SERIE: SISTEMAS DE ALMACENAMIENTO DISTRIBUIDO (II)

Grandes almacenes SAD: un proyecto a futuro[1]

Ricardo Marcelín Jiménez

Julio de 2011

Un Sistema de Almacenamiento Distribuido (SAD), básicamente, consiste en un procedimiento que distribuye la información en diversos discos que se encuentran conectados a la red. A partir de esta generalidad, la definición de lo que es un SAD tendría que tomar en cuenta cada caso en particular, pues un sistema de esta naturaleza es capaz de adaptarse a las necesidades de cada organización que echa mano de él. De esta manera, podemos plantear que hay diversos tipos de SAD en la misma medida en la que existen diferentes necesidades de almacenamiento. No podemos afirmar que estos sistemas son precisamente novedosos, pues cuentan ya con una historia y con una evolución específicas que se relacionan con las demandas del creciente mundo de la información. En este artículo exploraremos un proyecto específico de SAD que busca ajustarse a las necesidades actuales y que, además, tiene la capacidad de crecer junto con ellas.

Para decirlo en pocas palabras, este proyecto busca aportar soluciones a las necesidades de almacenamiento actuales. De ahí que se requiera una

[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 Ricardo Marcelín Jiménez con la asistencia de Reina Carolina Medina Ramírez y Diego R. Guzmán Santamaría. Los autores colaboran en proyectos de investigación aplicada del Fondo de Información y Documentación para la Industria INFOTEC.

arquitectura (construcción informática) que permita un orden adecuado en los datos y una accesibilidad óptima para los usuarios. Una buena organización supone múltiples beneficios. Si en un almacén los productos se encuentra bien jerarquizados y el orden corresponde a la naturaleza de los mismos, es más fácil encontrarlos, e incluso, aprender de ellos. Este proyecto, pues, apunta al mejor aprovechamiento de nuestros recursos de información y plantea una nueva relación con ellos, lo cual implica mejores prácticas y mayor beneficio.

Antes de pensar en cualquier tipo de propuesta debiéramos preguntarnos, ¿qué necesita un SAD para ser realmente funcional? Un SAD óptimamente diseñado toma en cuenta dos tipos de requerimientos: los *funcionales* y los *no funcionales*. Pensemos en el ejemplo del almacén; los *requerimientos funcionales* se refieren a las necesidades de los clientes o de los que acuden a él, y los *requerimientos no funcionales*, a la forma de solventar estas necesidades o, siguiendo el ejemplo, lo que debe hacer la gerencia para mantener contentos a sus clientes. En suma, los primeros responden a la pregunta: ¿qué se necesita?, y los segundos: ¿cómo resuelvo las necesidades?

Entre los requerimientos funcionales están, en primer lugar, *el monitoreo y control*, que buscan regular los parámetros de funcionamiento (redundancia, balance y disponibilidad), así como invocar los mecanismos de seguridad y reparación. *Los metadatos*, por su parte, son la creación de mapas de acceso a la información. Una base de metadatos centralizada es eficiente al plantear las rutas de acceso, pero proclive a quedar saturada (como un cuello de botella en el

tránsito). A su vez, una descentralizada garantiza la velocidad de acceso, pero diversifica y dificulta la creación de rutas. *La consistencia y la sincronización*, por su parte, engloban otro requerimiento funcional que debe tomarse en cuenta. Se trata de una característica que pretende regular las veces que un archivo puede ser abierto o escrito al mismo tiempo. Si existen varias copias de un documento, éstas deben coincidir con las modificaciones de escritura que se le hacen. De ahí que deba tomarse en cuenta una falla común: hay un momento en el que se hallan dos copias diferentes de un documento: la original y la que se está modificando.

A lo largo de sus viajes por el sistema distribuido, un archivo puede ir desgastándose y perder calidad, o bien, ser interceptado por usuarios no autorizados. *La integridad y seguridad*, por ello, constituyen un requerimiento básico. Para mantener un archivo intacto, la estrategia de redundancia es la más común, pero en dado caso pueden aumentar severamente los costos. Se puede solucionar este problema si se genera un sistema que revise la frecuencia de las fallas en cada archivo y produzca copias de acuerdo a ello. Si hablamos de seguridad, es necesario codificar los archivos para evitar su interrupción y crear estrategias para autorizar el acceso sólo a determinados usuarios. *La indexación y búsqueda* plantean la necesidad de que cada archivo tenga un código de localización único y un indicador que señale qué usuarios pueden acceder a él. Se necesita, pues, un sistema ágil y distribuido para emplazar y localizar archivos.

