desde registros destinados a tablas de hechos hasta estructuras de gráficos y redes neurales. (Aun así, en otros casos de uso, es preferible que la herramienta proporcione como resultado un índice de búsqueda de palabras clave).

Dada la relevancia de los casos de uso en que el texto se transforma en datos estructurados, las herramientas de procesamiento de lenguaje natural se denominan a veces "ETL para texto"; de ahí que ETL y otros tipos de herramientas de integración de datos se unan en las ofertas de los equipos modernos de integración de datos. La moraleja es que, si se impone la estructura justa sobre los datos no estructurados, se produce un resultado que puede consumir infinidad de herramientas y usuarios (tanto antiguos como nuevos) con el fin de obtener el máximo valor de negocio. Quizá los puristas lo contemplen con tono burlón, pero es coherente con la mayoría de los análisis, que suelen descubrir estructuras, relaciones y correlaciones que no se manifestaban se forma explícita en el formato original de los datos.

Metadatos: muchos tipos de datos nuevos y Big Data carecen de esquema y sus fuentes no revelan ningún diccionario de datos ni repositorio de metadatos accesible. Es lo que suele ocurrir con las entradas de los sensores y con cualquier clase de volcado de texto en lenguaje humano. Es más, la estructura implícita puede evolucionar de forma imprevisible o presentar numerosas variaciones, tal como se aprecia en los documentos JSON. Eso supone todo un reto para las herramientas tradicionales (y para los trabajadores tradicionales), ya que el acceso a los datos y su carga dependen en gran medida de los metadatos conocidos.

Aun así, los metadatos siguen desempeñando un papel importante con los nuevos formatos de Big Data y otras fuentes exóticas. En lugar de conocer y desarrollar los metadatos antes de crear la solución, se deducen *ad hoc* en tiempo de ejecución a partir de las configuraciones implícitas de valores hallados en los datos, por lo demás, no estructurados; es lo que, en ocasiones, se denomina "esquema al leerse". Una vez detectada o deducida la estructura, el desarrollador —o una función de herramienta automática— puede capturar y mejorar los metadatos. Algunos metadatos son pertinentes todo el tiempo, así que se deben registrar en un repositorio, mientras que otros solo son aplicables a una sesión de tiempo de ejecución, así que se pueden utilizar y desechar a continuación. Este es uno de los ajustes más significativos observados en la modernización de la integración de datos: las soluciones concebidas de cara al futuro deben admitir los enfoques tanto tradicional como novedoso en la gestión de los metadatos.

NÚMERO SIETE

PLANTÉESE LA MODERNIZACIÓN DE SU OFERTA DE HERRAMIENTAS DE INTEGRACIÓN DE DATOS CON UNA PLATAFORMA INTEGRADA DE VARIAS HERRAMIENTAS DE GESTIÓN DE DATOS

En un estudio de TDWI de hace unos años sobre la integración de datos de nueva generación, se preguntaba a los usuarios si utilizaban alguna herramienta de integración de datos que formara parte de una serie integrada de herramientas de gestión de datos de un proveedor. Apenas el 9 % de los encuestados contestó que usaba una, aunque el 42 % afirmó que preferiría hacerlo. En esa misma encuesta, se preguntó qué les motivaría a sustituir su herramienta de integración de datos principal. Esta fue la respuesta más elegida: "Necesitamos una plataforma unificada que admita integración de datos, así como calidad de datos, gobernanza, MDM, etc.".[4]

Desde entonces, TDWI ha entrevistado a muchos usuarios que han abandonado el enfoque de referencia por un conjunto de herramientas unificado, que es la tendencia mayoritaria entre quienes modernizan sus herramientas de integración de datos. Esta tendencia también cobra fuerza entre los principales proveedores de integración de datos, que han respondido a las exigencias de los usuarios suministrando más funciones de gestión de datos en una plataforma única estrechamente integrada.

Esa clase de plataforma integrada suele contar con una sólida herramienta de integración o calidad de datos como eje central y herramientas adicionales para la gestión de datos maestros, la gestión de metadatos, la administración, la gobernanza, la captura de cambios de datos, la replicación, el procesamiento de eventos, los servicios de datos, el perfilado de datos, la supervisión de datos, etc. Como ve, puede ser una lista bastante extensa, que llegue a acumular un impresionante arsenal de herramientas de gestión de datos relacionadas y funciones de herramientas. Sin embargo, esa dotación es una mera serie, no una plataforma integrada, a menos que las herramientas se complementen mediante una estrecha integración que posibilite las prácticas modernas.

