TDWI RESEARCH

TDWI CHECKLIST REPORT

Modernizando la Integración de Datos

para dar Cabida al Big Data y a los Nuevos Requerimientos de las Empresas

Por Philip Russom

Patrocinado por:

informatica sas

tdwi.org

tdwi
Advancing all things data.

DICIEMBRE DE 2015

TDWI CHECKLIST REPORT

Modernizando la Integración de Datos

para dar Espacio al Big Data y a los Nuevos Requerimientos de las Empresas

Por Philip Russom

CONTENIDO

tdwi
Advancing all things data.

555 S Renton Village Place, Ste. 700
Renton, WA 98057-3295

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org

PRÓLOGO

Todos los que trabajamos en la gestión de datos estamos viviendo una época de cambios extraordinarios a medida que el big data, otras categorías de nuevos datos y las nuevas plataformas de gestión de datos entran a nuestras organizaciones. En respuesta, la mayoría de las organizaciones se están esforzando por conocer las nuevas tecnologías y, lo más importante, cómo pueden aprovechar los nuevos datos y las plataformas para tener una ventaja competitiva. En consecuencia, muchos profesionales de la gestión de datos ahora enfrentan los requerimientos nuevos y los futuros que surgirán a medida que entren a escena nuevas fuentes de datos.

Estos cambios están haciendo que muchas organizaciones técnicas replanteen y modernicen su infraestructura, sus equipos y sus habilidades para gestionar los datos. Entre estos esfuerzos, la modernización de la integración de datos es uno de los más apremiantes debido al papel que juega la integración de datos (ID) para capturar, procesar y mover datos, tanto viejos como nuevos. Sin las soluciones de ID modernas, las organizaciones no pueden satisfacer los requerimientos nuevos y futuros del big data, la analítica y la operación en tiempo real.

La modernización de la ID puede adoptar muchas formas, dependiendo del estado actual de su infraestructura y de los tipos de nuevos datos o plataformas que usted debe adoptar. En lugar de tratar de englobar aquí a todos ellos, ofrecemos recomendaciones que pueden servir de guía a sus esfuerzos de modernización cuando llegue el momento de seleccionar los productos y actualizar sus diseños de soluciones:

1. Complemente la alta latencia de las viejas prácticas de la integración de datos (ID) con una gama más amplia de técnicas de ingestión de datos.
2. Adopte las nuevas prácticas y herramientas de preparación de datos para tener agilidad, velocidad, simplicidad y facilidad de uso.
3. Integre los datos de formas que permiten el acceso de autoservicio a datos nuevos y big data para una amplia variedad de usuarios.
4. Modernice su infraestructura de integración de datos al aprovechar los nuevos tipos de plataformas de datos.
5. Continúe agregando más funciones oportunas a medida que moderniza sus soluciones de integración de datos.
6. Modernice su funcionalidad de integración de datos para finalmente obtener valor de negocio y analítico de los datos multi-estructurados y no tradicionales.
7. Considere modernizar su portafolio de herramientas de ID con una plataforma integrada de múltiples herramientas de gestión de datos.

Este TDWI Checklist Report va al fondo de cada una de estas siete recomendaciones y discute muchos de los nuevos tipos de productos de los proveedores, su funcionalidad y las mejores prácticas que pueden contribuir a la modernización de la ID. Asimismo, presentamos el caso de negocio y las fortalezas tecnológicas de cada recomendación.

NÚMERO 1

COMPLEMENTE LA ALTA LATENCIA DE LAS VIEJAS PRÁCTICAS DE LA INTEGRACIÓN DE DATOS (ID) CON UNA GAMA MÁS AMPLIA DE TÉCNICAS DE INGESTIÓN DE DATOS

Uno de los cambios más radicales en la práctica de la integración de datos de los años recientes, es la modernización de la ingestión de datos. La ingestión es sencillamente cómo, dónde y con qué frecuencia los datos que entran a un entorno son integrados o cargados en los objetivos (como las áreas de preparación de datos, los almacenes de datos o los sistemas de archivos). Durante décadas, los procesos de ingestión de las prácticas de integración de datos al estilo de ETL han estado "latentes" (por ejemplo, consumiendo tiempo, a menudo ejecutándose por las noches). Los datos de la encuesta de TDWI muestran que la mayoría de los datos en la mayoría de los almacenes de datos se actualizan cada 24 horas. Sin embargo, el porcentaje de datos que se recolectan de forma rápida o frecuente, se preparan y se entregan para su presentación sigue incrementándose por varias razones:

Algunas nuevas fuentes de datos generan datos frecuentemente. Una categoría importante del big data son los datos de las máquinas. Los sensores se integran (o se añaden) a una lista cada vez más grande de máquinas, incluyendo vehículos, dispositivos portátiles y robots de las líneas de producción. Además de las máquinas, los sensores GPS están proliferando en los palés de embarque y otros recursos móviles. Algunos sensores generan y transmiten datos en un flujo continuo de eventos; otros transmiten únicamente cuando se conectan a una red (como con chips RFID) o cuando las máquinas de los sensores entran en acción (como cuando un robot instala un widget). El punto es que muchas organizaciones quieren capturar y aprovechar las nuevas fuentes para mejorar la logística, el monitoreo de los sentimientos, los acuerdos de nivel de servicio y el cumplimiento de las cuotas, la vigilancia de instalaciones, la analítica operativa y el monitoreo de las actividades del negocio.

