ADMINISTRACIÓN ELECTRÓNICA 24x7: ORACLE RAC Y LA ALTA DISPONIBILIDAD

*"El **mejor servicio al ciudadano** constituye la razón de las reformas que tras la aprobación de la Constitución se han ido realizando en España para configurar una **Administración moderna** que haga del principio de eficacia y eficiencia su eje vertebrador siempre con la mira puesta en los ciudadanos."*

(Extracto de la **Exposición de motivos** de la *Ley 11/2007, de Acceso electrónico de los ciudadanos a los Servicios Públicos*.)

Lydia López Sánchez
lydia.lopez@mpr.es
Jefa de Servicio de Sistemas
División de Proyectos Tecnológicos de la Admón. Pública
D.G. para el Impulso de la Admón. Electrónica
Ministerio de Presidencia

Palabras clave: 24x7, alta disponibilidad, bases de datos, SVDR, SVDI, Administración Electrónica

INTRODUCCIÓN

El reciente desarrollo normativo en materia de trámites administrativos se encamina hacia una Administración más sencilla y cercana al ciudadano, en donde los procedimientos sean eficaces, puedan realizarse por diversidad de canales, se eviten trámites innecesarios y estén disponibles a cualquier hora. En otras palabras, debe ser la Administración la que tienda a adaptarse al ciudadano, frente al enfoque tradicional en el que se le obligaba al ciudadano a ajustarse a la rigidez burocrática.

Así pues, las diversas Administraciones están realizando un notable esfuerzo a la hora de proveer servicios que cumplan con los principios de eficacia, eficiencia y simplificación en todo momento y en todo lugar. En este contexto, los Sistemas de Verificación de Datos de Residencia y Datos de Identidad aparecen como elemento clave en la simplificación de trámites administrativos.

Nos corresponde a los departamentos de Sistemas complementar y respaldar con la infraestructura adecuada a este tipo de proyectos. En esta línea se enmarca el proyecto presente, que proporciona alta disponibilidad a sistemas definidos como 24x7 mediante una infraestructura repartida en dos CPD con configuración activo-activo en todas sus capas. Para lograr esta alta disponibilidad, la capa de datos suponía un auténtico desafío, que finalmente fue resuelto mediante la implantación de un cluster extendido geográficamente basado en las tecnologías de Oracle RAC.

SITUACIÓN DE PARTIDA

EL PROYECTO INICIAL: CENTRO DE RESPALDO PARA LA SGPD

A lo largo de 2008, la entonces Subdirección General de Proceso de Datos de la Administración Pública, perteneciente al MAP, se embarca en el proyecto de implantación de un centro de respaldo para todos los servicios y aplicaciones para los que provee infraestructura de sistemas y comunicaciones.

El proyecto del centro de respaldo fue llevado a cabo conjuntamente por los departamentos de Comunicaciones y Sistemas. Entre las aplicaciones a respaldar, se establecieron dos grupos principales en función de sus requerimientos en cuanto a tiempo máximo de caída admisible.

GRUPO 1. Aplicaciones para la gestión de los recursos humanos de la Administración (Registro Central de Personal, Badaral, SIGP, FUNCIONA, etc.).

- Definidas como 5x12.
- Tiempo máximo de recuperación: 1 hora.
- Criticidad de los datos: alta.

GRUPO 2. Aplicaciones enmarcadas en el ámbito de la eAdministración: Sistema de Verificación de Datos de Residencia y Datos de Identidad (SVDR/I) y Servicio de Comunicación de Cambio de Domicilio (SCCD).

- Definidas como 24x7.
- Tiempo máximo de recuperación: 15 minutos.
- Criticidad de los datos: máxima.

OBJETIVOS

El proyecto del centro de respaldo se abarca teniendo siempre en mente los siguientes objetivos:

- Alcanzar el objetivo de disponibilidad para cada servicio, especialmente para aquellos definidos como 24x7.
- Garantizar la fiabilidad y la robustez de la información, aún en el caso de catástrofes.
- Optimizar el uso de los recursos disponibles, tratando de minimizar el número de sistemas ociosos en el centro de respaldo.
- Asegurar la escalabilidad del sistema y su continua evolución.

