Bases de datos hoy

Bruno Crespo - Trustserver SL

Una pequeña tienda

- Una pequeña tienda de juguetes en Sevilla
- Estamos a tres días de Nochebuena
- Son las diez de la noche y la tienda está vacía, casi han vendido hasta las estanterías
- Al día siguiente a las diez de la mañana está llena otra vez ¿como lo han conseguido?

Logística en acción

- Justo antes de cerrar el TPV de la tienda se conecta a los servidores centrales y vuelca las ventas del día desde su base de datos MySQL a la base de datos Oracle de los sistemas centrales.
- Un programa se conecta a Oracle, calcula el stock de la tienda a partir del stock del día anterior y las ventas de hoy, después calcula la reposición necesaria.
- La información pasa al sistema de gestión de almacén (SGA) apoyado en otra base de datos Oracle.
- El SGA calcula el recorrido óptimo de los operarios de almacén y los dirige durante su trabajo mediante terminales móviles, coordinando todo a través de Oracle.
- Los automatismos de almacén siguen a las mercancías por el almacén gracias a su base de datos Informix.
- A las 3:00 am los envíos están listos para salir.
- A las 7:00 am el repartidor deja la mercancía para llenar la tienda en Sevilla.
- Cuando el TPV de la tienda confirma la recepción, Oracle transmite a una base de datos MS SQL Server el contenido del envío para que realice la facturación.

Logística en acción

- Gracias a MySQL, Oracle, Informix y SQL Server los niños serán felices unas Navidades más.

Nuestra vida cotidiana

- Casi cualquier actividad en nuestra vida utiliza una base de datos, y casi siempre es relacional:
  - Comprar en el supermercado.
  - Sacar dinero en un cajero.
  - Consultar los contactos en el móvil.
  - Consultar el saldo de la tarjeta del móvil.
  - Mirar nuestro muro de Facebook.
  - Leer slashdot.org.
  - Mirar el saldo de puntos del carnet de conducir.
  - Pedir un taxi.
  - Concertar una cita médica.
  - Ver un partido de futbol en pay-per-view.
  - Pujar por algo en e-bay.
  - Recibir un paquete por mensajero.
  - Domiciliar un recibo en el banco.
  - Llamar a emergencias.

Bases de datos en la empresa

- En general cualquier sistema empresarial consiste en mover registros en una base de datos u obtener informes sobre estos registros.
- Mover registros: Sistema transaccional o CRUD: CReate, Update, Delete.
- Obtener informes: Data Warehouse.
- El 100% de los desarrolladores que conozco profesionalmente dedican el 100% de su tiempo a desarrollo sobre bases de datos.
- Podéis estar seguros que el 90% de vosotros trabajará toda su vida profesional administrando, instalando o desarrollando sobre bases de datos.