Las prácticas comerciales que requieren datos muy recientes siguen creciendo. Los usuarios han estado poniendo en práctica la inteligencia de negocio operativa (OpBI), que con frecuencia actualiza los dashboards de gestión y otros reportes operativos con datos cuya frescura es de varios minutos a algunas horas. Similares a OpBI, las prácticas en la gestión del desempeño, el reporteo, OLAP y la analítica avanzada demandan datos cada vez más frescos. Los datos frescos le dan a estas empresas una mayor ventaja competitiva, mejoran las relaciones con los clientes, mejoran la excelencia operativa, y permiten tomar decisiones tácticas ágiles pero informadas.

Las prácticas de ingestión de datos necesitan acomodar datos de muchas velocidades y frecuencias. No lo olvide, usted aún necesita de ETL latente y de ELT para conservar la precisión y para las uniones extremas, las transformaciones y la ruta de auditoría de datos típica para el almacenamiento de datos, la mayoría de los reportes estándares y muchas soluciones OLAP. El desafío radica en diseñar nuevas soluciones de integración de datos (ID) (o ajustar las anteriores) para capturar e ingerir nuevos datos con mayor rapidez y frecuencia. Esto es algo a lo que se llama algunas veces ingestión inicial o ingestión continua, que es muy rápida y frecuente en comparación con las cargas en lote nocturnas. Para compensar, la ingestión inicial realiza poca o ninguna transformación o agregación de datos antes de cargarse porque eso haría lenta la ingestión. Un beneficio es que los datos son capturados en su estado original, lo que significa que se pueden readaptar repetidamente a medida que surgen nuevos requerimientos para el reporteo y la analítica. El mayor beneficio es que los datos están listos lo antes posible para el reporteo, la analítica y las operaciones.

Una vez más, un pequeño desafío con este enfoque es que la readaptación de los datos se está realizando cada vez más sobre la marcha, durante el tiempo de ejecución (en lugar de hacerlo antes del tiempo de carga), cuando un analista de datos o un científico de datos explora los datos y desarrolla nuevos conjuntos de datos para la analítica. Como otro ejemplo, una rutina de DI puede analizar los datos que han cambiado (recientemente ingeridos) para actualizar los reportes operativos o a los análisis durante el día hábil. Un director de ventas puede refrescar un cubo para ver las ventas de hoy hasta ahora.

El hardware y el software modernos de hoy son veloces y escalables, por lo que el rendimiento de la ingestión continua y el procesamiento del tiempo de ejecución ahora son prácticos. Además, la funcionalidad de las herramientas para el procesamiento de datos sensible al tiempo ha alcanzado un nuevo nivel de madurez, como se ve en las funciones que se explicarán en este reporte, tales como el procesamiento de flujos de eventos, la federación de datos, el acceso a datos de autoservicio y la preparación de los datos.

Los datos ingeridos rápidamente también pueden procesarse de formas tradicionales. La ingestión de datos pone los datos a disposición de las tecnologías y de los usuarios que los necesitan de inmediato, mientras que la "operacionalización" de los datos capturados posteriormente recurre a las mejores prácticas establecidas en la calidad de datos, el modelado y la recolección. Por ejemplo, el procesamiento en línea de los datos de los robots de manufactura puede revelar lotes malos así como otros problemas con los materiales que requieren de atención inmediata. Los mismos datos estudiados fuera de línea revelan tendencias de largo plazo igualmente valiosas en el rendimiento de los proveedores relativo a la calidad de los productos.

NÚMERO DOS

ADOPTE LAS NUEVAS PRÁCTICAS Y HERRAMIENTAS DE PREPARACIÓN DE DATOS PARA TENER AGILIDAD, VELOCIDAD, SIMPLICIDAD Y FACILIDAD DE USO

Como ya se mencionó anteriormente, procesar datos en el momento se ha convertido en una práctica de integración de datos distinta. La práctica tiene muchos nombres, incluyendo el de discusión de datos, limpieza de datos y combinación de datos. Algunas personas la llaman "ID ligera" porque sus implementaciones normalmente son un pequeño subconjunto de la funcionalidad de ID, recortado por razones de usabilidad y de desempeño. Sin embargo, el nombre que se escucha más a menudo en TDWI es "preparación de datos". Varios tipos de herramientas soportan alguna forma de preparación de datos, incluyendo aquellas para la integración de datos, el perfilamiento de datos, la calidad de datos, la exploración de datos, la analítica y la visualización de datos.

En particular, la preparación de datos ahora es común para muchas formas de analítica. Permite que el analista de datos, el científico de datos, o un usuario similar, trabaje con datos detallados de las fuentes (llenos de detalles ricos) sin ser paralizados por los modelos de datos y por las estandarizaciones existentes. Después de todo, normalmente este tipo de analítica es una misión de descubrimiento, y la preparación rigurosa de los datos (como lo hace ETL para el almacenamiento de datos [DW]) puede eliminar los nuggets valiosos que un analista está tratando de descubrir, como los valores atípicos que sugieren un nuevo segmento de clientes o los datos no estándares que sugieren un fraude o el acceso no autorizado.

