Haciendo los Repositorios de Datos Utilizables y Útiles

Rogerio DePaula - depaula@colorado.edu
CSI 5817 Sistemas de Base de Datos

"Si los computadores son útiles para todos, no deben serlo mediante la producción de mayor información -ya tenemos suficiente para mantenernos ocupados desde el amanecer al ocaso- sino que ayudándonos a atender la información que consideremos más útil, ya sea por su interés o cualquier otro criterio".

Herbert Simon

Introducción

Es bastante preciso plantear cuán significativa es una tecnología en un momento dado al examinar el número de referencias disponibles. Por otro lado, para que dicha tecnología sea ampliamente adoptada tiene que demostrar su facilidad de uso y utilidad. Después de un entusiasmo inicial, los usuarios a menudo notan problemas más complejos que pueden afectar la utilización de la tecnología y que van más allá de las consideraciones técnicas.

Ciertas tecnologías se ajustan a esta categoría, por ejemplo, los sistemas expertos fueron considerados la solución para todos los problemas empresariales (ignorando los hechos reales). No obstante, después de un entusiasmo inicial, esta tecnología nunca fue adoptada por una audiencia más amplia. En realidad, el sistema experto tiende a prestarse para aplicaciones orientadas al dominio (para lo cual ha sido útil); sin embargo, la integración de estas aplicaciones se volvió muy compleja, previniendo que las organizaciones tomaran ventaja de ellas. La integración entre sus actuales tecnologías de información y otros sistemas basados en el conocimiento demostró ser muy difícil de lograr. Es decir, esta tecnología no ha probado ser útil y utilizable para hacerse cargo de la mayoría de las necesidades de gestión.

Aún así, los repositorios de datos han probado merecer su reputación como una tecnología de bases de datos que integra todo tipo de información, distribuida dentro de las organizaciones cuyas bases de datos usualmente van desde aplicaciones heredadas (base de datos no relacionales) a sistemas de bases de datos relacionales avanzados. De manera significativa, los repositorios de datos prometen entregar un marco de trabajo tecnológico para apoyar procesos de toma de decisiones al entrevar datos informativos.

Sin embargo, las organizaciones que invirtieron en tal tecnología a menudo no reciben la rentabilidad que esperaban [1]. Dos variables complementarias, indirectamente relacionadas con la tecnología de repositorio de datos pero que son inherentes a la naturaleza de los problemas que busca resolver, han demostrado añadir un nivel de complejidad al proceso de hacer que dicha tecnología sea utilizable y útil. Esta complejidad inhibe, por un lado, el acceso de información que los usuarios realmente necesitan y, por otra parte, que los administradores de datos diseñen la aplicación correcta para las necesidades de sus usuarios. Estas dos variables con la calidad de datos y los procesos de gestión.

En este sentido, este estudio señala que además de las consideraciones técnicas (tales como, arquitectura de repositorios de datos, soluciones de data-mart y repositorios de datos centralizados o distribuidos), los gerentes deben tener en consideración estos dos factores mientras se diseñan sistemas de bases de repositorios. Por ejemplo, la calidad de los datos puede ser inapropiada, incomprendida o ignorada y los desarrolladores podrían no tener una comprensión clara del tipo de información que los usuarios quieren adquirir a través del sistema. Esto es, ellos se arriesgan a diseñar la solución técnicamente "correcta" que se encarga de los problemas "equivocados", o peor aún, proveen las respuestas equivocadas para un problema.

Este ensayo define consideraciones técnicas importantes con respecto a la arquitectura de los repositorios de datos para poder dar contexto a la posterior discusión. Luego, presenta un análisis de ambos factores. Finalmente, discute algunos enfoques para hacer de los repositorios de datos útiles y utilizables y presenta la manera en que las bases de datos y el marketing de gestión de información han enfrentado ambos problemas.

¿Por qué un repositorio de datos?