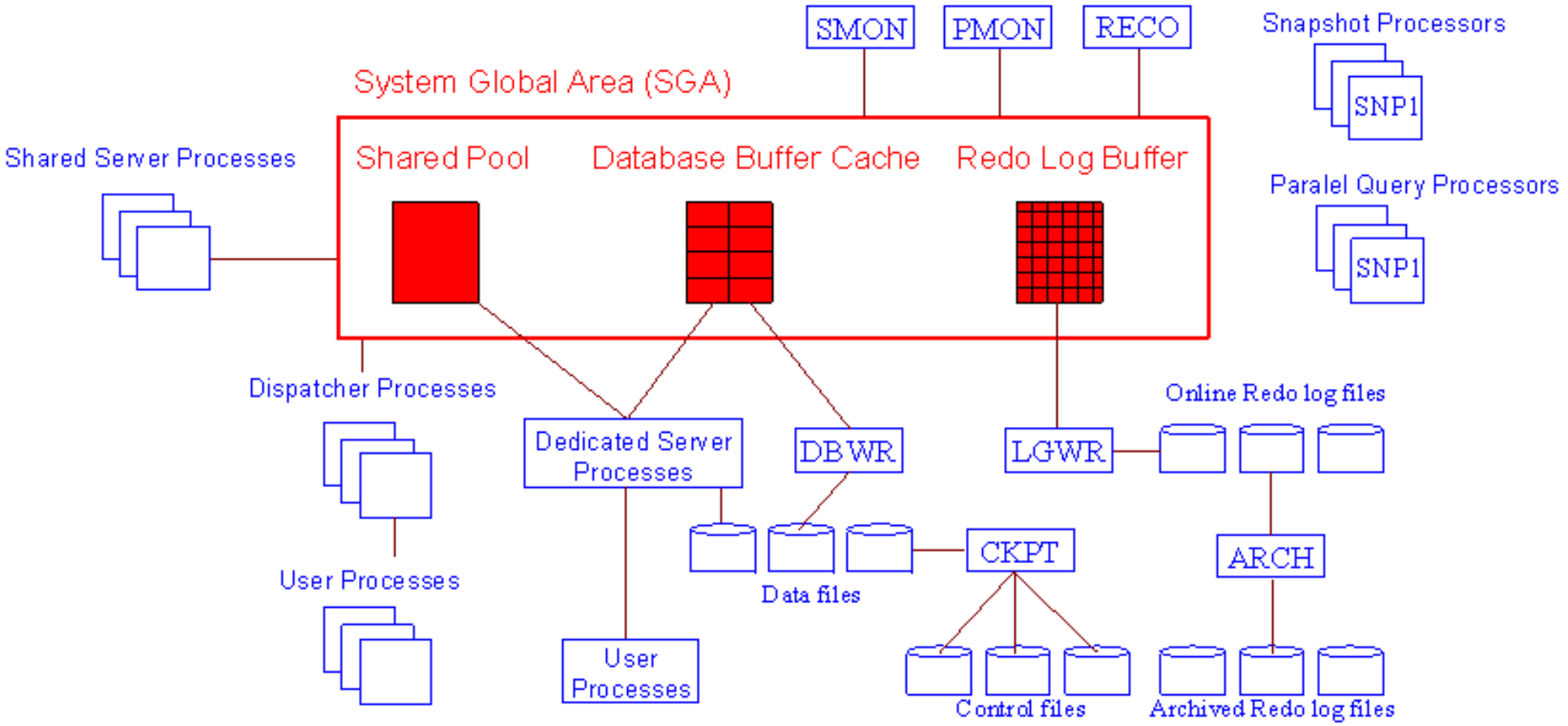
Sistemas RDBMS: Transaccional

- Oracle: La mejor.
- PostgreSQL: La mejor OpenSource.
- MySQL: Un juguete.
- DB2: La más veterana.
- MS-SQL Server: Un dolor.
- MS-Access: Todavía más doloroso.
- SQLite: La empotrada SQL más popular.
- BerkeleyDB: La empotrada más popular.
- Otras: Sybase, Informix, Firebird, Ingres...

Sistemas RDBMS: Data Warehouse

- Casi todas las bases de datos transaccionales tienen soporte para Data Warehouse pequeñitos (< 10TB y tablas con unos pocos miles de millones de filas)
- Teradata: La veterana.
- Netezza (IBM): Derivado de PostgreSQL y asistido en hardware.
- Vertica (HP): Derivado de PostgreSQL, asistido en hardware, organizado por columnas, Mike Stonebraker.

Estructura de una base de datos



Características habituales de las BBDD modernas

- ACID:Atomicity, Consistency, Isolation, Durability.
- Row Level Locking.
- MVCC: MultiVersion Concurrent Control.
- Hot Backup.
- Log shipping.
- Partitioning.

Las diferencias

- Cluster
  - Almacenamiento distribuído: muy útil para data warehouse (DB2 UDB).
  - Almacenamiento compartido: rendimiento en transaccional (Oracle, DB2 Mainframe).
- Oracle ASM: Hace un análisis de las zonas calientes del disco y distribuye los datos de forma óptima.
- Parallel query: Oracle EE, DB2, SQL Server.
- Integración con R: Oracle EE, Netezza, Vertica, Teradata.
- JVM en la base de datos: Oracle, DB2, PostgreSQL, SQLServer.
- Auditoría: Gran diferencia entre comerciales y open source.

Licenciamiento

- Las bases de datos son muy caras.
- Oracle es muuuy caro.
- Cluster dos nodos Intel, cabina de discos 12 TB
  - Hardware: 20K€
  - Oracle EE + RAC + partitioning: 230K€
  - No se incluyen los pack de administración, si los quieres +150K€
- El esfuerzo de entender y asegurar el cumplimiento de la licencia supone un coste extra significativo.

Licenciamiento

- El software que utilizas terminas pagándolo, si no quieres pagarlo, no lo utilices.
- No instales o desactiva todo aquello para lo que no tienes licencia, si no alguien terminará utilizándolo y, al final, habrá que pagarlo.
- Cuidado con el licenciamiento traidor!
- Estúdiate los contratos, licencias y documentación técnica sobre licenciamiento, aprovecha cada resquicio.
- Nunca mientas en una auditoría.

