XUNTA DE GALICIA

ESCOLA GALEGA DE ADMINISTRACIÓN PÚBLICA

59. SISTEMAS DE BACKUP: HARDWARE Y SOFTWARE DE BACKUP. ESTRATEGIAS DE BACKUP A DISCO. DISPONIBILIDAD DE LA INFORMACIÓN RPO, RTO. REPLICACIÓN LOCAL Y REMOTA, ESTRATEGIAS DE RECUPERACIÓN.

Fecha de actualización 26 de julio de 2011

XUNTA DE GALICIA

ESCOLA GALEGA DE ADMINISTRACIÓN PÚBLICA

Tema 59.- Sistemas de backup: hardware y software de backup. Estrategias de backup a disco. Disponibilidad de la información RPO, RTO. Replicación local y remota, estrategias de recuperación.

59.1 SISTEMAS DE BACKUP: HARDWARE Y SOFTWARE DE BACKUP

Un factor importante en todo sistema de backup es la elección de los sistemas hardware y software que lo componen.

59.1.1 Hardware de Backup

En la categoría de elementos hardware de backup tenemos:

59.1.1.1 Cintas

Tradicionalmente, los cartuchos de cinta magnética son los medios de comunicación más habituales en los sistemas de backup. Como soporte de almacenamiento de los respaldos de datos, la cinta magnética tiene una larga historia de uso y es el medio de copia de seguridad con mayor nivel de madurez. La cinta magnética, o de una forma más abreviada, la cinta, es un componente basado en cartuchos que se hace típicamente de algún tipo de plástico rígido. Contiene uno o más bobinas de plástico flexible que se han impregnado con un material con comportamiento magnético.

Los cartuchos de cinta están fabricados en varios formatos. Cada formato tiene unas características diferentes que responden a las diferentes necesidades de almacenamiento físico y de tiempo de preservación de la copia de seguridad, tanto en términos de la cantidad de datos almacenados, como de vida útil de los medios de almacenamiento o su coste. Los formatos de cinta de uso común son los siguientes:

- DLT/ SDLT
- LTO
- AIT
- STK 9840/9940/T10000

Según el tipo de cada cartucho este posee distintas capacidades o características como la velocidad de funcionamiento. El mercado está renovando continuamente este tipo de dispositivos con el fin de mejorar ambos aspectos. Sin embargo, existen tres formatos que podemos considerar de los más comunes y tienen características particulares que se describen aquí como ejemplos de elementos arquitectónicos de diseño: *DLT, LTO, T10000 y STK.*

El resto de formatos, aunque sean formatos comunes, se utilizan normalmente para entornos especializados, como el archivado y almacenamiento intermedio (nearline storage) empleado entre el almacenamiento online y el almacenamiento de backups.

59.1.1.1.1 Digital Linear Tape (DLT)

Digital Linear Tape (DLT) es el formato de cinta más antiguo y por lo tanto uno de los productos más maduros del mercado. Originalmente fue diseñado e implementado por DEC en 1984, para posteriormente ser adquirida por Quantum y redistribuido en 1994.

DLT es el primer cartucho de cinta compacta para copias de seguridad de sistemas abiertos en la empresa. Mientras que otros tipos de medios se encontraban en uso (como la cinta media pulgada, 4mm/8mm, y otros), DLT proporciona el mejor compromiso entre todos los factores debido a su tamaño, la fiabilidad de su almacenamiento, la capacidad, y disponibilidad relativa.

La conectividad de DLT, se limita a los tradicionales de SCSI, y está limitado a 300 GB de capacidad nativa de almacenamiento y 160 MB /seg velocidad de transferencia (SDLT600). Existían otras variantes disponibles, pero nunca llegaron a popularizarse con carácter general. Hoy en día, DLT se encuentra normalmente como copia de seguridad de larga duración en entornos pequeños que no requieren mayor capacidad.

59.1.1.1.2 Linear Tape Open (LTO)

Linear Tape Open (LTO) fue diseñado y concebido como una evolución y alternativa a los formatos DLT y otros ya existentes, y estaba destinado a proporcionar una plataforma común para los backups en cinta.

Seagate, HP e IBM fueron los iniciadores originales del consorcio LTO, encargado de realizar el desarrollo inicial y el cuál mantiene la licencia de la tecnología y la certificación del proceso. En teoría, se debería de haber

producido un formato estándar de cinta, con el cual los fabricantes podrían seguir trabajando con el estándar en el mercado e incorporando sus propias características y funciones adicionales.

Sin embargo, entre el original LTO-1 y los formatos de LTO-2 hubo problemas de compatibilidad. Estos problemas abarcaban desde bloqueos en las cintas cuando se utilizan medios adquiridos a dos proveedores distintos hasta la incapacidad de una unidad LTO de un fabricante a leer los datos escritos en un cartucho de otra.

El LTO-1 inicial proporcionaba 100 GB de almacenamiento nativo y 15 MB /seg; con los actuales sistemas de LTO-4 se proporcionan 400 GB de almacenamiento nativo de 160 MB / seg. Por su parte, el LTO-5 proporciona 800 GB de capacidad de almacenamiento nativo a 160 MB / seg.

59.1.1.1.3 Sun StorageTek T10000 (T10k)

El T10000 / StorageTek (T10k) de Sun representa uno de las tecnologías de almacenamiento en cinta que mejor se ha comportado en términos de capacidad. El T10k es un formato propietario producido únicamente por StorageTek y se encuentra normalmente en entornos en los que se empleaban las tecnologías anteriores de Sun como el STK (9840/9940). También se han utilizado en sistemas abiertos de servidores o mainframe. El T10k está diseñado para 500 GB de almacenamiento nativo de 120 MB / seg.

59.1.1.1.4 Características de almacenamiento en cinta

Aunque todos los datos anteriores indican un valor interesante en cuanto al rendimiento, todos los dispositivos de cinta con características similares de rendimiento deben tenerse en cuenta a la hora de diseñar entornos de backup.