DESARROLLO DEL PROYECTO

ARQUITECTURAS A RESPALDAR

En el grupo 1 de aplicaciones (aplicaciones para la gestión de los recursos humanos), encontramos una serie de sistemas heterogéneos muy interdependientes entre sí y con arquitecturas complejas, difícilmente escalables o portables. Estas aplicaciones son herencia de sistemas de cierta antigüedad, fueron diseñadas en su día para correr en entornos *mainframe* y sus sucesivas migraciones se han limitado a cambiar la infraestructura sin llevar a cabo un rediseño en profundidad de los flujos de datos.

La situación en el grupo 2, caso del Servicio de Verificación de Datos de Residencia e Identidad, es bastante distinta al ser sistemas desarrollados en fechas relativamente recientes y basados en arquitecturas web. De modo genérico, son aplicaciones proveedoras y consumidoras de servicios web desplegadas sobre granjas de servidores de aplicaciones que a su vez se apoyan sobre clusters de base de datos. La complejidad en este caso no venía de los problemas de escalabilidad o portabilidad de la arquitectura sino del condicionante de 15 minutos como tiempo máximo admitido de caída.

INFRAESTRUCTURA

El centro de respaldo se ubica relativamente cerca del principal, a unos dos kilómetros aproximadamente. Entre las acciones que se llevan a cabo, destacan:

- Se establece entre ambos un enlace de fibra que permite configurar una única SAN para los dos CPD.
- Igualmente, se implementa un enlace de comunicaciones dedicado entre ambos CPD. El departamento de Comunicaciones implementa VLANs extendidas de forma que los nodos pertenecientes a una misma subred se vean unos a otros sin necesidad de enrutamiento, independientemente de su ubicación en uno u otro CPD.
- Se adquiere una pareja de cabinas de discos HP XP 20000, una para cada CPD, que permiten replicación síncrona de datos entre ellas.

ALTERNATIVAS DE SOLUCIÓN

En un principio, el centro de respaldo se plantea como un centro de tipo *cold-site*, de configuración pasiva, pero con procedimientos de puesta en marcha ante contingencia en tiempos relativamente cortos.

Sin embargo, a la hora de escoger entre las distintas alternativas barajadas, nos encontramos con dos problemas muy distintos:

- Por un lado, la compleja arquitectura heredada del grupo 1 de aplicaciones presenta unas dificultades enormes a la hora de plantear soluciones balanceadas o de tipo activo-activo.
- Por otro lado, el establecimiento de 15 minutos como tiempo máximo de parada para el grupo 2 de aplicaciones supone todo un conjunto de retos en cuanto a los procedimientos de contingencia y recuperación.

Hay que tener en cuenta que los 15 minutos comienzan en el mismo momento en el que se produce el incidente, por lo que en este tiempo se incluye lo que se tarda en detectar el problema, en avisar a los responsables de poner en marcha los procedimientos y, por último, en llevar a cabo el procedimiento de paso a contingencia. Por tanto, dado lo ajustadísimo de la ventana temporal, se decide que la solución, al menos para este grupo de aplicaciones, debe tender hacia una configuración activa-activa.

Al tener que solventar dos problemas casi incompatibles entre sí --la dificultad de balancear los sistemas heredados del grupo 1 y el exigente SLA del grupo 2--, se opta por dos filosofías diferentes en cada caso, que se expondrán a continuación.

RESPALDO DEL GRUPO 1 DE APLICACIONES

Para este tipo de aplicaciones, ante la imposibilidad plantear siquiera redundancia de servidores en algunos casos, se opta por reproducir una réplica de todo el entramado de servidores y relaciones entre ellos en el centro de respaldo. Esta infraestructura estará inactiva en operativa normal y sólo pasará a funcionar en situación de contingencia.

Dada la enorme interrelación de todos los sistemas entre sí, el paso a contingencia debe ser completo, de todas las aplicaciones a la vez.