Por ejemplo, si varias organizaciones de usuarios coordinan diferentes equipos de gestión de datos y sus soluciones (como en un centro de competencia), tiene lógica que el equipo consolidado emplee una plataforma única con objeto de facilitar la colaboración. Por lo general, a esta clase de equipos coordinados le interesa compartir datos maestros y metadatos, perfiles de conjuntos de datos, reglas de negocio, métricas de calidad, lógica y otros artefactos de desarrollo.

Otro ejemplo son los conjuntos de herramientas integrados de algunos proveedores que permiten a los usuarios diseñar un "flujo de datos" (o una construcción similar) que, en diversos pasos, ejecuta funciones de ETL, federación de datos, calidad de datos y gestión de datos maestros. Los usuarios prefieren estos diseños modernos porque reflejan el hecho real de que la mayoría de los datos sometidos a integración precisa de mejora, estandarización, fusión y eliminación de duplicados. Con una oferta de herramientas de referencia, cuesta crear el flujo de datos unificado, ya que resulta complicado conseguir un funcionamiento fiable de gran rendimiento y funcionalidad con herramientas diversas de proveedores diferentes.

Por muy completa que sea, la plataforma de integración de datos integrada no suele ser la única herramienta en uso:

- Muchas organizaciones de usuarios cuentan con una plataforma o una herramienta de integración de datos principal y estándar para casi todas las soluciones, en especial, para las de ámbito empresarial. Quizá dispongan también de herramientas secundarias más sencillas y baratas, que se aplican a los proyectos pequeños, en particular, los específicos de los departamentos.
- Con todos los cambios que están teniendo lugar en torno a la gestión de datos, es posible que el equipo necesite otras herramientas para los datos nuevos (por ejemplo, datos de procesamiento de lenguaje natural para texto) o las nuevas plataformas de datos (sobre todo, Hadoop).

Tanto si se decanta por una herramienta de referencia como si elige una plataforma de integración de datos incorporada o una combinación de ambas, exija al proveedor una actualización constante a fin de admitir las fuentes y los tipos de datos nuevos, así como las interfaces de las nuevas plataformas de datos y su procesamiento en la plataforma.

[4] Consulte las explicaciones de las figuras 2 y 5 del informe de 2011 "TDWI Best Practices Report *Next Generation Data Integration*", disponible para descarga gratuita en www.tdwi.org/bpreports.

CONCLUSIÓN

Repasemos los siete problemas de máxima prioridad que acucian la modernización de la integración de datos que hemos abordado en este informe:

Diversidad de técnicas de incorporación de datos: facilita la transferencia de datos a su propia velocidad o frecuencia de generación. De ese modo, los datos llegan a las plataformas de datos de destino cuanto antes y se encuentran disponibles para el uso inmediato en cuadros de mando, informes y análisis con fines de negocio.

Preparación de datos: permite a los analistas de datos, los especialistas en datos o usuarios parecidos crear enseguida prototipos de conjuntos de datos, sin demoras por un modelado y una estandarización excesivos. Tal rapidez es fundamental en las prácticas modernas de análisis.

Acceso de autoservicio a los datos: permite a los usuarios trabajar con espontaneidad y rapidez porque no tienen que esperar a que el equipo de TI o de gestión de datos cree un conjunto de datos para ellos. Esto es clave en prácticas modernas como el desarrollo ágil, la exploración de datos y la detección de datos.

Nuevos tipos de plataformas de datos: cuando se incorporan a una infraestructura moderna de integración de datos, ofrecen otras opciones para capturar datos distintos a los tradicionales e ingentes volúmenes de datos, aparte de transformaciones de integración de datos y procesamiento analítico.

Transferencia oportuna de datos: es el ingrediente secreto que acelera muchas prácticas de negocio sujetas a plazos, entre otras, business intelligence operacional, gestión del rendimiento y una amplia gama de análisis en tiempo real. Como el momento "oportuno" para transferir los datos varía, lo correcto es ofrecer varias características de integración de datos que funcionen a diversas velocidades y frecuencias.