La primera y más importante de ellas es el hecho de que todas las unidades de cinta son entornos serie. A diferencia de los dispositivos de disco, los dispositivos de cinta escriben los bloques de datos de forma lineal, uno tras otro. Las unidades de cinta sólo tienen una cabeza de escritura que escribe un bloque de datos de cada vez en la cinta, a medida que ésta se mueve por ella. Los dispositivos de disco tienen una serie de dispositivos de escritura, o cabezas, que se mueven a varios puntos del disco giratorio para situar los datos de una manera óptima. Esto permite que los dispositivos de disco puedan leer cualquier trozo de información solicitada. Dado que los discos tienen varias cabezas para obtener bloques de datos en paralelo, varios sistemas pueden acceder al disco al mismo tiempo

La lectura de los datos de una cinta, se realiza mediante el proceso inverso: La cinta debe rebobinarse hasta el principio, hacia adelante hasta bloque que se necesita, y leer así el bloque de datos. Al poder devolverse únicamente un segmento de datos con cada lectura, los dispositivos de cinta no se pueden compartir de forma paralela entre sistemas sin un mecanismo para transferir el control entre los sistemas que usan dicho dispositivo.

El tipo de conectividad también tiene influencia sobre la utilización de dispositivos de cinta. Las unidades de cinta dependen de una conexión directa con el host para el transporte de los datos. Una vez más, esto se debe al hecho de que las unidades de cinta son dispositivos de serie que sólo aceptan una sola conexión a la vez.

59.1.1.2 Disco

La cinta proporciona un método muy maduro, muy conocido, y de bajo coste para almacenar copias de seguridad. Sin embargo, las debilidades, tales como la naturaleza secuencial de la cinta, la complejidad mecánica, y la gran variabilidad del rendimiento de los dispositivos de cinta están rápidamente relegando a la cinta a medio de almacenamiento secundario o terciario en muchos entornos.

Con todos los problemas con la cinta, los administradores buscaban un medio que permitiera un rápido acceso a las copias de seguridad y que

proporcionase una forma de tener un almacenamiento rápido y fiable: *el disco.*

Los backup a disco son simples sistemas de archivos que han sido situados aparte para que el software de backup los use. Aunque esto parece sencillo, la implementación y gestión de las soluciones basadas en disco pueden ser muy complejas.

El almacenamiento en disco supera algunas de las desventajas propias de las cintas. Por la capacidad de recibir datos de forma rápida, tiene múltiples flujos para almacenar copias de seguridad al mismo tiempo, y tiene la capacidad de presentar el almacenamiento de un número de maneras diferentes, dependiendo de la necesidad del sistema, por eso, el disco es muy empleado como un medio de almacenamiento de copia de seguridad primario.

Pero el disco también tiene sus debilidades, el coste de los medios de comunicación, la falta de portabilidad, y la dificultad de asegurar la plena utilización de los medios de comunicación hacen que el disco no sea tan satisfactorio como parece a priori.

59.1.1.3 Medios Virtuales

Los medios virtuales emulan el hardware físico de cinta con el objetivo de reducir o eliminar los problemas de gestión asociados a los medios físicos. Mediante la eliminación del hardware con una alta complejidad mecánica y de gestión y la eliminación de sus sistemas asociados y reemplazándolos por unidades de disco, los medios virtuales también tiene la ventaja de aumentar la fiabilidad general del entorno de backup. Los medios virtuales ofrecen estas ventajas sin cambiar los procedimientos operativos o exigir modificaciones del software de copia de seguridad. Además, en algunos casos, el rendimiento puede aumentarse a través de un mejor uso del ancho de banda en los medios de comunicación utilizados para conectar los medios virtualizados con los servidores de backup.

Los Medios virtuales de copia de seguridad se asocian tradicionalmente de forma exclusiva con bibliotecas de cintas virtuales (VTL) pero recientemente se han realizado nuevas implementaciones a través de protocolos que permiten la virtualización de otros tipos de sistemas de almacenamiento.

59.1.1.4 Medios Ópticos

Los medios ópticos se sitúan entre las ventajas de las cintas y las del disco. Sobresalen en las áreas de fiabilidad, flexibilidad, ciclo de trabajo e inamovilidad, mientras que sus retos los encontramos en las áreas de rendimiento, capacidad y coste.

59.1.1.4.1 CD

CD, o compact disk, es un soporte digital óptico que se utiliza para el almacenamiento de prácticamente cualquier tipo de datos. En la actualidad el uso del CD está decayendo a favor del aumento del uso de un nuevo medio de similares características como el DVD.

El CD ha servido y sigue sirviendo como medio de almacenamiento de copias de seguridad gracias a su fiabilidad e inamovilidad. Proporciona en comparación con otro medios como la cinta magnética, mayor seguridad y protección de los datos, dado que el propio medio es mucho más robusto frente a interacciones físicas externas (por ejemplo los campos magnéticos).

Además de ser un medio habitual para el almacenamiento de pistas de audio, los CDs se utilizan habitualmente para la generación de copias de seguridad relacionadas con la recuperación de los sistemas.

Los sistemas de CD utilizan un dispositivo hardware específico para grabar información, conocido como grabadora/regrabadora de CD. Existen también dispositivos hardware similares que solamente permiten la lectura de este medio.

Las capacidades habituales de los CD estándar abarcan desde los 650MB hasta los 900MB.

59.1.1.4.2 DVD

Los DVDs vienen a ser la evolución de la tecnología digital óptica de los CDs.

Al igual que los CDs, existen dos tipos de dispositivos para el uso de los DVDs que son las grabadoras y los lectores. Existen diferentes tipos de DVDs y diferentes categorizaciones, siendo la más importante la relativa al número de capas, factor que determina la capacidad final del dispositivo.

Las capacidades actuales abarcan desde los 4,3Gb hasta los 17Gb. Los DVD utilizan dos tipos de sistemas de ficheros que reemplazas el antiguo ISO 9660 de los CDs, y que son el UDF y el Joliet.

59.1.2 Software de Backup

En la categoría de elementos software de backup tenemos herramientas de código abierto o software libre y software privativo o comercial. Las herramientas más comunes a nivel de software son:

59.1.2.1 Herramientas de código abierto – AMANDA

Amanda (Advanced Maryland Automated Network Disk Archiver), es el software de código abierto de copia de seguridad más conocido. Amanda se desarrolló inicialmente en la Universidad de Maryland en 1991 con el objetivo de proteger los archivos de un gran número de estaciones de trabajo cliente con un servidor de copia de seguridad único. James da Silva fue uno de sus desarrolladores originales.