Al hablar de *los requerimientos no funcionales*, es decir, de los principios de diseño que buscan soluciones para los requerimientos funcionales, no podemos

dejar de lado *la modularidad* o diseño de sistemas por separado que se interconectan para poner en marcha el SAD. Una modularidad positiva está preparada para los cambios, incluso radicales, sin afectar la interfaz. La *interoperabilidad*, por su parte, provee un seguro intercambio de datos entre distintas fuentes. Para lograrla, es indispensable generar estándares que traduzcan eficientemente los archivos para hacerlos fácilmente accesibles. En lo que se refiere a *la escalabilidad*, es indispensable pensar en el crecimiento de las redes. No basta con planear un SAD funcional, es necesario tomar en cuenta que, con el tiempo, éste tendrá más archivos y requerirá un sistema de distribución, una política de intercambio y un sistema de localización preparados para el crecimiento. Finalmente, *la confiabilidad y la tolerancia a fallas* prometen un sistema que funcionará a largo plazo y que estará preparado para afrontar los problemas, ya que, entre más grande es un SAD, más proclive se encuentra a los errores.

En suma, al planear un SAD es necesario no sólo considerar los sistemas que estarán interconectados, sino la manera en que se conectarán. La siguiente propuesta se enfoca en redes pequeñas o medianas capaces de articularse entres sí para lograr una mayor complejidad. En primer término, se descarta la opción de almacenar en componentes individuales de la red (los discos duros de las computadoras, por ejemplo). Con el fin de garantizar una vida larga e independizarse de los manejos individuales, se propone crear una memoria virtual.

Por otra parte, se proyecta una construcción celular, es decir, células como unidades mínimas de interconexión que, a su vez, se enlazan a un conjunto de servidores web. Una estructura de este tipo brinda la posibilidad de tener una interfaz única con capacidad extendida y altos parámetros de funcionamiento. En otras palabras, podemos hablar de una serie de sistemas P2P (*peer to peer*, o entre pares) conectados a un servidor, el cual administra la información, delega actividades a los participantes (como la codificación, por poner un ejemplo) y a su vez funciona él mismo como dispositivo de almacenamiento. Una red pequeña funciona muy bien con un sistema P2P; en este caso, al conectar esos sistemas a un servidor, se posibilita una mayor articulación, y por lo tanto, un SAD más grande y con capacidad de crecer aún más. Bajo esta perspectiva de diseño, se favorecen *la modularidad*, *la interoperabilidad*, *la escalabilidad* y *la confiabilidad*, al tiempo que se brindan mejores servicios gracias a los mecanismos de indexado (etiquetas a la información) de cada célula.

De acuerdo con lo anterior, el proyecto de sistema permitiría dos tipos de almacenamiento: uno local y otro remoto, cada uno con distintos niveles de redundancia. De esta forma, el primero funcionaría como una memoria caché mientras que el segundo brindaría un servicio fuera de sitio. La idea es que ambas funcionen, en conjunto, con una estructura de células interconectadas con servidores. En resumidas cuentas, cada participante ahorra en costos gracias al almacenamiento local funcionando como caché que permite un rápido acceso a

información habitual, al tiempo que tiene acceso a información remota pero menos recurrente.

Finalmente, los siguientes pasos a seguir, una vez identificadas las necesidades y la manera de solventarlas, y después de que se ha planteado una propuesta específica, son: 1. desarrollar metodologías y herramientas para caracterizar la naturaleza de la información (anotarla semánticamente); 2. identificar las relaciones entre los requerimientos *funcionales* y los *no funcionales*, así como diseñar métodos de medición para ellos; y 3. seleccionar o desarrollar una metodología general para el diseño y construcción del SAD, tomando en cuenta todo lo anterior. Por ello, si algo tenemos que aprender de la historia de los SAD es que sus necesidades cambian y crecen. De ahí que sea indispensable tratar de anticipar los requerimientos futuros y planear sistemas capaces de asimilar el cambio y el crecimiento.

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 de investigación *"Hacia los sistemas de almacenamiento distribuido de información, orientados por la naturaleza de los contenidos".*

INFOTEC es:

- Investigación **-** Educación **-** Soluciones integrales **-**