Además de la analítica, la preparación y la exploración de datos a menudo van juntas, como cuando un usuario explora grandes colecciones de datos, normalmente aquellas que se manejan en lagos de datos, las bóvedas de datos, los centros de datos empresariales, y en algunos almacenes de datos. El usuario crea un conjunto de datos cuando realiza la exploración, el cual se utiliza para el análisis o la visualización de datos. En un ejemplo relacionado, la preparación de datos se combina a menudo con funciones para el acceso de datos de autoservicio y la creación o análisis de reportes de autoservicio, como se abordará en la siguiente sección de este reporte.

Tenga en cuenta que la preparación de datos es complementaria a las prácticas tradicionales de gestión de datos. Ambas se aplican a diferentes usuarios, aplicaciones y otros contextos. En general, la nueva preparación de datos es normalmente para la exploración de datos y para la analítica, no para los diseños permanentes o para reportes altamente precisos. Pueden trabajar juntas: los conjuntos de datos primero construidos a través de la preparación de datos (para apoyar la exploración de datos y las prácticas analíticas) pueden convertirse en un prototipo para los conjuntos de datos permanentes cuando el resultado de la exploración o del análisis se operacionalice. Durante la operacionalidad, se mejora considerablemente el resultado de la preparación de datos, usando funciones para la calidad, la transformación, el modelado y la suma de los datos. Por lo tanto, el moderno portafolio de ID debe incluir herramientas que soporten ambas prácticas.

Hablando de herramientas, la preparación de datos normalmente utiliza las funciones de federación y de virtualización de datos. Estas son ideales para las uniones de tablas, las transformaciones ligeras y el acceso a múltiples plataformas de datos que normalmente requieren de la preparación de datos. La federación y la virtualización de datos crean vistas dinámicas e integradas de los datos dispares que permiten que la preparación de datos opere virtualmente.

NÚMERO TRES

INTEGRE LOS DATOS DE FORMAS QUE PERMITEN EL ACCESO DE AUTOSERVICIO A DATOS NUEVOS Y BIG DATA PARA UNA AMPLIA VARIEDAD DE USUARIOS

Una creciente clase de usuarios finales está haciendo un uso excelente de las funciones de autoservicio en una amplia variedad de herramientas y plataformas de software, incluyendo herramientas para reporteo, analítica e integración de datos. Las funciones de autoservicio están distribuidas en diversas herramientas porque los usuarios mismos son diversos, desde el personal altamente técnico (analistas de datos, científicos de datos y otros profesionales de la gestión de datos) hasta la gente de negocios moderadamente técnica (administradores de datos, analistas y otros usuarios). Debido a esta diversidad, el autoservicio adopta múltiples formas, incluyendo el acceso a datos, la preparación de datos, la creación de reportes, la visualización y la analítica de autoservicio.

Las funciones de datos de autoservicio son importantes. Le permiten a los usuarios trabajar con los datos con espontaneidad, velocidad y agilidad porque los usuarios no están esperando a que TI o que un equipo de gestión de datos creen un conjunto de datos, un reporte o un análisis únicos para ellos. TI y otros equipos, a su vez, se liberan cuando los datos de autoservicio se preparan para que los usuarios puedan crear sus propios conjuntos de datos y los reportes y el análisis basados en ellos. De acuerdo con un reporte reciente de TDWI, las cuatro tareas que los usuarios de BI quieren realizar más a través del autoservicio son (en orden de importancia) el descubrimiento de datos, la visualización, la creación de dashboards, y la preparación de datos. Esto es más que un deseo; el mismo reporte revela que la mitad de los usuarios ya están practicando con éxito el autoservicio determinado por los datos.[1]

La modernización de la integración de datos debe mejorar los datos de autoservicio. Las empresas tienen gran interés en apoyar un mejor acceso a los datos y la preparación de datos de autoservicio, y que los usuarios realicen la exploración, el reporteo y la analítica de datos. Esto puede lograrse de formas distintas:

- **Integrar datos específicamente para el autoservicio.** Durante años, los almacenes de datos y los data marts cumplieron con este requerimiento. Sin embargo, los almacenes y marts son principalmente valores sumados y calculados. Estos aún son relevantes, aunque la tendencia apunta hacia los datos detallados de fuentes recolectados en bases de datos y sistemas de archivos. Las prácticas como la exploración de datos y la analítica avanzada funcionan bien con datos brutos e intactos. En respuesta, las soluciones modernas de integración de datos llevan los nuevos datos y el big data a los lagos de datos, a las bóvedas de datos y a los centros de datos empresariales que pueden albergarse en los ecosistemas Hadoop, bases de datos relacionales o sistemas de archivos. Note que estos complementan (pero no remplazan) a los almacenes, marts y cubos tradicionales.
- **Depender de las funciones de las herramientas de autoservicio.** En las herramientas de integración de datos, las funciones involucran la alta facilidad de uso y las vistas de datos amigables con las empresas, que pueden permitir el acceso de datos y la preparación de datos de autoservicio. Aunque está diseñada para usuarios menos técnicos, TDWI también ve que los usuarios altamente técnicos están aprovechando estas funciones porque todos se benefician de la agilidad y de la autonomía. Tenga en cuenta que aun cuando la facilidad de uso es alta, los usuarios menos técnicos necesitan capacitarse sobre la herramienta y las mejores prácticas de la gestión de datos.