El proyecto Amanda se registró en SourceForge.net en 1999. Jean-Louis Martineau, de la Universidad de Montreal ha sido el líder del desarrollo de Amanda en los últimos años. Durante años, más de 250 desarrolladores han contribuido al código fuente de Amanda, y miles de usuarios aportan pruebas y comentarios, lo que lo convierte en un paquete robusto y estable. Amanda se incluye con la mayor parte de las distribuciones Linux.

En un principio, Amanda fue utilizado mayoritariamente en las universidades, laboratorios técnicos, y departamentos de investigación. Hoy, con la amplia adopción de Linux en los departamentos de informática, Amanda se encuentra en muchos otros lugares, sobre todo cuando la atención se centra en aplicaciones LAMP (Linux+Apache+MySQL+PHP). Con los años, Amanda ha recibido múltiples premios de los usuarios.

Amanda permite configurar un único servidor backup maestro para realizar múltiples copias de seguridad de equipos Linux, Unix, Mac OS X, y Windows en una amplia variedad de dispositivos: cintas, discos, dispositivos ópticos, bibliotecas de cintas, sistemas RAID, dispositivos NAS, y muchos otros.

Las principales razones para la adopción generalizada de Amanda son:

- Se puede configurar un único servidor de copia de seguridad de varios clientes en red con cualquier dispositivo de almacenamiento: una cinta, disco o sistema de almacenamiento óptico.
- Está optimizado para el backup en disco y cinta, permitiendo escribir simultáneamente backup a cinta y disco.
- No utiliza drivers propietarios, cualquier dispositivo soportado por un sistema operativo también podrá funcionar en Amanda.
- Utiliza herramientas estándar, como dump y tar. Puesto que no son formatos propietarios, los datos se pueden recuperar con esas mismas herramientas.
- Se utiliza un planificador que optimiza niveles de seguridad para los diferentes clientes, de tal manera que el tiempo total del backup es aproximadamente el mismo para cada ejecución.
- Existe una amplia y activa comunidad de usuarios que crece día a día.
- El coste total de propiedad (TCO) de una solución de backup basada en Amanda es significativamente menor que el TCO de cualquier solución que utilice software privativo.

59.1.2.2 Herramientas de código abierto – BackupPC

BackupPC es un sistema de alto rendimiento que permite realizar copias de seguridad de sistemas Unix, Linux, Windows y MacOS en un disco. Es por tanto una herramienta basada totalmente en disco.

Ofrece una serie de ventajas como son:

- **Soporta cualquier sistema operativo cliente.** Esto se debe a que se utilizan herramientas estándar que o vienen con el SO o se pueden añadir al SO, sin necesidad de instalar cliente. Así resulta más fácil integrar un nuevo cliente.
- **Interfaz Web** con control de usuario para acceder a copias de seguridad. La mayoría de los SO trae un navegador web, así que usar

una interfaz web es otra manera de acelerar el proceso de incorporación de nuevos clientes con diferentes sistemas operativos. La interfaz web está diseñada para dar el máximo control posible al cliente de forma segura. El usuario puede solicitar restauraciones, y navegar fácilmente y restaurar archivos individuales. Sin embargo, el usuario no podrá ver las máquinas de otro usuario.

- **Soporte de clientes DHCP**. Mediante el uso de servicios estándar, BackupPC soporta clientes DHCP, siempre y cuando el cliente esté registrado con un servicio de nombres como DNS, Active Directory o LDAP.

Funcionamiento de BackupPC

El modelo de BackupPC tiene un usuario por cliente. Esto es así porque BackupPC fue específicamente diseñado para realizar copias de seguridad de PCs de varios usuarios (de ahí el nombre).

Normalmente, el usuario es el propietario de los datos de la máquina. Si se trabaja con un servidor de ficheros, el usuario deberá ser un administrador.

BackupPC envía mensajes de correo electrónico al propietario si no puede realizar la copia de seguridad después de un tiempo configurable; el propietario puede gestionar las restauraciones de las copias a través de una interfaz web.

En los siguientes puntos se describen algunas de las características proporcionadas por BackupPC:

- **Directo al disco**. BackupPC almacena todas sus copias de seguridad directamente en el disco. Los archivos idénticos en cualquier directorio o cliente se guardan sólo una vez, lo que reduce drásticamente los requisitos de almacenamiento del servidor. Estos archivos se almacenan en un conjunto de discos. Además del conjunto de discos, las copias de seguridad están en un árbol de directorios organizados por host.

BackupPC también tiene un proceso (que se lanza por las noches) que recupera espacio del conjunto de discos que no está referenciado por

ningún backup, lo que evita un uso inadecuado del espacio en disco. Este es un proceso automático que el administrador no tiene que configurar.

- **Sistema operativo del servidor** La parte del servidor de BackupPC está diseñada para ejecutarse en un sistema tipo Unix con Perl y mod_perl. Ofrece el mejor rendimiento con Apache, pero se puede ejecutar en cualquier servidor web que soporte Perl (se requiere mod_perl o Perl setuid.) El servidor debe tener un disco con gran capacidad o RAID para almacenar los backups.
- **Sistema operativo del cliente.** Como se comentó anteriormente, soporta cualquier SO. Las versiones más modernas de las variantes comerciales de Unix (Solaris, AIX, IRIX,HP-UX) traen en la propia distribución las herramientas tar, compress, gzip, rsync, y rsh y / o ssh. Otros sistemas operativos tipo Unix (Linux, FreeBSD, OpenBSD, NetBSD, Mac OS X) también cuentan con estas herramientas.

Los clientes de Windows pueden hacer copias de seguridad de diferentes formas dependiendo de si las políticas locales permiten o no la instalación de software. Si no se permite, BackupPC utilizar parte de la suite Samba (http://www.samba.org) para hacer backup de la información compartida mediante SMB o CFIS. Si se permite instalar software, se utiliza rsync junto con el conjunto de herramientas Cygwin (http://www.cygwin.com).

- **Soporte para herramientas nativas**. BackupPC utiliza las herramientas estándar de Unix para su funcionamiento interno. Esto incluye programas como Perl, tar, rsync comprime, gzip, bzip2, zip, apache y samba.