Recientemente, gracias al rápido desarrollo de las tecnologías computacionales, las organizaciones se dieron cuenta de la necesidad de entender sus procesos en profundidad. Al hacerlo, crearon la necesidad de recolectar datos informativos que apoyaran los procesos de planificación, gestión y toma de decisiones. Sin embargo, cuando trataron de extraer información útil de sus bases de datos operacionales, experimentaron los siguientes problemas [2]:

- Un modelo operacional que fue diseñado para transacciones de actualizaciones cortas y predecibles de datos.
- Los sistemas funcionan a través de datos no verificados (blandos) y puntuales (a menudo no almacenados en bases de datos OLTP).
- Tuvieron que operacionalizar sistemas heredados y sistemas prerrelacionales.
- Finalmente, encontraron datos heterogéneos, inhibiendo la rápida integración de datos necesarios para los procesos de toma de decisiones.

Inicialmente, los gerentes intentaron emplear sistemas de gestión de bases de datos relacionales tradicionales. Aún cuando los RBDMS proveen una significativa independencia de datos y un lenguaje de búsqueda, apoyando tanto los procesos operacionales como los de toma de decisiones y contar con la base de OLTP, este enfoque habría sido marginalmente exitoso, por las siguientes razones [2]:

- La complejidad del esquema de búsqueda.
- La necesidad de información más interesante, lo que implica un análisis de datos sumamente complejo.
- El bloqueo de contenido necesario para el apoyo mixto a la toma de decisiones el deterioro del volumen de trabajo de los sistemas de bases de datos.

En este sentido, los depositarios de datos fueron desarrollados para permitir que las organizaciones integren todo tipo de información a través de sus sitios. En resumen, los repositorios de datos contienen datos de síntesis, de historial y de detalle para apoyar las actividades de toma de decisiones. Los datos son extraídos de fuentes operacionales (RDBMS o de

otro tipo), transformados, limpiados, combinados, agregados y resumidos para ser usados por una aplicación de repositorios de datos.

En los repositorios de base de datos hay una importante diferencia entre datos informativos y operacionales. Básicamente, los datos informativos proveen información para las actividades de apoyo a las decisiones, mientras que los datos operacionales son organizados alrededor de operaciones de negocios. Esta diferencia claramente muestra la esencia de los repositorios de datos. Gardner [3, p.54] presenta una definición interesante de repositorios de datos, que apoya el punto de vista de este documento:

> "Los repositorio de datos son un proceso, no un producto, para recopilar y administrar datos de varias fuentes con el propósito de ganar una única visión detallada de parte o todo un negocio".

Los desarrolladores de repositorios de datos deben comprender los procesos de negocio y las necesidades de los usuarios, además de las herramientas de hardware y software. Esencialmente, la información depende en gran medida de las necesidades y objetivos de la empresa, su estructura organizacional y las limitaciones de costos y tiempo. De esta manera, para que un repositorio de datos entregue efectivamente la información correcta para la pregunta apropiada en el momento indicado, debe ser desarrollado tomando en cuenta estos hechos. en relación a esto, este estudio plantea que es fundamental comprender la calidad de datos y los procesos de negocio además de los detalles técnicos del sistema o, de otro modo, esta tecnología puede caer en la categoría de ideas inútiles y poco utilizables.

Consideraciones Iniciales para el Diseño de Sistemas de Repositorios de Datos

Los negocios hoy están sobrecargados con todo tipo de datos, pero tienen poca información disponible. En la era de la información, los datos ya no son un recurso escaso y se ha vuelto un problema porque no sabemos exactamente qué hacer con ella. Además, tales datos son almacenados en sistemas heredados y su calidad a menudo se ve amenazada. De hecho, éstos no tienen valor si no pueden ser convertidos en información [3].