[1] Ver la Figura 9 del 2015 TDWI Best Practices Report, Tecnologías Emergentes para la Inteligencia de Negocio, la Analítica y el Almacenamiento de Datos. Disponible como una descarga gratuita en www.tdwi.org/bpreports.

NÚMERO CUATRO

MODERNICE SU INFRAESTRUCTURA DE INTEGRACIÓN DE DATOS AL APROVECHAR LOS NUEVOS TIPOS DE PLATAFORMAS DE DATOS

Uno de los desarrollos más interesantes de los años recientes para los profesionales de los datos es la llegada de varias nuevas plataformas de datos, como la familia de productos de código abierto Hadoop y los nuevos sistemas de gestión de bases de datos (DBMSs, dispositivos, bases de datos de gráficos y NoSQL). La mayoría de estos ya están disponibles localmente o en la nube, lo que muestra que la nube y el SaaS son ahora componentes importantes de la infraestructura para ID, DBMSs y otras plataformas de datos. Si bien estas son DBMSs y otros tipos de plataformas de datos (tenga en cuenta que Hadoop no es un DBMS), todas tienen ramificaciones positivas para modernizar su infraestructura de ID.

Por ejemplo, un cluster Hadoop tiene varios papeles que jugar en la modernización de la ID, especialmente cuando la ID soporta un entorno de almacenamiento de datos (DWE) multiplataforma, como en los siguientes ejemplos:

- **Hadoop es un área de aterrizaje efectiva para muchas velocidades de alimentación (feed) y tipos de datos.** El Hadoop Distributed File System (HDFS) es adecuado para lotes de baja latencia, micro lotes de baja latencia, captura de flujos y la ingestión continua. Además, el HDFS basado en archivos puede capturar, gestionar y procesar los datos que puedan almacenarse en un archivo.
- **Hadoop es un área de preparación de datos escalable y poderosa.** El HDFS es conocido por la escalabilidad lineal con terabytes y petabytes de datos. También es una plataforma poderosa de procesamiento paralelo que puede aplicarse para analizar, fusionar, transformar y preparar conjuntos masivos de datos.
- **Hadoop también es adecuado para archivar datos.** En muchos entornos de almacenamiento de datos, el área de preparación de datos también sirve como un archivo de datos detallados de las fuentes; en muchos casos, el volumen de datos de este archivo supera a los del almacén real. Hadoop puede ser una buena opción si usted está archivando grandes cantidades de datos brutos como muchas organizaciones lo hacen para la analítica.
- **Hadoop escala con el procesamiento pushdown.** La práctica de ELT popular normalmente empuja el procesamiento de datos a una base de datos relacional como la que está bajo un almacén de datos o de un almacén de datos operacional. Sin embargo, muchos tipos de procesamiento pushdown también funcionan (a una escala masiva) con Hadoop.
- **Hadoop puede descargar su plataforma o centro de ID.** Hadoop libera la capacidad del centro que puede aplicarse a otras rutinas de ID o nuevas soluciones, ayudando así al centro a escalar.[2]

Además de Hadoop, otras plataformas de datos relativamente nuevas pueden contribuir al procesamiento de datos en un contexto de ID. Por ejemplo, mucho del procesamiento pushdown es inherentemente relacional, algo en lo que Hadoop es débil actualmente. Sin embargo, la mayoría de las plataformas basadas en columnas y dispositivos son relacionales y están optimizadas para dichos pushdowns. En otro ejemplo, quienes las adoptaron primero están utilizando bases de datos NoSQL para procesar datos libres de esquemas y datos estructurados impredecibles (como es típico de las nuevas fuentes de datos como sensores, aplicaciones Web, y medios sociales).

[2] Para consultar detalles adicionales sobre los usos de Hadoop en múltiples arquitecturas de datos, vea el 2014 TDWI Best Practices Report La Evolución de las Arquitecturas de los Almacenes de Datos y el 2015 TDWI Best Practices Report, Hadoop para la Empresa, ambos están disponibles como descargas gratuitas en www.tdwi.org/bpreports.

NÚMERO CINCO
CONTINÚE AGREGANDO MÁS FUNCIONES OPORTUNAS A MEDIDA QUE MODERNIZA SUS SOLUCIONES DE INTEGRACIÓN DE DATOS

Términos como analítica en tiempo real (real-time), dashboard en tiempo casi real (near-time), y reportes en el momento adecuado (right-time) son engañosos. La mayor parte del tiempo no son la analítica, los dashboards o los reportes los que son en tiempo real, casi real o en el momento adecuado. Normalmente es la infraestructura de integración de datos y sus interfaces especializadas las que se mueven rápidamente y con frecuencia. Asimismo, las metodologías de negocio como BI operacional, empresa de cero latencia y gestión del desempeño del negocio dependen enormemente de las funciones en tiempo real de la ID. En aras de estas prácticas técnicas y de negocio populares e importantes, los usuarios de muchos contextos continúan modernizando la ID – pero también el reporteo, la analítica y las plataformas de almacenamiento de datos – para infundirles más funcionalidad de tiempo real.