Comparar bases de datos

- Asistir a demostraciones, presentaciones comerciales o leer documentación comercial provoca daños cerebrales irreversibles. No lo hagas.
- Para evaluar un producto de software lee los manuales técnicos, no la documentación comercial.
- Evaluar un producto de software necesita trabajo y esfuerzo, normalmente de meses.
- Una licencia de Oracle puede costar millones de euros, dedica unos cuantos miles a asegurarte que lo necesitas.
- Sólo hay dos tipos de documento en que los fabricantes no mienten: a) los formularios de la SEC y b) el „Full Disclosure Report“ del Transaction Processing Council: familiarízate con ellos y aprende a entenderlos

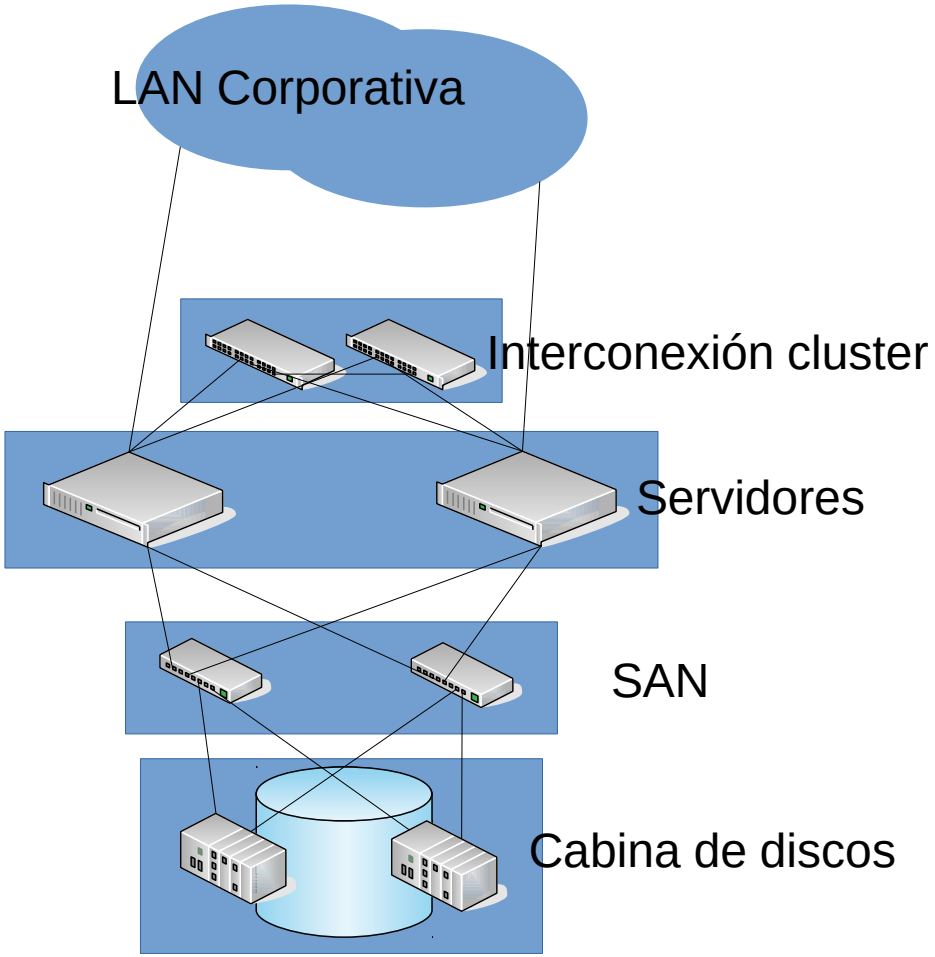
El futuro

- Transaccionales: Bases de datos basadas en RAM:
  - Base de datos en RAM con REDO LOG en disco y backup en caliente.
  - Cuando cierras la base de datos haces una copia de la RAM a disco, si no la haces recuperas el último backup y aplicas REDO.
  - SAP Hana ha logrado mejoras en instalaciones reales frente a Oracle x100000.
- Data Warehouse: Apache Hadoop.
  - Basado en la arquitectura Map-Reduce de Google.
  - Implementado en Java.
  - Utilizado en producción por Yahoo y Microsoft.
  - Apache Hive es un front-end SQL.
  - Casi todos los motores comerciales pueden integrarse con Hadoop para utilizarlo como back-end de cálculo.