BackupPC no utiliza una base de datos o catálogo para almacenar la información de respaldo. En su lugar, utiliza el árbol de directorios para almacenar esta información. Esto simplifica las actualizaciones del sistema operativo del servidor de BackupPC o de la propia aplicación BackupPC.

- **Control de los backups y restauraciones a través de interfaz web**. La Web es la interfaz principal de BackupPC. Tras la configuración inicial, no es necesario acceder al servidor mediante línea de comandos para administrar BackupPC. La interfaz web está escrita en

Perl y fue diseñada para funcionar tanto con mod_perl como con CGIs o con Perl setuid.

La interfaz permite a los usuarios identificarse, acceder y controlar los respaldos y las restauraciones.

El usuario puede solicitar copias de seguridad de tipo one-time, de tipo completa, o de tipo incremental.

Se pueden utilizar varias opciones para recuperar ficheros:

- Los archivos individuales se recuperan mediante selección.
- Los grupos de archivos o directorios se pueden restaurar a su ubicación original.
- El usuario puede descargar los archivos como un archivo tar o zip.

El usuario tiene control absoluto sobre qué archivos o directorios se restauran y donde hay que restaurarlos. Un histórico muestra que archivos se han modificado durante cada copia de seguridad en cada directorio.

- **Soporte para clientes DHCP**. Los clientes BackupPC se referencian por nombre de host. Si la red de la copia de seguridad utiliza DHCP y se permite la resolución de nombres dinámica, no hay que hacer nada más para que el servidor BackupPC respalde a los clientes DHCP. Si este no es el caso, y los clientes son máquinas Windows, BackupPC se puede configurar para buscar un conjunto de direcciones de los clientes, localizándolos mediante SMB.

Si el cliente no está en línea durante el período de copia de seguridad normal, el servidor BackupPC no genera un error a menos que haya transcurrido un período de tiempo establecido desde la última copia de seguridad. En este punto, el servidor envía un email al propietario del cliente y le recuerda que se asegure que la máquina está en la red para hacer una copia de seguridad. (El servidor también puede enviar cualquier error al administrador.)

Los clientes que residen en otra LAN pueden ser gestionados a nivel local asumiendo que hay conectividad de entre las redes. Esto significa

que se puede hacer backup de los clientes conectados a través de una red privada virtual (VPN).

Si el usuario no desea realizar copias de seguridad en un momento dado, se conectaría a través de la interfaz web para cancelar la copia de seguridad.

- **Pool de Backups**. Cuando los clientes utilizan el mismo sistema operativo se duplican los archivos respaldados. Si se quiere mantener múltiples copias de seguridad completas aumenta el número de archivos duplicados, lo que aumenta los requisitos de capacidad de almacenamiento para el servidor. BackupPC almacena un árbol de directorios por cliente respaldado, pero comprueba si los archivos se han almacenado antes. Si es así, BackupPC utiliza un enlace que apunta al fichero existente en el conjunto de discos común, ahorrando una gran cantidad de espacio. Además, BackupPC puede comprimir opcionalmente para ahorrar más espacio.
- **Fácil configuración por cliente.** Una vez que el administrador haya definido cuáles deberían de ser las políticas de backup del sitio, es muy fácil anular cualquier opción de configuración en base a un cliente. Esto permite una gran flexibilidad sobre qué, cuándo, y cómo hacer copia de seguridad de un cliente. No hay clases de clientes por sí mismo.

59.1.2.3 Herramientas de código abierto – Bacula

Bacula es un conjunto de programas Open Source, listos para ser utilizados en un entorno doméstico y profesional, que permiten administrar los backups, restauración y verificación de datos en una red heterogénea. Bacula es relativamente fácil de usar y eficiente, a la vez que ofrece muchas funcionalidades avanzadas para la administración de los datos almacenados, lo cual facilita hallar y recuperar archivos perdidos o dañados. En términos técnicos, Bacula es un sistema de backups Open Source, orientado a la red y listo para la empresa.

Es capaz de realizar copias de seguridad en disco, cinta o medios ópticos. Bacula fue escrita originalmente por John Walker y Kern Sibbald en el año 2000. John dejó el proyecto no mucho tiempo después de su creación, y Kern, trabajó en él desde mediados del 2000 hasta el primer lanzamiento público de Bacula en abril de 2002. Desde entonces, otros desarrolladores han contribuido a su desarrollo.

Bacula está disponible bajo licencia AGPL versión 3. La página web del proyecto se encuentra en http://www.bacula.org, y los archivos descargables y un repositorio CVS se alojan en SourceForge.

Bacula Arquitectura

Bacula es una solución distribuida de backups. Esto significa que Bacula está compuesto por varios elementos, que pueden o no residir en el mismo host. Por ejemplo, se puede tener un host con el catálogo y en otro el storage.

Se basa en una arquitectura Cliente-servidor que resulta eficaz y fácil de manejar, dada la amplia gama de funciones y características que brinda: copiar y restaurar ficheros dañados o perdidos. Además, debido a su desarrollo y estructura modular, Bacula se adapta tanto al uso personal como profesional.

Se puede utilizar TLS (Transport Layer Security) para proteger los datos durante la transmisión.

Los componentes principales de esta arquitectura son:

- **Director (DIR)** es el encargado de gestionar de forma centralizada la lógica de los procesos de backup y los demás servicios. Trabaja en base a una unidad básica denominada JOB (un cliente, un conjunto de archivos, ...) de tal forma que el Director planifica, inicia y supervisa todos los jobs.

  También es el encargado de mantener el catálogo, por lo que el servidor de la base de datos debe estar accesible desde la máquina que ejecuta el Director.

- **Storage** es el encargado de gestionar los dispositivos de almacenamiento; esto exige que esté instalado en la máquina que posea la conexión física a los dispositivos de almacenamiento, tales como: discos locales, grabadoras, unidades de cinta, volúmenes NAS o SAN, autocargadores o librerías de cinta.
- **File Daemon** es el agente que corre del lado del cliente, es decir, en la máquina cuyos datos se van a respaldar, y que tiene como objetivo empaquetar los datos y enviarlos al Storage, donde serán almacenados.
- **Consola** es la herramienta que permite al usuario o administrador controlar Bacula. Se comunica con el director vía red, iniciando los jobs, revisando la salida del job, haciendo consultas y modificaciones en el catálogo.