El término en el momento adecuado supone que hay muchas velocidades y frecuencias necesarias porque cada paso del proceso de negocio (o cada dato en una base de datos) puede tener su propio grado de urgencia o marco de tiempo de actualización. Es por eso que modernizar la DI para el momento adecuado implica soportar varias funcionalidades técnicas. Éstas incluyen el alto desempeño (para consultas, la actualización de dashboards, carga de almacenes) el micro lote (ejecutando frecuentemente durante el día para complementar el procesamiento en lote nocturno), y la federación de datos (para buscar pequeñas cantidades de datos para métricas sensibles al tiempo). Muchas funciones son flexibles y pueden configurarse para correr a múltiples velocidades en el momento adecuado, como con la réplica de datos, la sincronización de datos y la captura de datos cambiados. Si el procesamiento en lote es el extremo bajo del momento adecuado, entonces el otro extremo involucra el "verdadero tiempo real" (respuestas en milisegundos), que es posible gracias a las herramientas para el procesamiento de eventos, el procesamiento de eventos complejos, la inteligencia operativa y el procesamiento de flujos.[3]

Son muchas funciones y opciones del momento oportuno. Por suerte, las plataformas de integración de datos modernas soportan múltiples tipos de herramientas y funcionalidades en un entorno integrado de desarrollo y de implementación. Los usuarios de estos entornos integrados de múltiples herramientas tienen muchas opciones a su disposición, de modo que sus soluciones pueden manejar datos a la velocidad o frecuencia ideales.

Usted probablemente notó que muchas de las prácticas modernas de la integración de datos de las que se habló anteriormente en este reporte tienen un requerimiento de momento oportuno:

- **La ingestión de datos** depende de muchas velocidades de momento oportuno, desde el procesamiento en lotes nocturno tradicional hasta la ingestión continua requerida para el procesamiento de flujos, más los muchos puntos intermedios.
- **La preparación de datos** teóricamente podría aprovechar cualquier tipo de función de ID, además de aquellas para la calidad de datos, pero tiende hacia las técnicas de tiempo casi real como la federación de datos y el micro lote.
- **Exploración de datos** (como otras variaciones del acceso de datos de autoservicio) supone una respuesta inmediata para el usuario, que normalmente se realiza a través de consultas de alto desempeño).

[3] Para una discusión más detallada, lea el 2014 TDWI Best Practices Report, Datos, BI y Analítica en Tiempo Real, disponible como una descarga gratuita en www.tdwi.org/bpreports.

NÚMERO SEIS

MODERNICE SU FUNCIONALIDAD DE INTEGRACIÓN DE DATOS PARA FINALMENTE OBTENER VALOR DE NEGOCIO Y ANALÍTICO DE LOS DATOS MULTI-ESTRUCTURADOS Y NO TRADICIONALES

Durante años hemos estado hablando del servicio, diciendo que sabemos que hay información valiosa en los tipos de datos que no están en los formatos estructurados o relacionales usuales. Pero pocas organizaciones han actuado, mucho menos han utilizados estos formatos en producción. Los usuarios entrevistados por TDWI regularmente hablan de cómo han madurado sus habilidades y sus portafolios de herramientas para los datos relacionales y otros pocos tipos de datos estructurados, además de las interfaces asociadas a estos. El caso es que estos no se aplican directamente, como tales, a los datos no tradicionales o a los "datos novedosos", esto es, los datos que están fuera del paradigma estructurado establecido.

Sin embargo, un punto de inicio exitoso es modernizar las habilidades y las herramientas de integración de datos para permitir la funcionalidad que es clave para obtener el valor de negocio de los nuevos big data y de otros datos exóticos:

Capturar: Los datos continuos son el caso extremo. Las fuentes de flujos (principalmente máquinas de diferentes tipos) empujan los datos a su entorno de ID, que es inverso al paradigma usual de jalar que adoptan las soluciones de TI. Como resultado, su plataforma de ID necesita interfaces que puedan capturar los grandes números de pequeños mensajes que la mayoría de los flujos generan, y después almacenarlos o procesarlos adecuadamente. Esto es importante para obtener el valor de negocio de explotar varios sensores que están integrados o que se están agregando a casi todo en el Internet de las Cosas (IoT), incluyendo pozos petroleros, camiones, vagones, palés de embarque, plantas físicas, intersecciones de tráfico y dispositivos móviles.

Otros escenarios de captura son más conocidos para los casos de ID tradicionales. Durante décadas, las soluciones de ID han retomado y procesado archivos planos que contienen datos ligeramente estructurados, normalmente archivos que contienen un vertedero de tablas, un registro de aplicaciones, registros de datos cambiados o un documento de intercambio de datos. Hoy, los datos basados en archivos están aumentando debido a un mayor uso de formatos de archivos estandarizados (como XML o JSON) y de registros de las aplicaciones empresariales y Web. Las organizaciones han adquirido durante mucho tiempo datos de terceros para las demografías de los consumidores, pero ahora muchos también adquieren de datos de los medios sociales, que tienen sus propios formatos. Los usuarios necesitan plataformas de ID modernas que puedan capturar y manejar nativamente los formatos de archivos viejos, nuevos y en evolución, además de permitirle a los desarrolladores diseñar su propio formato para los formatos no estándares.