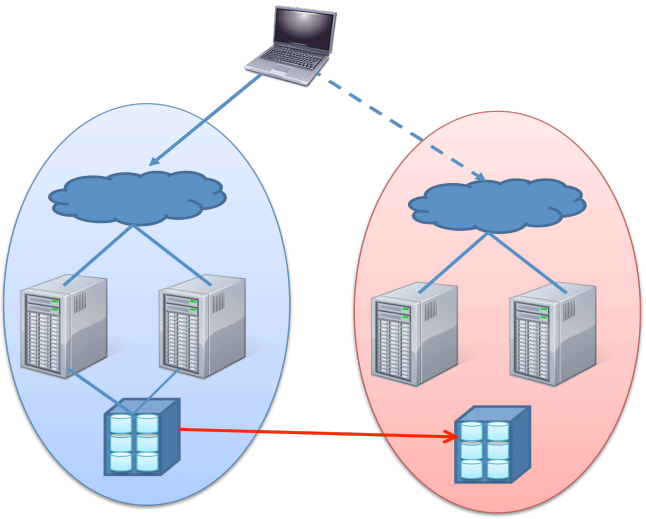
Otro problema de esta solución es la infrautilización de recursos, ya que obliga a tener parte de la infraestructura inactiva. Para minimizar este problema, se opta por disminuir el número de servidores en respaldo. Ello supondrá un menor rendimiento en caso de contingencia, pero por el contrario serán menos los recursos inactivos que compondrán nuestra infraestructura.

La réplica de las bases de datos se realiza a bajo nivel, mediante la replica síncrona ente las cabinas de discos (*disk array mirroring*).

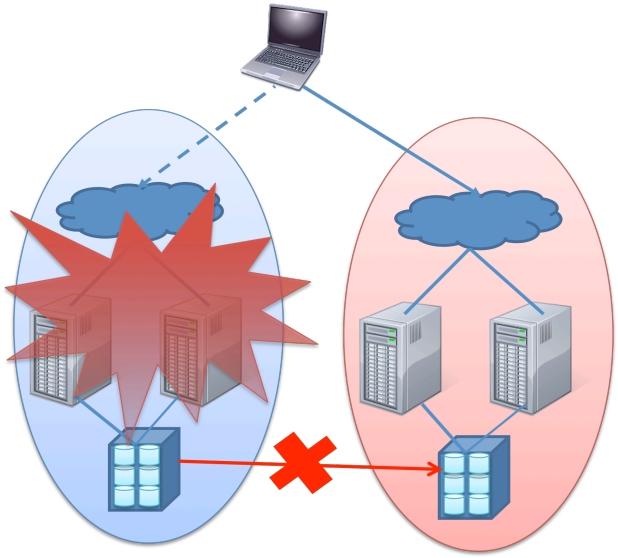
El procedimiento a seguir en caso de contingencia comprende:

1. Detener la réplica entre cabinas.
2. Montar los sistemas de ficheros en los servidores de base de datos.
3. Levantar las bases de datos.
4. Levantar los servidores de aplicaciones.
5. Redirigir el tráfico hacia el CPD de respaldo actuando en los balanceadores que proveen el acceso a los servicios.

Poniendo el foco en la capa de base de datos, el esquema simplificado en situación normal de operativa sería el siguiente:



En caso de contingencia, la réplica entre cabinas se detiene y se pasa a activar la infraestructura de respaldo:



Se trata de un proceso fundamentalmente manual, si bien se han automatizado ciertos pasos mediante el uso de *scripts*. El tiempo de parada es medio, dentro de los márgenes establecidos en el SLA inicial del proyecto.

No obstante, se está trabajando conjuntamente con los departamentos de Desarrollo para ir migrando progresivamente las aplicaciones de forma que la arquitectura se racionalice y permita una configuración activa-activa.

RESPALDO DEL GRUPO 2 DE APLICACIONES

La solución anterior, además de ofrecer el inconveniente del despilfarro de recursos, resultaba del todo inapropiada para el grupo 2 de aplicaciones debido a que incumplía el tiempo de parada máximo en caso de contingencia.

En este caso, los aplicativos basados en apache-tomcat ofrecían un escenario ideal para tratar de buscar una configuración activa-activa. El balanceo y el *failover*, en el caso de servidores apache-tomcat, es fácilmente configurable. No obstante, la configuración activa-activa pura ofrecía dos dificultades fundamentales:

- Cómo resolver el balanceo entre CPD en la capa de acceso de comunicaciones.
- Lograr el activo-activo puro en la capa de datos, donde la pérdida de la cabina de discos en una situación de contingencia planteaba un serio problema como veremos más adelante.

La primera cuestión se resolvió mediante el uso de una pareja de balanceadores, uno en cada CPD, que presentan en SARA una VIP por cada una de las aplicaciones.

Dichos balanceadores se encargan de dirigir el tráfico desde esa VIP a los distintos servidores que constituyen el grupo de balanceo.