Existen consolas en modo texto, modo GUI para Windows y Linux/UNIX e interfaces web.

- **Catálogo** es una base de datos donde se guarda información sobre los jobs y sobre los datos respaldados. El catalogo permite dos cosas:
  - Por un lado, como guarda información de los jobs, pools y volúmenes, Bacula lo usa para saber si hay un backup completo para un job, y si no lo hay, realizará para ese backup una copia completa.
  - Por otro lado, el catálogo tiene todos los nombres de archivo (y sus atributos, como fecha de última modificación, etc.) que se respaldaron, y eso es lo que permite hacer una recuperación selectiva, es decir, seleccionar (marcar, en la jerga de Bacula) individualmente qué archivos y/o directorios restaurar.

59.1.2.4 Software Propietario CommVault Simpana

Simpana comenzó como un proyecto dentro de AT & T Labs en 1987 y posteriormente fue adquirido por la empresa CommVault.

Simpania es un software de backup que realiza copias de seguridad de entornos Unix, Windows, Linux, servidores de correo Exchange, Lotus Notes, bases de datos Oracle, MySQL, SQLServer y máquinas virtuales

VMware. Además permite funciones avanzadas como puede ser el archivado, la deduplicación y la replicación de ficheros.

El funcionamiento de la aplicación se basa en el uso de los bloques de disco, por lo que todos los módulos no utilizan la información del archivo, si no que trabaja a más bajo nivel. Con ello consigue mejores ratios de compresión y una importante reducción de la ventana de backup, al utilizar únicamente los bloques modificados y no el fichero entero para realizar estas operativas.

Otra característica que incide en el uso de almacenamiento de bajo coste es la capacidad de generar políticas como las de archivado, mediante las que automáticamente permite mover ficheros de un almacenamiento a otro con mayor capacidad a menor coste. De esta forma, por ejemplo se podrían pasar los datos de una cabina de fibra a otra con discos SATA, pudiendo llegar a un tercer nivel a cinta, en base a unos requisitos (fecha del archivo, último acceso al archivo, etc.). Todos los movimientos se realizan de forma transparente para el usuario, tanto en el archivado como en su recuperación (si fuese necesario).

A estas funcionalidades hay que sumar la capacidad de deduplicación, que realiza una compresión de los datos aprovechando las duplicidades de los datos a nivel de bloque, consiguiendo alcanzar ratios de hasta el 50% de ahorro en el uso de almacenamientos en datos de segundo nivel y hasta el 90% en los de tercer nivel.

Para terminar el repaso a las principales funcionalidades, la replicación, permite la utilización de snapshots a nivel de cabina permitiendo volver el almacenamiento replicado a un estado anterior o montar la imagen snapshot como un recurso compartido.

Todo se administra desde una única consola centralizada, que simplifica toda la administración de la plataforma. Adicionalmente el motor de búsqueda ofrece la opción de buscar rápidamente y recuperar datos sin necesidad de saber donde se ubican.

59.1.2.5 Software Propietario Symantec NetBackup

Symantec NetBackup es actualmente el titular de la mayor cuota de mercado del entorno de software de copia de seguridad.

Netbackup 7 es la nueva versión de la solución de copia de seguridad y recuperación de datos orientada a grandes corporaciones. Esta herramienta trata de simplificar la gestión de la información reduciendo el volumen de almacenamiento de datos con técnicas de deduplicación en los ordenadores cliente de la red además del propio servidor, ofreciendo protección para entornos virtualizados. Todo ello con el único propósito de agilizar los procesos de backup y recuperación de datos.

La nueva herramienta incluye eliminación de datos duplicados nativos dentro del cliente NetBackup y permite a los clientes multiplicar por diez la velocidad de las copias de seguridad en oficinas remotas, el propio centro de datos y los entornos virtuales. Esta eliminación de datos duplicados en el cliente y en el destino ofrece una mayor cobertura con menos herramientas.

El proceso de deduplicación se contempla para todos los sistemas físicos y virtuales, independientemente del método de copia de seguridad. De este modo se integra una mayor protección para los cada vez más extendidos entornos virtualizados bajo las plataformas Hyper-V y VMware. Es en el caso de esta última en la que se ha podido observar un incremento de velocidad de hasta el 50% a la vez que disminuye el volumen de almacenamiento necesario en un 40%.

Otro de los aspectos notablemente mejorados en Netbackup 7 es la velocidad de recuperación de datos ante desastres. Permitiendo la restauración de grandes volúmenes de información en pocos segundos desde cualquier lugar y punto en el tiempo. Esta gestión se facilita al administrador de TI mediante un sistema centralizado de supervisión y alerta, que integra la administración de varios dominios de archivos con sus respectivas políticas de salvaguarda de datos.

La tecnología incluida en NetBackup acelerará la transición a un entorno virtual para las organizaciones empresariales que instalen un gran número

de máquinas virtuales o que decidan crear una infraestructura de nube privada.

La solución NetBackup también ofrece una elaboración de informes simplificada y un mayor soporte a las aplicaciones de bases de datos de Oracle y MySQL.

Algunas de las prestaciones y beneficios incluidos en la última versión de la herramienta son:

- La tecnología Virtual Machine Intelligent Policy incorpora la automatización a la localización y la protección de máquinas virtuales y minimiza los esfuerzos de administración necesarios para realizar copias de seguridad de máquinas virtuales VMware de alto rendimiento.
- Un 50% más de rapidez en copias de seguridad de máquinas virtuales gracias a que la tecnología Granular Recovery Technology (GRT) se encuentra ahora disponible para sistemas Linux en entornos VMware. Esto permite a los clientes reducir los tiempos comparables de copias de seguridad de máquinas virtuales en un 50%, además de simplificar la administración y mejorar la velocidad de recuperación de archivos individuales.
- Recuperación “a la carta” desde cualquier lugar con la nueva tecnología de replicación de imagen que permite a los clientes que replican datos entre múltiples sitios o dominios de NetBackup realizar backup de datos en un sitio alternativo.
- Recuperación acelerada: NetBackup RealTime ofrece soporte a entornos VMware para eliminar el espacio de tiempo entre copias de seguridad, además de reducir el impacto para grandes hosts de VMware y permitir la recuperación casi instantánea de sistemas completos.
- Satisfacer los requisitos normativos y de cumplimiento para seguimiento de auditorías.
- Incorpora informes mejorados de las políticas del ciclo de vida del almacenamiento, del seguimiento de las auditorías y del estado de las licencias.