Almacenamiento: Si todos los datos que llegan a su entorno de ID son totalmente o casi relacionales, entonces tiene sentido almacenarlos en un DBMS relacional. Sin embargo, ya hay una historia de fallas entre los usuarios que han intentado transformar estructuras de datos únicas para que encajen en el modelo relacional. Los casos de uso fallidos incluyen nivelar las estructuras jerárquicas en estructuras tabulares y almacenar cantidades masivas de texto en lenguaje humano como grandes objetos binarios. Dichas prácticas tergiversan los datos originales, limitan la viabilidad de las consultas y búsquedas, y oscurecen el linaje y la auditoría de los datos. En un problema relacionado, la mayoría de los formatos de archivos planos ligeramente estructurados se transforman fácilmente y con exactitud en tablas relacionales. Esto no es universalmente bueno debido al costo de la transformación y al costo relativamente alto del almacenamiento relacional.

La tendencia en el aterrizaje y preparación determinados por la ID es almacenar los datos en su forma original cuando es posible de modo que los datos puedan procesarse y transformarse de nuevas formas cuando surgen nuevos requerimientos de las aplicaciones. Así, los datos se aplican a más situaciones en lugar de estar limitados por los formatos de almacenamiento que tergiversan los datos e inhiben la exploración y la analítica de descubrimiento.

Procesamiento: Estos problemas se encuentran entre las razones por las que los usuarios están implementando una gama más amplia de plataformas de datos, como se discutió anteriormente en este reporte. El punto de las plataformas de datos diversas es satisfacer los múltiples requerimientos de almacenamiento y de procesamiento en las plataformas de los nuevos datos multi-estructurados y no tradicionales de hoy. Estos problemas afectan al almacenamiento de datos y a la integración de datos, que es por qué sus arquitecturas que se superponen comparten nuevas plataformas de datos, desde los dispositivos hasta Hadoop.

Almacenar datos en los formatos nativos en nuevas plataformas se vuelve aún más práctico gracias a las nuevas plataformas de datos (basadas en columnas, dispositivos, Hadoop, NoSQL) que pueden procesar conjuntos de datos masivos in situ con poco o ningún procesamiento o movimiento de datos. Una de las tendencias de modernización más fuertes (que afectan a la ID, la analítica, los almacenes de datos, etc.) es llevar algoritmos y otra lógica de procesamiento a los datos en lugar del viejo hábito de mover los datos a una herramienta de procesamiento. Las nuevas plataformas se hicieron para esto, y las viejas marcas relacionales de DBMSs han sido readaptadas con la analítica en la base de datos y otro procesamiento in situ.

Estructurar: A pesar de la tendencia hacia el procesamiento en el sitio, aún hay muchos escenarios en los que una herramienta independiente necesita acceder a los datos y procesarlos en una variedad de plataformas. Como un caso especial, considere las herramientas para la minería de textos, la analítica de textos y otras formas de procesamiento de lenguaje natural (NLP). Esta clase de herramientas está optimizada para los datos basados en archivos, y la mayoría de las herramientas tienen una estrecha integración con Hadoop. Debido a que los datos en la que operan tienen una estructura gramatical – pero no una estructura relacional – estas herramientas a menudo se configuran y se programan para analizar el lenguaje humano y generar estructuras de datos que puedan ser leídas por otras herramientas. Estas estructuras van desde registros que se integran en una tabla de hechos hasta una red neural y las estructuras gráficas. (Incluso, otros casos de uso pueden preferir un índice de búsquedas de palabras clave como resultado de la herramienta).

Debido a la importancia de los casos de uso que transforman el texto en datos estructurados, las herramientas de NLP algunas veces son denominadas "ETL para texto", y por tanto se están uniendo a ETL y a otros tipos de herramientas ID e los portafolios de equipos modernos de integración de datos

Aquí la lección es que imponer la estructura suficiente en los datos no estructurados produce un resultado que muchas herramientas y usuarios – nuevos y viejos – pueden consumir para maximizar el valor de negocio. Los puristas pueden mofarse, pero esto es consistente con la mayoría de las analíticas, las que regularmente revelan estructuras, relaciones y correlaciones que no eran explícitas en el formato original de los datos.

Metadatos: Muchos tipos de nuevos datos y big data están libres de esquemas, y sus fuentes no exponen un repositorio de metadatos accesible o un diccionario de datos. Esto es típico de muchas aportaciones de los sensores (sensor feeds) y de todo lo que involucre los vertederos de texto de lenguaje humano. Además, la estructura implícita puede evolucionar de manera impredecible o tener muchas variaciones, como se ve en los documentos JSON. Este es un desafío para las herramientas tradicionales – ¡y para los trabajadores tradicionales! – donde el acceso y la carga de datos dependen enormemente de los metadatos conocidos.