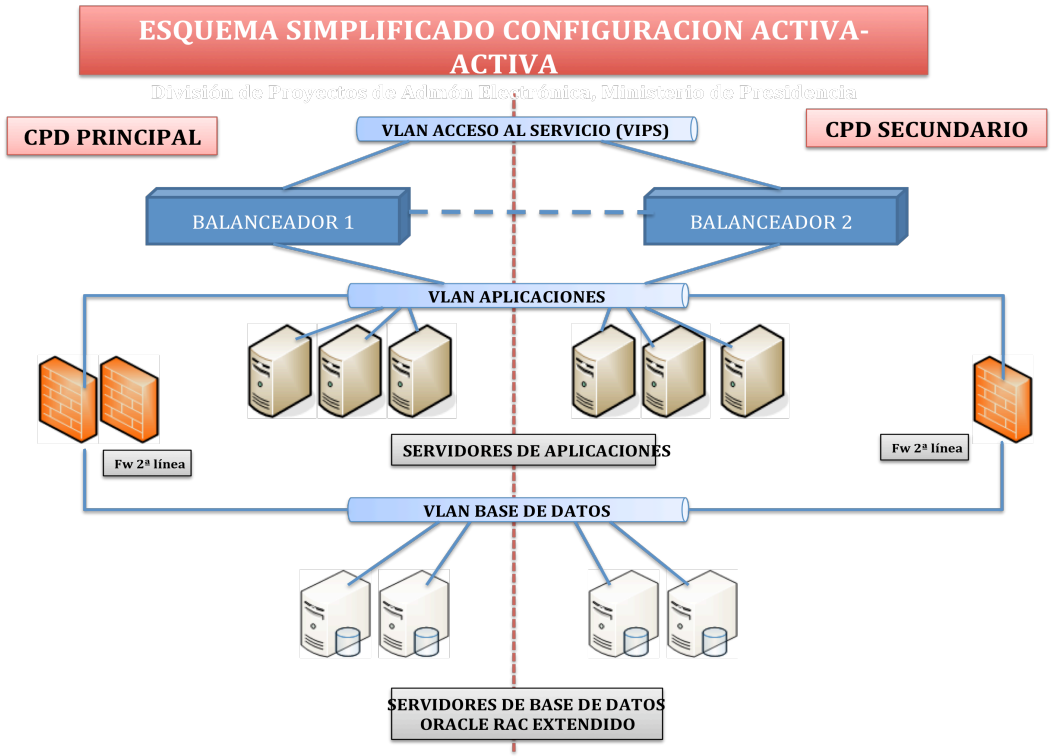
Se definen tres VLAN: una de acceso al servicio para la VIP, presentada en SARA; otra, por debajo de los balanceadores, de donde cuelgan los servidores de aplicaciones; y una tercera por debajo de los firewall de segunda línea donde se encuentran los servidores de bases de datos.

Las tres VLAN están extendidas, de forma que todos los elementos de una misma subred se ven entre sí sin necesidad de enrutamiento, independientemente de su ubicación en uno u otro CPD.

De este modo, una petición dirigida a una VIP es recogida por uno de los balanceadores y dirigida a cualquiera de los servidores de aplicaciones, en cualquiera de los dos CPD. La petición del servidor de aplicaciones de un CPD puede a su vez ser atendida por un nodo del cluster de base de datos que se encuentre en el otro CPD.

En cuanto a la solución en la capa de datos, se resolvió con un cluster extendido implementado con Oracle RAC 11g, como se verá con mayor profundidad en el siguiente punto.

Esta arquitectura supone un aprovechamiento total de los recursos, está completamente redundada y carece de punto único de fallo. Además, es completamente escalable puesto que la adición de nuevos servidores de aplicaciones o de nuevos nodos al cluster de base de datos es sencilla y puede realizarse sin pérdida alguna de servicio.



ORACLE RAC Y LA BÚSQUEDA DEL ACTIVO-ACTIVO

Uno de los grandes problemas a la hora de implementar soluciones activo-activo se encuentra en la capa de base de datos, ya que la pérdida de la cabina de discos supone la caída total del servicio.

