SERIE: SISTEMAS DE ALMACENAMIENTO DISTRIBUIDO (IV)

Los SAD puestos en perspectiva: evolución y estado actual[1]

Ricardo Marcelín Jiménez

Noviembre de 2011

Basta con poner atención a nuestro entorno para darnos cuenta de que el manejo de la información, en la actualidad, es un problema significativo, en tanto bien colectivo e intangible. Conforme pasa el tiempo, mayor número de organizaciones elevan sus necesidades de manejo de información y, por lo tanto, de almacenamiento. Estas organizaciones y estas necesidades seguirán creciendo con el tiempo y no parece que haya razón para pensar que se detendrán en algún momento cercano.

Bajo este panorama, no podemos esperar que los sistemas de almacenamiento usados hasta hoy tengan un alto rendimiento, pues su operación puede resultar muy costosa, su capacidad de crecimiento limitada, su seguridad vulnerada y su sistema volverse susceptible a fallas. En la actualidad la información ya no se encuentra en un solo lugar, sino que se halla distribuida; de ahí la importancia de considerar los mecanismo de distribución y de recuperación,

[1]Este artículo fue redactado por Fernando Barajas con base en la investigación *Hacia los sistemas de almacenamiento distribuido de información, orientados por la naturaleza de los contenidos*, cuyo responsable es el Dr. Ricardo Marcelín Jiménez, quien colabora en proyectos de investigación aplicada del Fondo de Información y Documentación para la Industria INFOTEC.

así como el almacenamiento de volúmenes de información que se multiplicarán con el tiempo para prevenir que suceda un desastre.

De esta manera, conforme pasa el tiempo y la información se multiplica, el desempeño de los sistemas de almacenamiento es cada vez más exigido, sobre todo en dos aspectos. Por un lado, existe una creciente necesidad de compartir información entre organizaciones en relaciones particulares, donde cada miembro tenga acceso a determinado volumen de información. Por otro, se requiere un movimiento intenso en los datos que les permita ser recuperados de acuerdo con parámetros específicos, o en otras palabras, se necesita que entre una gran cantidad de información llegue la correcta a la persona correcta de acuerdo con sus necesidades específicas y sin importar que ignore la existencia o la naturaleza de dicha información.

La primera de estas necesidades demanda mecanismos de interoperabilidad entre sistemas. Esto es, en lugar de pedir que todos los sistemas en los que se ha invertido sean tirados a la basura en nombre de un único sistema, se trata de generar "traductores" que optimicen su funcionamiento sin la necesidad de perder inversiones pasadas. La segunda, por su parte, supone la anotación semántica de los contenidos. De tal forma, la información se moverá de acuerdo con las necesidades particulares de los usuarios, quienes podrán recuperarla en arreglo a determinados parámetros de búsqueda.

Toda propuesta de los Sistemas de Almacenamiento Distribuido (SAD) debe tomar decisiones al menos en tres aspectos imprescindibles: 1. necesidades del usuario y principios de diseño del sistema, 2. parámetros de desempeño que

garanticen la calidad y 3. entidad básica de almacenamiento. En este último aspecto es importante insistir en que no parece conveniente recargarse demasiado en dispositivos individuales, muy proclives a fallas, sino que es mejor generar una memoria virtual que tenga un mayor tiempo de vida útil.

Por su parte, en términos de diseño, puede ser sumamente positivo que un SAD responda cuatro preguntas: 1. ¿Cómo se identifica el lugar donde se encuentra la información?, 2. ¿Cómo se solicita esta información?, 3. ¿Cómo se intercambia la información entre repositorios?, y 4. ¿Cómo se preserva la consistencia de la información? La respuesta a cada una de estas interrogantes debe ser considerada a la luz de las necesidades de almacenamiento, recuperación, seguridad y naturaleza de la información.

Historia de los SAD:

Desde que el hardware de cada computadora comprende capacidades de comunicación, los Sistemas de Almacenamiento Distribuido han evolucionado en distintos aspectos. A continuación, de manera sumaria, presentamos algunos ejemplos:

- DAS (*Direct Attached Storage*, lo que se puede traducir como sistema de almacenamiento conectado directamente): el sistema de almacenamiento se encuentra conectado a la computadora y reservado para su uso exclusivo. El acceso se obtiene mediante el sistema de archivos que la computadora tenga habilitado.

- SAN (*Storage Area Network*, o almacenamiento de redes locales): el usuario tiene acceso remoto a la información como si estuviera almacenada localmente, además de contar con la funcionalidad de un sistema de archivos. Se utiliza en redes locales de conexión.
- NAS (*Network Attached Storage* o almacenamiento conectado a una red): los usuarios se encuentran conectados por una red de comunicación y pueden leer y escribir archivos como si se tratara de operaciones locales. La diferencia con los SAN es que éstos no ofrecen un sistema de archivos.
- iSCSI (*Internet SCSI*): los usuarios acceden al sistema de almacenamiento por medio de Internet.