Incluso, los metadatos siguen jugando un papel importante dentro de los nuevos formatos de big data y otras fuentes exóticas. Sin embargo, en lugar de conocer y desarrollar metadatos antes de crear una solución, los metadatos pueden ser deducidos ad hoc en el tiempo de operación de las configuraciones implícitas de los valores encontrados en los datos no estructurados. A esto se le llama algunas veces "esquema en la lectura". Una vez que se descubre o se deduce una estructura, un desarrollador o la función de una herramienta automática pueden capturar y mejorar los metadatos. Algunos metadatos son relevantes con el tiempo y por lo tanto deben ser registrados en un repositorio, mientras que otros metadatos sólo se aplican a una sesión de tiempo de operación y por tanto pueden usarse y descartarse. Este es uno de los ajustes más profundos vistos en la modernización de la integración de datos – las soluciones innovadoras deben soportar los enfoque tradicionales y nuevos de la gestión de metadatos.

NÚMERO SIETE

CONSIDERE MODERNIZAR SU PORTAFOLIO DE HERRAMIENTAS DE ID CON UNA PLATAFORMA INTEGRADA DE MÚLTIPLES HERRAMIENTAS DE GESTIÓN DE DATOS

Hace algunos años, una encuesta de TDWI sobre la integración de datos de próxima generación le preguntó a los usuarios si estaban "usando una herramienta de ID que fuera parte de una suite integrada de herramientas de gestión de datos de un proveedor". Sólo 9 por ciento de los entrevistados reportó que estaba usando una, aunque 42 por ciento preferiría una. En la misma encuesta se preguntó qué haría que los usuarios remplazaran su principal herramienta de ID. La respuesta más popular fue: "Necesitamos una plataforma unificada que soporte ID, además de la calidad de datos, el gobierno, MDM, etc."[4]

Desde entonces, TDWI ha entrevistado a muchos usuarios que han abandonado el mejor enfoque a favor de un juego de herramientas unificado, haciendo a esta la tendencia más fuerte entre los usuarios que modernizan sus herramientas de ID. Esto también es una tendencia entre los proveedores de ID líderes; estos proveedores han respondido a la demanda de los usuarios al proveer funciones adicionales de gestión de datos en una sola plataforma estrechamente integrada.

Dicha plataforma integrada normalmente tiene una herramienta de ID y/o de calidad de datos en su interior, con herramientas adicionales para la gestión de datos maestros, la gestión de metadatos, la administración, el gobierno, la captura de datos cambiados, la replicación, el procesamiento de eventos, los servicios de datos, el perfilamiento de datos, el monitoreo de datos, etcétera. Como puede ver, la lista puede ser bastante larga, reuniendo un arsenal impresionante de herramientas de gestión de datos relacionadas y funcionalidades. Sin embargo, el arsenal es una simple suite – no una plataforma integrada – a menos que las herramientas se integren estrechamente de una forma que permita las prácticas modernas.

Por ejemplo, cuantas más organizaciones usuarias coordinen diversos equipos de gestión de datos y sus soluciones (como en un centro de competencia), tiene sentido que el equipo consolidado utilice una sola plataforma para una colaboración más sencilla. Los equipos coordinados de este tipo generalmente quieren compartir metadatos y datos maestros, los perfiles de los conjuntos de datos, las reglas de negocio, las métricas de calidad, la lógica y otros artefactos de desarrollo.

Como otro ejemplo, considere los conjuntos de herramientas integrados de algunos proveedores que le permiten a los usuarios diseñar un "flujo de datos" o una construcción similar, que en varios pasos del flujo realiza funciones para ETL, la federación de datos, la calidad de datos y la gestión de datos maestros. A los usuarios les gustan estos diseños modernos porque reflejan el hecho del mundo real de que la mayoría de los datos que se están integrando necesitan múltiples mejoras, estandarizaciones, fusiones y deduplicaciones. Tratar de alcanzar un flujo de datos unificado con un portafolio de herramientas es problemático debido a los desafíos de hacer que múltiples herramientas de múltiples proveedores interoperen de manera confiable con el alto desempeño y una funcionalidad rica.

[4] Ver las discusiones sobre las Figuras 2 y 5 en el 2011 TDWI Best Practices Report, Integración de Datos de Próxima Generación, disponible como una descarga gratuita en www.tdwi.org/bpreports.

Aunque completa, una plataforma de ID integrada es raramente la única herramienta en uso:

- Muchas organizaciones usuarias tienen una herramienta o plataforma ID principal, que es el estándar para la mayoría de las soluciones, especialmente aquellas en el ámbito empresarial. También pueden tener herramientas secundarias que son más sencillas, más accesibles y que se aplican a proyectos más pequeños, especialmente a proyectos departamentales.
- Con todos los cambios que ocurren en la gestión de datos, un equipo podría necesitar herramientas adicionales para los nuevos datos (como NLP para datos de texto) o nuevas plataformas (especialmente Hadoop).

Ya sea que usted opte por una plataforma de ID integrada, o una combinación de ambos, exija que su proveedor de herramientas tenga soporte actualizado para las nuevas fuentes de datos y tipos de datos, así como con interfaces a las nuevas plataformas de datos y al procesamiento en la plataforma para ellos.