En un principio, las soluciones que se contemplaron fueron:

- **Réplica síncrona basada en cabina (*disk array mirroring*).** Es la opción que se decidió para el grupo 1 de aplicaciones. Se trata de una solución sencilla de implementar y que no requiere licencias adicionales al mantener la base de datos de respaldo inactiva. Por el contrario, supone una infrautilización de recursos y un tiempo de parada medio-alto en caso de contingencia. Además, si la contingencia es brusca, sólo se conservarán aquellos datos que estén consolidados en disco y podrían perderse transacciones.
- **Oracle Datagard.** Este producto de Oracle requiere tener una infraestructura de base de datos paralela a la de producción, si bien puede ser inferior en recursos. En este caso, cada transacción de la base de datos se va replicando automáticamente mediante el software de Oracle en la base de datos de datagard. En caso de contingencia, esta segunda base de datos podría actuar como base de datos principal, si bien requiere al igual que en el caso anterior una serie de procedimientos manuales para su puesta en producción. La ventaja frente al anterior es que, al ser el mismo software de Oracle el que gestiona la réplica, la pérdida de datos no consolidados en caso de contingencia se ve muy disminuida. En contra, sí requiere licencia adicional y supone, como en el caso anterior, una infrautilización de recursos, si bien es cierto que las últimas versiones de Datagard permiten tener la base de datos de respaldo levantada en modo de solo lectura.

Ambos casos seguían sin resolver el problema de los 15 minutos de tiempo máximo de parada. Ante esta circunstancia, se optó por una tercera vía de solución consistente en la implantación de un RAC extendido con tecnología Oracle RAC 11g.

Las características generales de un cluster extendido de este tipo son:

- Cada uno de los nodos tiene visibilidad de los dispositivos de almacenamiento que le presentan las dos cabinas. Para evitar el punto de fallo, los nodos tiene tarjetas de fibra redundadas y ven ambas cabinas por dos caminos diferentes.

- Existe una capa de software, Oracle ASM (*Automatic Storage Management*), que gestiona el almacenamiento de forma que todos nodos ven los discos que les presentan las cabinas como un único pool de almacenamiento.

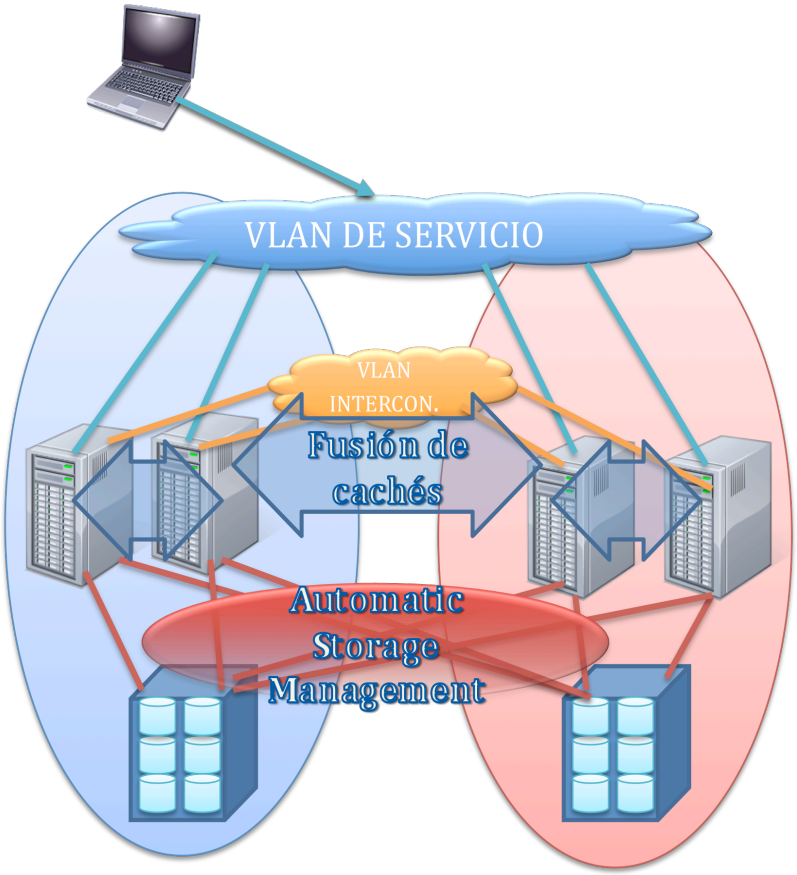
- Los nodos se comunican entre sí mediante una red privada (red de *interconnect*) a Giga. A través de esta red intercambian información de lo que está haciendo cada nodo y realizan una fusión de cachés (todos los nodos comparten la caché), que es lo que permite que la caída de uno de ellos sea transparente para la sesión del usuario.