En este sentido, se ha planteado que los repositorios de datos proveen la solución para este problema, no sólo permitiendo a los usuarios **encontrar** las respuestas a sus inquietudes sino, además, **comprender** cómo y por qué son recibidas respuestas específicas [3]. Sin embargo, no puede garantizar tales resultados si la calidad de los datos se ve comprometida y, aún más importante, si el diseño del sistema no responde a las preguntas específicas de los usuarios. "Nada se inclina más a deteriorar el rendimiento y valor de un negocio que una calidad de datos inapropiada, incomprendida e ignorada" [4, p.73].

Se han desarrollado diferentes marcos de trabajo para poder facilitar el proceso de diseño de sistemas de repositorios de datos. Ellos se orientan básicamente a resolver los siguientes problemas [3]:

- **Planificación.** Servicios de descubrimiento de información, los que intentan comprender de mejor manera los procesos de negocio y los problemas que deben resolver.

- **Diseño e implementación.** Cuando se identifica y se comprende la información sobre el negocio, los desarrolladores dan partida al primer proyecto de repositorio de datos.
- **Soporte y mejoramiento.** Comprende una serie de operaciones: apoyo diario a las operaciones en ejecución, asistencia a la expansión del uso de la solución, expansión del sistema para abordar nuevos problemas y, por último, ayudar a mantener el sistema continuamente actualizado y en crecimiento a la vez que apoya mejores decisiones de negocios.

Esta directriz provee una hoja de ruta para que los desarrolladores piensen y desarrollen sistemas de repositorios de datos. Sin embargo, no es suficiente para abordar la complejidad de los casos reales. Los usuarios a menudo no saben exactamente el tipo de información que necesita e incluso si acaso la necesitan de verdad. No hay una mirada clara de la organización como un todo que pueda permitirles crear un sistema que integre de manera efectiva toda esta información y que les facilite conocer la calidad de los datos disponibles. Finalmente, no hay un uso de tecnologías estandarizado que facilite a los diseñadores idear una solución que se ocupe de la mayoría de las necesidades dentro de una organización. A continuación, este documento presenta dos problemas que también deben ser solucionados durante el proceso de diseño de sistemas de repositorios de datos.

Mejorando la Calidad de los Datos

Tal como se dijo antes, la calidad de datos inapropiada, incomprendida o ignorada tiene un efecto negativo en los negocios, sus decisiones y rendimiento, así como el valor de los sistemas de repositorios de datos. En otras palabras, una empresa podría estar en mejores condiciones al no usar tal sistema que usándolo de manera equivocada. Últimamente existe una creciente necesidad de información en el apoyo de los procesos de toma de decisiones y los sistemas de repositorio de datos entregan la infraestructura tecnológica necesaria para adquirirla. Subsecuentemente, dicho sistema debe entregar medios para los desarrolladores así como a los usuarios para que comprendan de mejor manera los procesos de su empresa y cuán efectivamente ellos usan la información entregada por estos sistemas. Para hacerlo, deben evaluar primero la calidad de los datos que están usando.

El problema entonces es la manera en que se mide la calidad de los datos. En primer lugar, los datos informativos son los más apropiados para el proceso de toma de decisiones y, en consecuencia, son los más efectivos para las aplicaciones de repositorios de datos. Luego, los datos pueden ser caracterizados a través de múltiples atribuciones: precisión, exhaustividad, consistencia, temporalidad, credibilidad, valor añadido, interpretabilidad y accesibilidad [4].

Estos atributos pueden entonces estar agrupados en categorías más amplias: intrínseca, contexto, representación y accesibilidad. Por ejemplo, la precisión es un atributo intrínseco; la exhaustividad y la temporalidad son contextuales; la consistencia es representacional; y la disponibilidad es de accesibilidad. Este esquema ayuda a los desarrolladores y usuarios a evaluar la calidad del grupo de datos (es decir, tipos de datos claramente diferenciados). Entonces, deben definir los grupos de datos necesarios para apoyar un esfuerzo de repositorios de datos. Sin embargo, el problema es identificar estos grupos de datos en el contexto de una organización. En este sentido, los atributos guían esta determinación. Por ejemplo, si los datos no existen, su calidad es deficiente en el atributo de disponibilidad.