Estos sistemas y su patente evolución son los antecedentes directos de los SAD, pero es evidente que conforme la información se multiplique se volverán inoperantes o difíciles de costear, al tiempo que sus funciones se encontrarán sensiblemente disminuidas.

Sistemas destacados de almacenamiento distribuido

Existen muy diversas maneras de pensar el almacenamiento distribuido, todo depende de las necesidades y los volúmenes de información que se quieran atender. Los ejemplos que presentamos a continuación pueden representar pistas tanto para entender los sistemas distribuidos de almacenamiento (e identificar los momentos en los que los hemos utilizado), como para pensar en nuevas posibilidades:

- *Napster*: popular sistema *P2P* (*peer to peer* o entre pares) para compartir archivos MP3. Cada usuario contribuía con un espacio de almacenamiento y sus propios recursos de red. El indexado y la búsqueda se realizaban en un sitio principal, mientras que las descargas iban directamente de usuario a usuario.
- *Gnutella*: una consulta difunde una cadena de datos que cada receptor puede usar como le parezca, de acuerdo con el nombre del archivo, el contenido de un documento, etc. Las transferencias también se efectúan de usuario a usuario.
- *Farsite*: sistema que destaca por su confidencialidad. El acceso a archivos se efectúa mediante firma electrónica y todos los archivos se encriptan. Un directorio distribuido registra los sitios en donde se guardan los originales.
- *OceanStore*: sistema a escala global que presenta alto desempeño y disponibilidad. Aun bajo condiciones fuertes de trabajo, es resistente a fallas, ataques o cambio de condiciones en la red de comunicaciones.
- *BigTable*: sistema construido por encargo que almacena datos estructurados. El cliente puede desarrollar nuevas aplicaciones.
- *Dynamo*: también es un sistema construido por encargo. En él los clientes pueden programar su propio nivel de calidad de servicios.
- *Ceph*: sistema de almacenamiento de código abierto que ofrece capacidades masivas de almacenamiento en el orden de los petabytes.
- *Keyspace*: sistema de código libre diseñado para satisfacer tres requerimientos: fuerte consistencia, resistencia a fallas y alta disponibilidad.

Sistemas P2P

Dentro de la evolución de los SAD debemos destacar la aparición de sistemas *P2P*. En resumen, en la historia de los SAD se pueden identificar dos momentos clave: el primero fue cuando se migró a un sistema centralizado que hacía distribuciones estáticas, y el segundo cuando los dispositivos cobraron movilidad y ya no era posible sostener una estructura cliente/servidor. Bajo esta última situación se crearon los sistemas *P2P*, nodos interconectados capaces de auto-organizarse y que se usan para compartir recursos, tales como ciclos de CPU, almacenamiento y ancho de banda. El aspecto positivo de estos sistemas es que se encuentran descentralizados, además de que cada participante que se adhiere aporta más recursos y no representa ninguna carga adicional. El reto es evitar que cada usuario que se ausente ponga en riesgo la seguridad y la confiabilidad del sistema.

Para responder a las búsquedas que se realicen en los *P2P* se pueden implementar dos tipos de sistemas: una organización por llaves (como un mapa semántico) que distribuye los recursos y dirige las búsquedas, u otra que provoca una “inundación” de los datos con el nombre del recurso que se busca. Al mismo tiempo, estos sistemas pueden ser anotados semánticamente, con lo que las búsquedas serían más completas y los resultados más enriquecedores.

Aunque no siempre lo supimos, a lo largo del tiempo hemos trabajado con diferentes sistemas de almacenamiento que nos brindan determinados servicios.

Es importante ser conscientes de esa historia para conocer el estado actual de la práctica de SAD. En estos días, en los que a cada momento la importancia de la información aumenta, conocer los sistemas en los que se administra nos brinda mejores perspectivas y nos ayuda a pensar en nuevas posibilidades.

Si te interesó el artículo, también puedes consultar:

- Investigación *"Hacia los sistemas de almacenamiento distribuido de información, orientados por la naturaleza de los contenidos"*
- Artículos de Divulgación INFOTEC
- Proyectos de Investigación Aplicada en INFOTEC

Ricardo Marcelín Jiménez

r.marcelin.jimenez@gmail.com

Doctor en Ciencias Computacionales por la Universidad Autónoma Metropolitana Unidad Iztapalapa. Miembro del Sistema Nacional de Investigadores Nivel I, otorgado por el CONACYT. Profesor del Área de Redes y Telecomunicaciones, del Departamento de Ingeniería Eléctrica de la Universidad Autónoma Metropolitana, Unidad Iztapalapa. Como investigador, sus intereses son: el almacenamiento distribuido, las redes inalámbricas de sensores y la simulación de eventos discretos. Actualmente, entre otras actividades, colabora en proyectos de investigación en INFOTEC, dirigiendo el proyecto *"Hacia los sistemas de almacenamiento distribuido de información, orientados por la naturaleza de los contenidos"*.