- Deduplicación para Oracle mejorando el rendimiento de las copias de seguridad.
- Se añade un nuevo agente que presta soporte a MySQL para centralizar y automatizar las copias de seguridad y la recuperación de datos de las bases de datos de MySQL.
- Actualización simplificada de clientes con LiveUpdate que permite mejoras en equipos cliente para UNIX, Linux y Windows respecto a la versión NetBackup 6.5 y posterior desde una política única controlada por el administrador de NetBackup.

59.2 ESTRATEGIAS DE BACKUP A DISCO

Las estrategias de backup definen el plan que se ha de seguir para garantizar la integridad de la información. Los motivos por los que se debe establecer una correcta estrategia antes de comenzar a realizar las copias de seguridad pueden ser muy diversos, pero en esencia se trata de determinar la mejor manera para asegurar la información teniendo en cuenta las posibles dificultades de recuperación de parte de los datos, el coste de los medios que se emplearan y el tiempo que se necesitara.

Como no todos los sistemas son iguales, no todas las estrategias de backup son adecuadas para todos los sistemas. Partiendo de unas características comunes, algunas de las propiedades básicas de una estrategia backup son:

Tiempo de almacenamiento. Define el tiempo máximo que una copia permanece almacenada en un dispositivo. Al finalizar este tiempo la copia puede cambiar de dispositivo o ser borrada para liberar espacio en el medio de almacenamiento y poder hacer uso del mismo.

Almacenamiento alternativo. Posibilita realizar una o varias copias de seguridad en una ubicación externa al sistema y a la localización geográfica del mismo, manteniéndola durante un elevado período de tiempo, aumentando la seguridad ante cualquier catástrofe, ya sea a nivel de software o de hardware.

Protección ante fallo de los dispositivos. Establece el número de medios que se emplean. Cuanto mayor es el número de medios utilizados, mayor es la seguridad contra posibles pérdidas de información producidas por un fallo en el dispositivo de almacenamiento.

Tiempo de restauración. Esta característica especifica el tiempo de regeneración del sistema en caso de producirse algún fallo.

El coste. Suele ser un factor determinante a la hora de seleccionar la estrategia a realizar.

Las estrategias para la realización de copias de seguridad pueden ser muy distintas, dependiendo del sistema en cuestión sobre el cual se realizan.

En algunos casos, solamente se efectúa un backup de todo el contenido. Esto se produce cuando por algún motivo especial y muy específico o por algún motivo técnico, cuestiones de tiempo o por que existe un elevado riesgo para los datos. Alguno de estos casos especiales pueden ser:

No disponer del software original.

Desconocimiento de la ubicación de los ficheros de configuración.

Cambiar un disco de almacenamiento rígido.

Realizar cambios en las particiones de uno o más discos de almacenamiento rígidos.

Es habitual que este tipo de situaciones concretas se produzcan a la hora de llevar a cabo tareas de reparación o actualización sobre sistemas no controlados.

Cuando se trata de cubrir alguno de estos casos la estrategia de backup a seguir es sencilla, realizar un resguardo o copia de seguridad de todo el contenido de las unidades involucradas para así garantizar que no se perderá ninguna información y que será posible realizar la restauración completa del sistema.

Por otro lado, cuando realmente se ha de diseñar un plan estratégico para la realización de las copias de seguridad de un sistema propio o de una organización externa, se deben tener en cuenta una serie de pautas que ayudan a que el plan estratégico de backups sea el más conveniente y conseguir la mejor relación coste/beneficio posible.

Estas pautas aportan una reducción en el tiempo de respuesta a la hora de realizar una recuperación en caso de que se producirse cualquier tipo de contingencia.

Al intentar definir un plan de backups, surgen una serie de dudas:

¿Qué datos se deberían resguardar en cada backup? Datos a resguardar.

Es un factor determinante para una estrategia de backup que se determine el grado o grados de importancia de la información, es decir, establecer que información resulta de mayor valor para la organización. No tienen la misma transcendencia un documento de trabajo que una copia de respaldo de la configuración de una aplicación.

¿Cada cuánto se debería efectuar un backup de los datos? Frecuencia del backup.

Para determinar la periodicidad con la que se deben realizar las copias de seguridad no existe un criterio claramente definido. Sin embargo si se tienen en cuenta factores como:

Tiempo empleado en la creación de la información.

Coste invertido en la creación de la información.

Posibles consecuencias derivadas de su pérdida.

¿Cuánto tiempo deberían permanecer guardadas las copias de seguridad? Tiempo de Almacenamiento.

El período máximo de tiempo de estancia de una copia de seguridad en un dispositivo, es decir, el tiempo de retención, está directamente relacionado con los medios de almacenamiento disponibles, y por consiguiente por el presupuesto de la estrategia de backup.

Otra de las decisiones importantes a tomar durante la elaboración de una estrategia para la realización de copias de seguridad es la de seleccionar y planificar los distintos tipos de copias de seguridad.

Los backups son copias exactas de la información. Se pueden definir como instantáneas de los datos en un momento determinado, almacenados en un formato estándar, se puede realizar un seguimiento a lo largo de su periodo de utilidad y con cada nueva copia se mantiene la independencia con copia inicial. Se pueden crear múltiples niveles de backups, siendo los principales:

Copias de seguridad completas (Full backups): representan una copia exacta en un momento dado, de los datos que se pretende proteger. Proporcionan la base para todos los demás niveles de backup.

Por otro lado, están dos niveles de backup que capturan únicamente los cambios realizados sobre una copia de seguridad completa.

1. ***Copia de seguridad diferencial***, también conocida como la *copia de seguridad incremental acumulativa*, captura copias de seguridad que se han producido desde el último backup completo y suele utilizarse en entornos en los que no se produce un elevado número de cambios. La copia de seguridad diferencial se debe utilizar con cuidado debido a que puede crecer con rapidez e igualar e incluso superar el tamaño de la copia de seguridad completa.