Además, Ballou y Tayi [4] presentan diferentes maneras en que la calidad de datos puede ser mejorada. Por ejemplo, se pueden solucionar las diferencias a través de formatos de datos. Otro recurso puede ser obtener datos en momentos más precisos. Otro podría ser identificar y aplicar una definición común de conceptos clave en la organización (por ejemplo, "ventas", en lo que concierne a los repositorios de datos). Otro podría ser obtener datos externos. Cada uno ejerce una influencia sobre la calidad del grupo de datos, el cual influye, a su vez, en la calidad de datos disponibles para las decisiones de negocios.

Ser un Experto en el Negocio - Creando Soluciones Efectivas de Repositorio de Datos

Entender la calidad de datos es entender el contexto en el que se usará la información obtenida de tales datos. Entonces, tanto los usuarios como los desarrolladores (las personas involucradas en el esfuerzo de repositorio de datos) deben tener una visión clara de los contextos de negocios que requerirán el uso de esta información. En este sentido, ellos deben pensar sistemáticamente sobre lo que quieren obtener con la implementación de un sistema de repositorio de datos. Para lograrlo, deben adquirir la experticia necesaria con respecto a su empresa. Además, cada grupo debe entender las necesidades del otro.

Los administradores de datos (desarrolladores) saben lo que ocurre detrás de un proyecto de base de datos - la optimización SQL, modelo de datos, extracción de legado e integridad de datos, diseño de esquema, índices y rendimiento de carga. Por otro lado, los usuarios conocen las necesidades del negocio y qué tipo de información es mayoritariamente relevante para ciertos procesos de toma de decisiones, aunque pueden ser un repositorio de datos simplemente como un esquema y herramientas entregadas por administradores de datos que les permiten buscar información, crear reportes y analizar datos. Entonces, ambos grupos deben, al menos, reconocer uno al otro para diseñar de manera realista un sistema de repositorio de datos que se ocupe de las necesidades reales de los usuarios y, por otro lado, imponga los requisitos del sistema.

A propósito de esto, Classey [1] delinea los criterios clave que apuntan a apoyar un diseño centrado en el usuario:

- Organizar y estructurar una base de datos más legible para los usuarios al usar terminología empresarial comúnmente utilizada.
- Limpiar los datos en el repositorio en vez de usar complejas herramientas de reglas de transformación por el lado de los clientes.
- Almacenar y reutilizar los metadatos en el almacén de datos.
- Buscar RDBMS optimizados que funcionen como la piedra angular del almacén de datos; el rendimiento, la eficiencia de respuestas y la escalabilidad son esenciales.
- Aplicar diferentes estrategias que hagan la información disponible para los usuarios, por ejemplo, el uso de repositorios de datos en la Web.

Por otro lado, los desarrolladores necesitan tener una imagen clara del tipo de información que entregarán a los usuarios. Con el objetivo de proveer herramientas que permitan a los usuarios obtener la información correcta para apoyar los procesos de toma de decisiones, los desarrolladores deben comprender los procesos y estructuras empresariales. Esto es, ellos deben ser capaces de ver la empresa como un todo para poder discernir los datos a ser usador y, del mismo modo, el tipo de arquitectura de repositorio de datos que deben emplear para apoyar a la

organización de mejor manera tanto hoy como en el futuro. Un aspecto importante en los procesos de toma de decisiones es que las decisiones hoy en día probablemente serán distintas a las que se tomen mañana. Por ello, los desarrolladores también deben ser capaces de anticiparse, en alguna medida, a la dirección en la que se dirige la organización.

