

Almacenamiento inteligente: acceso rápido y preciso

UN CONCEPTO PUNTERO EN BASES DE DATOS PARA OPERACIONES Y ALMACENAMIENTO QUE
AYUDA A LAS EMPRESAS A MANEJAR EFICIENTEMENTE INGENTES VOLÚMENES DE DATOS.

A medida que aumentan los volúmenes de datos, las empresas se enfrentan a nuevos retos en sus operaciones de TI, en sus análisis de inteligencia empresarial (IE) y en otras áreas que dependen del almacenamiento de datos. Las soluciones tecnológicas deben adaptarse a las necesidades específicas de estas áreas de negocio.

El extraordinario crecimiento de los volúmenes de datos constituye un desafío crítico para las empresas que siempre han tenido que manejar un gran capital de datos. Según un estudio realizado el año pasado por la empresa de investigación de mercados IDC, las grandes bases de datos son especialmente proclives a experimentar un rápido crecimiento. Setenta y dos de las empresas que participaron, cuyas bases de datos ocupaban más de 1,5 terabytes, declararon unas tasas del 20% en cuanto a crecimiento anual de los datos. Es más, para el 45% de estas empresas el incremento a soportar era superior al 40%.

CAUSAS DE LA INFLACIÓN DE DATOS

Las causas de esta rápida inflación de datos están relacionadas con los volúmenes de operaciones y con el almacenamiento de los datos. Por otra parte, sigue aumentando el número de aplicaciones operativas que tienen que manejar un número de registros que se multiplica a gran velocidad. Cientos de empresas se ven afectadas por dicha inflación, como las empresas de telecomunicaciones, que deben conservar ciertos datos sobre cada llamada telefónica para almacenarlos en tablas de registro de detalles de llamadas (Call Detail Records, CDR) que fácilmente pueden contener varios miles de millones de registros; o el comercio minorista, donde existen aplicaciones de planificación de recursos empresariales (Enterprise Resource Planning, ERP), gestión de almacenes, logística o cadenas de suministro

que utilizan ingentes cantidades de datos sobre establecimientos, ventas, inventarios y previsiones.

En cuanto al almacenamiento, las tasas de crecimiento de los datos pueden ser aún más sorprendentes. Por lo general, los almacenes de datos guardan todos los datos históricos necesarios para efectuar profundos análisis de inteligencia empresarial (IE). Datos que se enriquecen con frecuencia con fuentes de datos de terceros. En la negociación de valores basada en algoritmos, además de datos en tiempo real, son necesarios series de datos históricos con el fin de llevar a cabo análisis combinados en fracciones de segundo.

Hoy en día, los análisis tradicionales, estratégicos y tácticos de IE, basados primordialmente en datos históricos, se combinan cada vez más con la IE operativa, que observa la evolución casi en tiempo real para poder realizar una intervención inmediata en caso necesario. Sin embargo, esto introduce grandes volúmenes adicionales de datos operativos en la base de datos analítica. Como consecuencia, el volumen de datos que gestiona un almacén de datos medio que crece más de un 125% al año. En esta línea, los volúmenes de datos de terabytes son ya algo habitual para muchos sectores.

Esta problemática se agrava con la calidad y la legislación vigente. La puesta en común de ciertos tipos de datos desestructurados dentro del conjunto total aumenta a medida que se añaden documentos en muy diversos formatos (gráficos, diagramas, grabaciones de vídeo digital o de audio, páginas html y otros). En particular, el elevado número de documentos adjuntos de correo electrónico está dando lugar a que estos datos crezcan de forma exponencial, y hay que tener en cuenta que, en España, ciertos sectores empresariales están obligados por la legislación comercial y fiscal a guardar la correspondencia electrónica

ALMACENAMIENTO DE DATOS POR COLUMNAS

El concepto básico de que las particiones reducen la cantidad de datos que hay que leer para una operación dada puede aplicarse también a los almacenes de datos. Sin embargo, dado que los análisis de IE, al igual que los de series temporales, suelen exigir la exploración de todos los datos para detectar interdependencias, es preferible adoptar otro planteamiento técnico para obtener mejores resultados. Es recomendable recurrir a una arquitectura que difiera esencialmente de las bases de datos de OLTP: el almacenamiento y acceso a los datos por columnas en lugar de por filas. Puesto que las consultas generalmente se relacionan con valores por columnas, este método presenta dos ventajas esenciales:

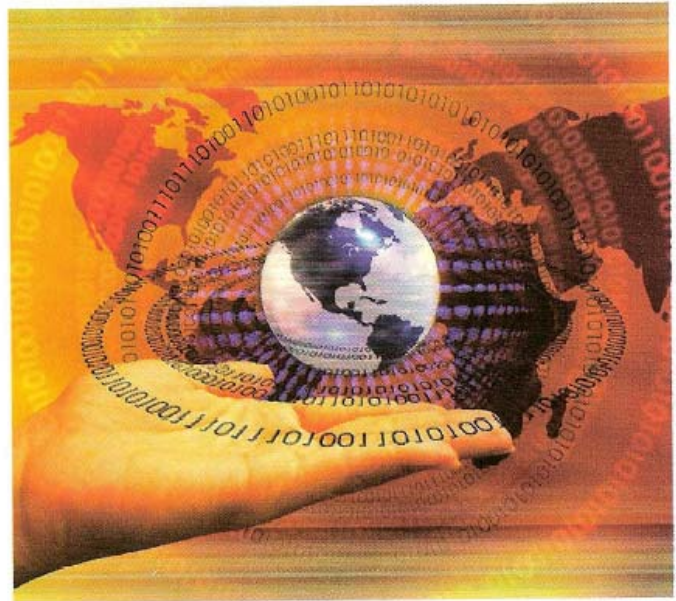
- Las operaciones de lectura se limitan a los datos que son realmente necesarios, con lo que se reduce el número de operaciones de E/S. Al mismo tiempo, los datos se pueden comprimir de manera más eficiente, porque todos los datos de una columna son del mismo tipo, mientras que los de las filas suelen ser heterogéneos. Por tanto, el algoritmo de compresión debe basarse en el mínimo común denominador.
- Una base de datos puede indexarse por completo, ya que cada campo sirve automáticamente como clave de búsqueda. Esto proporciona al usuario la máxima flexibilidad. Los archivos de índice adicionales que pueden inflar la base de datos son superfluos.

relevante hasta diez años, de manera que se pueda acceder a ella en cualquier momento dentro de un breve plazo razonable.

LAS GRANDES BASES DE DATOS Y SUS CONSECUENCIAS

En las operaciones diarias, el funcionamiento de las aplicaciones se ve afectado por el creciente volumen de datos y la complejidad cada vez mayor de datos e índices. Para cada operación o consulta individual es necesario recorrer todos los datos de una tabla. Operaciones simultáneas con diferentes clases de datos dentro de una misma tabla pueden entrar en conflicto con otras operaciones en la base de datos, originando un bloqueo de páginas u otros conflictos de acceso. Esto dará lugar a una ralentización de los procesos o a su total paralización.

Otras áreas afectadas incluyen la gestión y el mantenimiento de las bases de datos en el curso de operaciones de rutina. Las comprobaciones, como los análisis de errores de datos, así como las grandes tablas de índices o las actualizaciones de estadísticas, requieren mucho más tiempo. El tiempo de que se dispone para la realización de copias de seguridad de los datos (limitado en muchas compañías a tan sólo unas horas durante la noche) suele ser demasiado breve. Además, puede verse afectada la disponibilidad de los datos.



Por lo que respecta al almacenamiento, la inflación de datos tiende a causar graves problemas cuando la tecnología que emplean está orientada de forma exclusiva hacia el procesamiento de operaciones. Un sistema así requiere espacio de almacenamiento, no sólo para los datos sin procesar, sino también para índices y tareas administrativas.

Una segunda consecuencia son los prolongados tiempos de búsqueda. En las bases de datos orientadas a operaciones, el método tradicional de almacenamiento de los datos por filas exige leer íntegramente cada registro. Las empresas que tratan de paliar el problema preagregando los datos que se van a analizar, acaban limitando considerablemente la flexibilidad de las consultas ad hoc. Por otro lado, el almacenamiento y el procesamiento de gigantescos volúmenes de datos resulta cada vez más exigente en cuanto a potencia de procesamiento y funcionamiento de los discos duros y del sistema de refrigeración, incrementando el consumo de energía y por tanto las emisiones de CO₂. Un contexto dentro del cual se ha abierto un debate a cerca de la eco-eficiencia.

SOLUCIONES

Muchas empresas se deciden por el planteamiento más directo (y costoso) para hacer frente a esta avalancha de datos adquiriendo hardware en forma de más memoria, procesadores más rápidos y espacio en disco adicional, lo que no elimina la verdadera causa del problema de rendimiento. La raíz de la cuestión sólo puede abordarse mejorando significativamente la tecnología de base de datos, para lo que el principio clave debe ser que el acceso se limite a los datos que realmente sean necesarios.



Vicente Moncho Brunengo es director de marketing de la Región Sur de Sybase