Datos distintos a los tradicionales: constituyen la promesa del valor de negocio óptimo para el análisis y la toma de decisiones. Para alcanzar ese objetivo, una plataforma moderna de integración de datos debe capturar los datos que se insertan en ella, gestionar tipos de datos no estructurados, admitir enfoques nuevos de los metadatos y coordinarse con las herramientas de procesamiento de lenguaje natural.

Plataformas de herramientas integradas: incluyen numerosos tipos de herramientas para la integración de datos, la calidad de datos y la gestión de datos maestros. Se trata de herramientas con una estrecha integración para facilitar la colaboración entre los desarrolladores y para fomentar el diseño de soluciones modernas de integración de datos que activen funciones muy diversas.

ACERCA DE NUESTROS PATROCINADORES

www.informatica.com/es

Informatica es un proveedor de software independiente líder centrado en ofrecer innovaciones transformadoras para el futuro de todos los aspectos relacionados con los datos. Empresas de todo el mundo confían en Informatica para aprovechar su potencial de información y cumplir los principales imperativos de negocio. Más de 5800 empresas dependen de Informatica para aprovechar al máximo sus activos de información guardados en entornos locales, en el cloud y en Internet, incluidas las redes sociales.

http://www.sas.com/es_es/

SAS Data Management es una solución líder en el sector basada en una plataforma integrada común que permite mejorar, integrar y gobernar los datos. Independientemente de la ubicación en la que se almacenen los datos, desde sistemas heredados hasta Hadoop, SAS Data Management franquea a las organizaciones el acceso a los datos necesarios.

Las organizaciones que están modernizando sus sistemas heredados de hardware o software consideran SAS Data Management una solución indispensable para integrar y gestionar datos de todo tipo de fuentes tanto estructuradas como no estructuradas. Con sus ofertas principales de SAS Data Loader for Hadoop, SAS Event Stream Processing, SAS Data Management y SAS Federation Server, SAS satisface todos los nuevos requisitos de negocio descritos en el presente informe.

ACERCA DEL AUTOR

Philip Russom es el director de TDWI Research en materia de gestión de datos y supervisa muchos de los eventos, servicios y publicaciones orientados a la investigación de TDWI. Es una figura muy conocida en el campo de data warehousing y business intelligence y tiene en su haber más de 500 publicaciones entre informes de investigación, artículos para revistas, columnas de opinión, discursos, webinars, etc. Antes de incorporarse a TDWI en 2005, Russom trabajaba como analista del sector en materia de business intelligence en Forrester Research y Giga Information Group. También dirigía un negocio propio como analista del sector y consultor de business intelligence independiente, además de colaborar como redactor con las principales revistas de TI. Anteriormente, Russom desempeñó puestos técnicos y de marketing para varios proveedores de bases de datos. Para ponerse en contacto con él, escríbale a prussom@tdwi.org o bien búsquelo en Twitter, @prussom, o en LinkedIn, linkedin.com/in/philiprussom.

ACERCA DE TDWI RESEARCH

TDWI Research se especializa en la investigación y el asesoramiento destinados a los profesionales de business intelligence de todo el mundo. TDWI Research se centra exclusivamente en cuestiones relacionadas con business intelligence y data warehousing y se asocia con profesionales del sector para proporcionar una explicación amplia y profunda de los problemas técnicos y de negocio que rodean la implantación de las soluciones de business intelligence y data warehousing. TDWI Research ofrece informes, comentarios y servicios de consulta mediante un programa mundial de socios, así como estudios personalizados, creación de referencias y servicios de planificación estratégica a organizaciones de proveedores y usuarios.

ACERCA DE LAS LISTAS DE COMPROBACIÓN ELABORADAS POR TDWI

Las listas de comprobación elaboradas por TDWI proporcionan información general sobre los factores de éxito para un proyecto específico en el campo de business intelligence, data warehousing o una disciplina de gestión de datos relacionada. Las empresas pueden utilizar dicha información para organizarse antes de iniciar un proyecto o para identificar los objetivos y las áreas de mejora de los proyectos actuales.