Además, existen otras inquietudes que desarrolladores y usuarios deben responder de manera conjunta al momento de diseñar un sistema de repositorio de datos. Hay numerosos proveedores y cada uno de ellos ofrece un tipo específico de solución que puede o no ser la más apropiada para el proyecto, por lo que tanto desarrolladores como usuarios deben considerar cuidadosamente las siguientes categorías [3]:

- Costos: ¿Cuánto puede y debe gastar la empresa en el proyecto?
- Tiempo: ¿Cuánto tiempo tomará?
- Usuarios: ¿Cuáles son los perfiles de los usuarios finales que tomarán ventaja del sistema?
- Mantenimiento: ¿Quién construirá el sistema? ¿Quiénes serán los que lo mantendrán eventualmente?
- Hardware, software y herramientas. ¿Con qué herramientas cuenta la empresa hoy? ¿Cómo se utilizan?
- Servicios. ¿Cómo puede este sistema mejorar los procesos de trabajo de la empresa?

En resumen, tanto los usuarios como desarrolladores deben ser expertos no sólo en su propio campo de trabajo, sino que deben ser capaces de extender esta experticia hacia las especialidades del otro grupo. Al hacerlo, serán capaces de mantener los más altos beneficios que los repositorios de datos pueden entregar, ya que podrán equilibrar sus necesidades con las ventajas y desventajas tecnológicas del sistema.

Haciendo los Repositorios de Datos Utilizables y Útiles

Los diferentes proveedores de soluciones de repositorios de datos han reconocido los problemas presentados en este artículo. Ellos han integrado consultorías para poder ayudar a sus clientes a comprender de mejor manera sus necesidades reales y, asimismo, obtener los más altos beneficios de estas tecnologías. Además, han comprendido la importancia de entregar sistemas fáciles de usar en la medida que se han percatado del amplio rango de usuarios que eventualmente manipularían datos de estos sistemas para poder obtener la información que necesitan. Al contrario de los anteriores sistemas de base de datos, donde un administrador de datos entregaba los reportes a los consumidores de información, en los casos de repositorios de datos, los consumidores son capaces de manipular los datos directamente con el fin de analizarlos y, en consecuencia, obtener la información necesaria.

Con el fin de mejorar el uso de sistemas de repositorios de datos, IBM desarrolló DataGuide, una herramienta orientada al usuario que facilita el proceso de reunir información antigua de metadatos de base de datos de empresas. IBM desarrolló una herramienta de almacenaje visual que apunta a facilitar tareas comúnmente complejas, tales como mapeo de datos, extracción de fuentes heterogéneas de datos, programación de agenda de procesos y monitoreo de operaciones de almacenaje. Por otro lado, formaron sociedades con otras corporaciones para poder expandir las funcionalidades de su sistema de repositorio de datos y los servicios ofrecidos. Por ejemplo, se unieron a Evolutionary Technologies Internacional of Austin, cuya herramienta Extract genera

programas de 3GL para transformar, consolidar y extraer datos de prácticamente cualquier fuente de datos. También formaron una sociedad con Vality Technology of Boston, cuyo programa de ambiente de reingeniería de datos, Integrity, realiza limpieza de datos [2].

Aún cuando Corporación Oracle se ha atraso en la entrega de una nueva herramienta de repositorio de datos, lo han hecho con el propósito de mejorar el acceso de los clientes a su solución de almacenaje datos. Ellos están desarrollando Warehouse Builder, un complejo paquete para extraer, transformar y cargar datos de sistemas heredados, aplicaciones comprimidas, entre otras fuentes, para formar un repositorio de datos que sirva para el análisis. Su problema radica en mejorar el sistema que maneja los metadatos en entornos distribuidos. Al hacerlo, ellos buscan facilitar la extracción de datos de cualquier tipo de fuente de datos [5].

Por otra parte, NCR, el reconocido líder en las soluciones de repositorios de datos, anunciaron una sociedad con ETI Extract, quienes proveen una solución completa para la administración de integración de datos al automatizar el proceso de consolidar datos entre sistemas incompatibles, ayudando de esta manera a que los clientes completen sus proyectos de manera más rápida y eficiente desde el punto de vista de los costos. Se trata de un esfuerzo en conjunto para ofrecer no sólo el desarrollo e implementación de soluciones técnicas, sino también soluciones de marketing y consultoría. Ellos dicen que serán capaces de proveer soluciones sólidas y un soporte al cliente superior [6].