CONCLUSIÓN

Revisemos los siete problemas de alta prioridad de la modernización de la integración de datos (ID) discutidos en este reporte:

Las múltiples técnicas de ingestión de datos permiten que los datos se muevan a su propia velocidad y frecuencia de generación. Así, los datos llegan a las plataformas de datos objetivo tan rápido como sea posible y están disponibles para usarse de inmediato en los dashboards, reportes y la analítica.

La preparación de datos permite que los analistas de datos, los científicos de datos o un usuario similar construyan un prototipo de conjunto de datos rápidamente sin que el modelado y la estandarización excesivos los haga lentos. Dicha velocidad es crítica para las prácticas modernas de la analítica.

El acceso a los datos de autoservicio ayuda a los usuarios a trabajar con espontaneidad y velocidad porque no están esperando a que el equipo de TI o de gestión de datos construya un conjunto de datos para ellos. Esto es clave para las prácticas modernas como son el desarrollo ágil, la exploración de datos y el descubrimiento de datos.

Los nuevos tipos de plataformas, cuando se incorporan a una nueva infraestructura de integración de datos, ofrecen nuevas opciones para capturar datos no tradicionales y volúmenes masivos de datos, así como para el procesamiento analítico y las transformaciones de ID.

El movimiento de datos en el momento adecuado es el ingrediente secreto que acelera muchas prácticas de negocio sensibles al tiempo, incluyendo el BI operacional, la gestión del desempeño y una amplia gama de analítica en tiempo real. Debido a que hay muchos momentos "adecuados" para mover los datos, la habilitación adecuada normalmente implica múltiples funciones de integración de datos que operan a múltiples velocidades y frecuencias.

Los datos no tradicionales prometen un gran valor de negocio para la toma de decisiones y la analítica. Para apoyar ese objetivo, una plataforma moderna de integración de datos debe capturar los datos que se empujan hacia ella, manejar los tipos de datos no estructurados, soportar nuevos enfoques para los metadatos, y coordinarse con herramientas para el procesamiento del lenguaje natural.

Las plataformas de herramientas integradas incluyen muchos tipos de herramientas para la integración de datos, la calidad de datos y la gestión de datos maestros. Las herramientas se integran estrechamente para facilitar la colaboración entre los desarrolladores y para alentar el diseño de soluciones de ID modernas que contengan múltiples funciones altamente diversas.

ACERCA DE NUESTROS PATROCINADORES

www.informatica.com

Informatica es un proveedor de software independiente líder enfocado en brindar innovación transformadora para el futuro de todos los datos de las cosas. Organizaciones alrededor del mundo dependen de Informatica para aprovechar el potencial de su información y satisfacer los principales imperativos del negocio. Más de 5,800 empresas dependen de Informatica para aprovechar sus activos de información que residen en Internet, incluyendo las redes sociales

www.sas.com

SAS Data Management es una solución líder de la industria creada sobre una plataforma integrada común que ayuda a mejorar, integrar y gobernar sus datos. No importa dónde se almacenen sus datos – desde los sistemas legados hasta Hadoop – SAS Data Management ayuda a las organizaciones a tener acceso a los datos que necesitan.

Para las organizaciones que están modernizando sus sistemas de hardware/software legados SAS Data Management es indispensable para integrar y gestionar una variedad de datos de las fuentes estructuradas y no estructuradas. Con ofertas como SAS Data Loader for Hadoop, SAS Event Stream Processing, SAS Data Management, y SAS Data Federation, SAS satisfice los nuevos requerimientos de negocio descritos en este reporte.

ACERCA DEL AUTOR

Philip Russom es director de TDWI Research para la gestión de datos y supervisa muchas publicaciones, servicios y eventos orientados a la investigación de TDWI. Es una figura reconocida en las áreas del almacenamiento de datos y de la inteligencia de negocio, y ha publicado más de 500 reportes de investigación, artículos para revistas, columnas de opinión, discursos, webinars, y más. Antes de integrarse a TDWI en 2005, Russom era analista de la industria que cubría BI en Forrester Research y Giga Information Group. También administra su propio negocio como analista independiente y consultor de BI y fue editor de revistas de TI líderes. Antes de eso, Russom ocupó puestos técnicos y de marketing en varias compañías de bases de datos. Puede contactarlo en prussom@tdwi.org, @prussom en Twitter, y en LinkedIn en linkedin.com/in/philiprussom.

ACERCA DE TDWI RESEARCH

TDWI Research ofrece investigación y asesoría a los profesionales de BI de todo el mundo. TDWI Research se enfoca exclusivamente en asuntos de BI/DW y se asocia con profesionales de la industria para ofrecer un entendimiento amplio y profundo de los problemas de negocio y técnicos alrededor de la implementación de soluciones de inteligencia de negocio y de almacenamiento de datos. TDWI Research ofrece reportes, comentarios y servicios de consulta a través de un programa de Membresía mundial y ofrece investigación a la medida, benchmarking y servicios de planeación estratégica a usuarios y proveedores.

ACERCA DE TDWI CHECKLIST REPORTS

TDWI Checklist Reports ofrecen un panorama de los factores de éxito para un proyecto específico en inteligencia de negocio, almacenamiento de datos o una disciplina de gestión de datos relacionada. Las compañías pueden utilizar este panorama para organizarse antes de iniciar un proyecto o para identificar los objetivos y áreas de mejora de los proyectos actuales.