- Todas la interfaces de red, tanto las del *interconnect* como las de la red de servicio deben estar redundadas (mediante *bonding* en Linux, IPMP en Solaris, etc.) para evitar que constituyan un punto de fallo.

- Otra capa de software por debajo de ASM, Oracle Clusterware, gestiona todos los recursos del cluster, mantiene al nodo dentro del grupo y proporciona a ASM los recursos necesarios para su funcionamiento.

- Toda la información de estado del cluster se ubica en un dispositivo denominado *voting disk*. Para evitar que este constituya un punto de fallo, Oracle recomienda mantener tres *voting disk*, uno en cada cabina y un tercero en un servidor ubicado en otro CPD al que accedan los nodos mediante NFS.

Un esquema simplificado de un cluster de este tipo sería el siguiente:



La decisión de optar por un cluster extendido nos permitió, en este caso, lograr:

- Una configuración activo-activo pura, con aprovechamiento de todos los recursos disponibles.
- Completamente tolerante a fallos gracias a la redundancia de todos sus elementos.
- Alta disponibilidad total: 24x7 garantizado.
- Fácilmente escalable con posibilidad de añadir nuevos nodos en caliente.

SITUACIÓN ACTUAL

INFRAESTRUCTURA DEL SERVICIO DE VERIFICACIÓN DE DATOS DE RESIDENCIA E IDENTIDAD

Como se ha venido comentando a lo largo de este documento, tras la puesta en marcha del nuevo centro de respaldo, el Servicio de Verificación de Datos de Residencia e Identidad funciona actualmente en una configuración activa-activa pura sobre la siguiente infraestructura:

- Una pareja de **balanceadores Nortel ALTEON 2424** en la capa de acceso al servicio.

- **Cuatro servidores SPARC** (dos en cada CPD) con **Solaris SunOS 5.10** para la **capa de aplicación**. En cada uno de ellos existen dos zonas virtuales Solaris (ocho servidores en total), en las que corren un servidor apache-tomcat.

- La capa de base de datos se compone de otros cuatro servidores SPARC con **Solaris SunOS 5.10** con las siguientes características:
  - Acceso a la SAN redundado en todos los elementos.
  - Bond de interfaces de red (servicio e interconnect) mediante **IPMP**.
  - *Voting disk* en **dispositivos raw** ubicados en ambas cabinas. El terecero de los *voting* se haya en un servidor Solaris de gama baja ubicado en el CPD del MEPSYD.
  - Versión de Oracle Clusterware: **Oracle CRS 11.1.0.6.0** (próxima actualización a 11.1.0.7.0).
  - Versión de ASM: **Oracle Database 11g Enterprise Edition Release 11.1.0.6.0 – 64 bits** (próxima actualización a 11.1.0.7.0).
  - Versión de Oracle: **Oracle Database 10g Enterprise Edition Release 10.2.0.4.0 64 bits**.

PRÓXIMOS PASOS

Continuar en la racionalización de recursos mediante la implantación de una infraestructura común para todos los aplicativos, separada en dos grandes grupos: aplicaciones de recursos humanos y aplicaciones de eAdministración. Esta infraestructura se apoyará en:

- Separación estricta de los entornos de Producción, Preproducción e Integración-Desarrollo, tanto en cuestión de VLAN como de recursos de infraestructura (servidores, bases de datos, etc.).
- Configuración activa-activa pura para la Producción, con una pareja de balanceadores (uno en cada centro de proceso de datos), proveyendo acceso balanceado a las aplicaciones, las cuales estarán presentadas en SARA.
- Virtualización en la capa de servidores de aplicaciones (ESXserver para entornos de Producción y Preproducción, ESXi para entornos de Integración y Desarrollo).
- Consolidación de bases de datos: dos grandes clusters extendidos para la Producción, dos clusters locales para la Preproducción y dos grandes servidores *standalone* para la Integración.

Migración progresiva de los aplicativos hacia dicha infraestructura, en estrecha colaboración con los grupos de desarrollo de aplicaciones en aquellos casos en los que dicha migración exija un rediseño más profundo del flujo de datos.