Finalmente, Brio Technology se ocupa del problema de las interfaces orientadas a los usuarios. Esto es, ellos buscan reducir la sobrecarga de trabajo de los departamentos de informática al proveer herramientas que facilitan la interacción de los usuarios finales con los sistemas de repositorios de datos. Para cumplir este objetivo, redujeron la complejidad del soporte del usuario final y permiten la compatibilidad con un amplio rango de plataformas como Unix, Macintosh y Windows. Ellos integraron de manera estrecha el proceso de recolección de información con las herramientas de repositorio de datos. Al hacerlo, bajaron los costos de los departamentos de informática al permitir que los usuarios finales construyan informes, determinen el acceso, realicen agendas, impriman y envíen reportes al mismo tiempo que automatizan la instalación y el mantenimiento de la herramienta [1].

Conclusión

> “La información consume la atención humana, así que una riqueza de información crea una pobreza de atención humana. Los enfoques de diseño adecuados para un mundo donde la información es el recurso escaso son exactamente los equivocados para un mundo en donde el recurso escaso es la atención humana”.

Esta última cita muestra que vivimos en un mundo donde los datos se han vuelto un problema simplemente porque no sabemos exactamente qué hacer con ellos. Además, los viejos paradigmas tecnológicos probablemente no funcionan en esta realidad actual. En consecuencia, como desarrolladores de software, debemos ser capaces de cambiar nuestro previamente aceptado tecnocentrismo por un enfoque más orientado al usuario al momento de desarrollar soluciones técnicas. Esto es, debemos percibir que las viejas grandes soluciones pueden no ser las adecuadas para la realidad actual. En relación a esto, los esfuerzos de repositorios de datos

pueden no tener éxito por varias razones, pero nada asegura un fracaso más que la falta de preocupación por la calidad de los datos y la negligencia por la compresión de las necesidades de los usuarios.

Finalmente, los repositorios de datos prometen entregar la solución correcta para las actuales necesidades empresariales por una información de alta calidad en el formato correcto en el momento adecuado. Para hacerlo, tanto los desarrolladores como los usuarios deben ser cuidadosos y no sobreestimar los valores de la tecnología, sin tener en cuenta la complejidad del problema. Esto es, que la herramienta debe ser útil y utilizable. Al reconocer cuidadosamente este hecho, esta tecnología tiene el potencial de recompensar toda la inversión y esfuerzo que requiere, sin arriesgar convertirse en otra gran idea que simplemente no pudo prosperar.

Referencias

[1] K. Glassey, “Seduciendo al Usuario Final,” Communications of the ACM, vol. 41, pp. 62-69, 1998.

[2] C. Bontempo and G. Zagelow, “La Arquitectura de Repositorio de Datos de IBM,” Communications of the ACM, vol. 41, pp. 38-48, 1998.

[3] S. R. Gardner, “Construyendo del Repositorio de Datos,” Communications of the ACM, vol. 41, pp. 52-60, 1998.

[4] D. P. Ballou and G. K. Tayi, “Mejorando la Calidad de Datos en los Ambientes de Repositorio de Datos,” Communications of the ACM, vol. 42, pp. 73-78, 1999.

[5] M. Hammond, “Oracle Atrasa Entrega de Herramienta de Repositorio de Datos,”, 1999. http://www.zdnet.com/pcweek/stories/news/0,4153,396792,00.html

[6] J. E. Brawner, “ETI y NCR Anuncia Esfuerzo Conjunto Para Mejorar las Soluciones de Repositorio de Datos y Fortalecer el Soporte a los Clientes”, vol. 1999: ZDNew, 1999. http://www.zdnet.com/products/stories/reviews/0,4161,1002986,00.html