   La ventaja de utilizar las copias de seguridad diferenciales viene dada en el momento de la restauración puesto que en el momento de restaurar una copia de seguridad diferencial sólo se necesita el backup completo y la última copia diferencial realizada. Debido a que únicamente se precisan dos imágenes para la restauración, la probabilidad de que ambas imágenes sufran algún percance, perdida, corrupción, etc., se reduce significativamente.

2. ***Copia de seguridad incremental***, es capaz de capturar los cambios que se han producido desde la última copia de seguridad realizada, independientemente del tipo que sea. Es la forma más utilizada para la realización de copias de seguridad, evidentemente combinada con una copia de seguridad completa.

   Este tipo de copia de seguridad contiene la menor cantidad de datos necesarios durante cada ciclo de backup, reduciendo la cantidad de datos que se transfieren y el tiempo que se necesita para la creación de una copia de seguridad.

Sin embargo las copias de seguridad incrementales tienen aspectos negativos. Si se está recuperando un grupo de archivos de un conjunto de copias de seguridad completas e incrementales, es probable que se requieran más de dos imágenes de copias de seguridad diferentes para completar la restauración, lo que aumenta la probabilidad de que alguna de estas partes sufra algún tipo de problema y no se pueda completar la restauración.

59.3 DISPONIBILIDAD DE LA INFORMACIÓN RPO, RTO

La información representa uno de los activos más importantes en el contexto actual, y como tal, debe existir siempre un plan estratégico y de contingencia que proporcione los mecanismos necesarios para garantizar la seguridad y disponibilidad de la misma.

Existen en el mercado una gran cantidad de soluciones tecnológicas y metodologías que nos permiten aplicar o instaurar protocolos específicos de protección y garantía de disponibilidad de la información corporativa, independientemente de la entidad en la que nos encontremos. Para establecer un criterio de selección entre toda esta gran cantidad de soluciones, existen un conjunto de indicadores técnicos que nos proporcionan un mecanismo estándar para poder establecer comparativas objetivas sobre cuál de las diferentes soluciones es la más conveniente. Estos dos indicadores son el Objetivo de Punto de Recuperación (RPO) y el Objetivo de Tiempo de Recuperación (RTO).

A grandes rasgos, podemos definir ambos conceptos de la siguiente manera:

Objetivo de Punto de Recuperación o RPO: Es la cantidad máxima de información que puede ser perdida cuando el Servicio es restaurado tras una interrupción.

Objetivo de Tiempo de Recuperación o RTO: Es el tiempo máximo permitido para la recuperación de un servicio de TI tras una interrupción.

Dentro de los planes de contingencia desarrollados para prevenir y paliar los casos de caída de servicio o pérdida de información en una organización, pueden existir diferentes alternativas aplicables en función de determinados criterios relacionados con el flujo y la cantidad de información con la que se trabaja. Para evaluar cuales son las técnicas más apropiadas, los conceptos anteriores marcan un punto inicial que se debe tomar como referencia para la implantación de las políticas adecuadas en el tratamiento de los datos.

59.3.1.1 Objetivo de Punto de Recuperación

El indicador RPO es una manera objetiva para comparar diferentes productos, sistemas o metodologías de recuperación de información cuando lo que interesa controlar es la cantidad de información que podría llegar a perderse en caso de contingencia. Como se ha definido con anterioridad, RPO establece un indicador que evalúa la cantidad de información que puede llegar a perderse sin graves consecuencias, es decir, es un indicador de riesgo.

Este indicador debe tomarse con cautela dado que es altamente dependiente del contexto en el que se encuentre la organización, así como de su volumen de generación de datos.

Para ilustrar el ejemplo, se plantea la situación de una organización en la que se realiza una copia de seguridad incremental cada noche. En este escenario, la pérdida máxima de datos que podrían llegar a perderse en caso de contingencia (fallo de los servidores, etc...) sería como máximo un día hábil, dado que asumimos que cada noche se realiza la copia de seguridad incremental.

Esta medida puede ser válida para determinados modelos de negocio en los cuáles el volumen de datos con el que se trabaja a lo largo de un día hábil no es demasiado elevado. Sin embargo, el mismo modelo aplicado a

una entidad bancaria en la cual se realizan millones de transacciones diarias, la pérdida máxima de un día hábil no es aceptable.

59.3.1.2 Objetivo Tiempo de Recuperación

El RTO determina el tiempo de recuperación frente a una contingencia, es decir, el tiempo que se puede estar sin el servicio operativo. Para ello es necesario identificar al inicio todas las funciones críticas del negocio y su apoyo a los componentes de Tecnologías de la información. Una vez identificadas, podemos establecer el tiempo necesario en caso de fallo para poder reanudar las operaciones normales.

El RTO es un indicador íntimamente relacionado con el BIA (Business Impact Analysis) y se suele expresar en términos de tiempo (horas, minutos,...). De hecho, en muchas organizaciones ya se establecen por norma primas sobre la disponibilidad de los sistemas y acceso a datos corporativos.

RTO y RPO están también enteramente vinculados. A la hora de diseñar un plan de contingencia, es necesario saber qué estrategia RPO se implantará dado que el volumen de datos a recuperar en caso de fallo, que está ligado a la estrategia RPO, influye directamente en el indicador RTO, es decir, en el tiempo que se tardará en recuperar el sistema.

59.4REPLICACIÓN LOCAL Y REMOTA, ESTRATEGIAS DE RECUPERACIÓN

La replicación es el proceso de creación de una copia exacta de los datos. La creación de una o varias réplicas de los datos de producción es una de las maneras de proporcionar continuidad al del negocio (BC).

Estos modelos pueden ser utilizados para operaciones de recuperación y reinicia de los sistemas en caso de que se produzca una pérdida de datos.

Una réplica ha de proporcionar:

La capacidad de recuperación: permite la restauración de los datos de los volúmenes de producción en caso de que se produzca una pérdida

de los datos. Se ha de proporcionar un mínimo de y RTO y un RPO concreto que nos garanticen la reanudación de las operaciones comerciales en los volúmenes de producción.

La capacidad de reinicio: garantiza la coherencia de los datos de la réplica, posibilitando la reanudación de las operaciones de negocio utilizando para ello la información contenida en las réplicas.

La replicación se pueden clasificar en dos grandes categorías: ***locales y remotos***

59.4.1 Replicación Local

La replicación local hace referencia al proceso creación de réplicas dentro del mismo array de discos o el mismo centro de datos.

59.4.1.1 Tecnologías de replicación local

Las replicaciones Host-based (basadas en replicación en host local) y Storage-based (basadas en almacenamiento) son las dos principales tecnologías adoptadas para la replicación local. La replicación de archivos del sistema y la replicación basada en LVM son ejemplos de la tecnología Host-based de replicación local. La replicación de almacenamiento basada en matrices de disco puede llevarse a cabo con soluciones distintas, la duplicación de todo el volumen, la replicación pointer-based de todo el volumen, y la replicación basadas en punteros y virtual.

59.4.1.1.1 Basada en replicación en host local

En este tipo de replicación, los administradores del sistema llevan a cabo el proceso de copia y restauración en la propia máquina, pudiendo basarse la recuperación en una replicación integral del volumen mediante LVM (Logical Volume Manager), o bien mediante instantáneas del sistema de ficheros.

Replicación del volumen mediante LVM: El LVM se encarga de crear y controlar el volumen de host a nivel lógico y está formado por tres componentes: los discos físicos, los volúmenes lógicos y los grupos de volúmenes. En la replicación de volúmenes basado en LVM, cada partición lógica en un volumen se asigna a dos particiones físicas en dos discos diferentes. De esa forma se consigue un espejo que permite redundancia y recuperación directa en caso de necesitar replicar.

Instantánea de archivos del sistema: Consiste en crear una réplica a base de instantáneas del sistema de ficheros mediante la utilización de metadatos almacenados en un mapa de bits. Estos metadatos van reflejando el cambio que se va produciendo en el sistema de ficheros y van almacenando un registro de las direcciones accedidas mediante operaciones de lectura/escritura. Este sistema requiere de una fracción del espacio utilizado por el sistema de ficheros original.

59.4.1.1.2 Basada en arrays de discos

En este tipo de replicación se hace uso de matrices de discos que pueden estar distribuidas dentro del CPD. El entorno operativo es el que lleva a cabo el proceso de replicación de un determinado sistema de ficheros, sin necesidad de que los recursos de acogida (CPU y memoria) del anfitrión intervengan en el proceso de replicación.

59.4.2 La replicación remota

La replicación remota consiste en el proceso de creación de replicas del conjunto de datos en lugares con otra ubicación física. Las réplicas remotas ayudan a las organizaciones a mitigar los riesgos asociados a las interrupciones regionales del servicio, que pueden estar provocadas por diferentes causas, por ejemplo, desastres naturales. La infraestructura en la que los datos se almacenan inicialmente se llama fuente. La réplica, o infraestructura remota en la que se almacena la copia se le llama blanco.

59.4.2.1 Tecnologías de replicación remota

La más habitual es la tecnología de replicación basada en host remoto, que utiliza uno o más componentes de la máquina para realizar y gestionar la operación de replicación. Existen dos enfoques fundamentales para la replicación basada en host remoto: Replicación remota basada en LVM y replicación de bases de datos a través de trasvase de registros.

59.4.2.1.1 Replicación remota basada en LVM

En este modelo, la replicación se efectúa y gestiona a nivel de grupo de volúmenes. El LVM de la máquina origen es el encargado de gestionar y transmitir la información del volumen al LVM de la máquina remota. El LVM de la máquina remota se encarga de recibir los datos y realiza la operación de réplica del volumen.

Antes del inicio de la replicación, se deben configurar los sistemas fuente y remoto para que los sistemas de archivos, los volúmenes y la agrupación de volúmenes sea idéntica en ambos. El punto de partida, o sincronización inicial, se puede realizar de diferentes formas, siendo la más frecuente la restauración en el punto remoto de una copia de seguridad de los datos de origen.

En la replicación remota basada en LVM se soportan dos modos de transferencia de datos, que son el sincrónico y el asincrónico. En el modo asíncrono, las operaciones de escritura se van almacenando en una cola de registros gestionada por el LVM y se van enviando al host remoto en el orden en el que son recibidas. En caso de fallo de la red, las operaciones siguen acumulándose en la cola de registros.

En la replicación síncrona, las operaciones de escritura deben estar comprometidas tanto en origen como en destino. Las operaciones de escritura consecutivas no pueden ocurrir en fuente ni destino hasta que las operaciones previas hayan finalizado. Esto garantiza que los datos de la fuente y destino son exactamente los mismos en todo momento. Esto hace posible que el RPO en caso de fallo sea cero o cercano a cero. Sin embargo, como contraprestación al nivel de seguridad, el tiempo de respuesta es mucho mayor. El grado de impacto en el tiempo de respuesta depende de la distancia entre ambos sitios (fuente y destino), del ancho de banda disponible y de la infraestructura de conectividad de red.

59.4.2.1.2 Basada en trasvase de registros

La replicación de bases de datos a través de trasvase de registros consiste en la captura de las transacciones realizadas en la base de datos fuente, que son almacenadas en registros que se transmiten periódicamente de un host fuente a un host destino. El host destino recibe el conjunto de registros y realiza las operaciones oportunas en la base de datos replicada. El proceso inicial de producción y reproducción requiere que todos los componentes importantes de la base de datos se repliquen en el sitio remoto.

Los sistemas gestores de bases de datos permiten definir un intervalo de tiempo para el envío de los ficheros de registro, o bien configurar un tamaño predeterminado de los mismos. Cuando un registro supera el intervalo de tiempo establecido o alcanza su tamaño máximo, se cierra, y se abre un nuevo fichero para registrar las transacciones. Los registros cerrados van siendo enviados desde la fuente al destino garantizando que la base de datos replicada en destino sea consecuente con la fuente hasta el último registro cerrado. El RPO en el sitio remoto dependerá del tamaño del fichero de registro y de la frecuencia de cambio de registro en la fuente.

59.5 BIBLIOGRAFÍA

- System & Disaster Recovery Planning. Richard Dolewski

- Information Storage and Management: Storing, Managing, and Protecting Digital Information. G. Somasundaram y Alok Shrivastava.
- Backup & Recovery. W. Curtis Preston y O'Reilly Media.

Autor: Francisco Javier Rodriguez Martínez

Subdirector de Sistemas Escola Superior Enxeñaría Informática Ourense

Colegiado del CPEIG