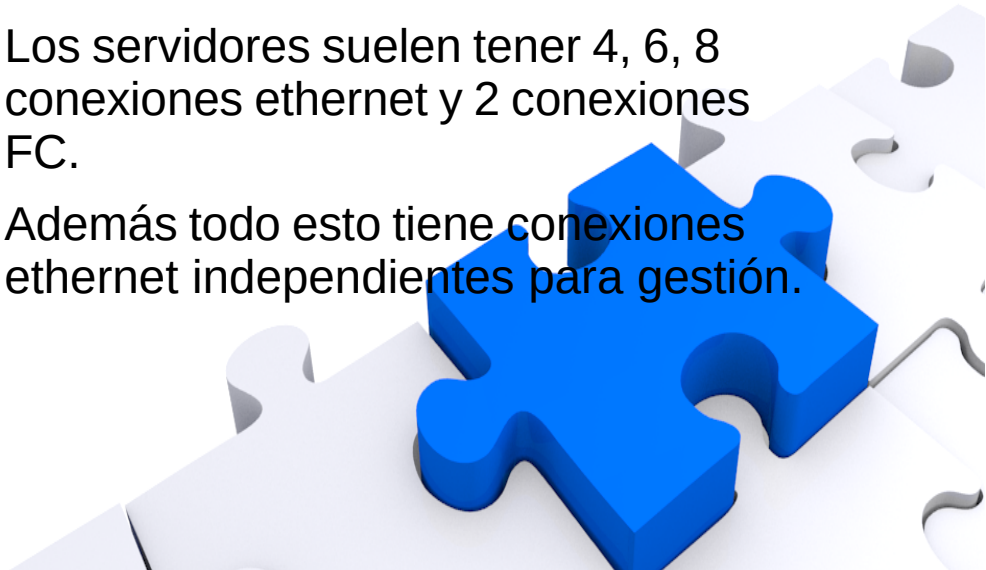
Hierro

- Una instalación moderna de base de datos rara vez es un equipo con discos internos.
- Las instalaciones de BBDD son críticas para las operaciones del negocio, así que tienen que ser resistentes a fallos.
- Todo está duplicado, se eliminan puntos simples de fallo.
- Para instalaciones críticas sólo se utiliza hardware „Enterprise Level“, nada de material de uso doméstico.
- Todo el hardware es gestionable: switches, servidores, cabinas, librerías de cintas...

Todo se duplica



- Todo está duplicado
- Dos servidores
- Dos switches ethernet para interconexión
- Dos switches FC en la SAN
- Cabina de discos con doble controladora
- Los discos se organizan en RAID 1, 5, 6, 10, 50 o 60
- Los servidores suelen tener 4, 6, 8 conexiones ethernet y 2 conexiones FC.
- Además todo esto tiene conexiones ethernet independientes para gestión.



Un armario de verdad

- Armario con dos servidores VMware y dos servidores Oracle, cabina de discos y dos switch ethernet.
- En este caso los servidores son un cluster activo-pasivo.
- Cabina de discos con 20 discos. La conexión de la cabina es SAS y no hacen falta switch FC.
- 200 metros de cable.
- Da servicio a 4 plantas industriales y más de 2000 usuarios interactivos.

Hierro: Mainframes

- IBM zSeries y predecesores: s/390, 3090, s/360.
- Virtualmente indestructibles: son capaces de sobrevivir a la destrucción de una CPU SIN que el proceso que está corriendo en ella cuando falla llegue a saberlo.
- Capacidad para decenas de miles de dispositivos conectados de forma simultánea.
- Rendimiento pobre en data warehouse, pero sobresaliente en transaccional.
- Caro no, carísimo
- Prácticamente ya sólo se utilizan para el transaccional de banca minorista.

Hierro: UNIX

- De las más de 35 versiones distintas a mediados de los 80, tras las devastadoras „Guerras UNIX“ y el efecto Linux sólo quedan 3 de uso corporativo: AIX (IBM), HP-UX (HP) y Solaris (Oracle).
- El hardware actual escala desde equipos similares a un servidor Intel hasta servidores bastante parecidos a los Mainframe.
- Todos tienen integradas soluciones de virtualización y/o particionamiento hardware.
- Son difíciles de ver en empresas con menos de 1000 M€ de ventas anuales.

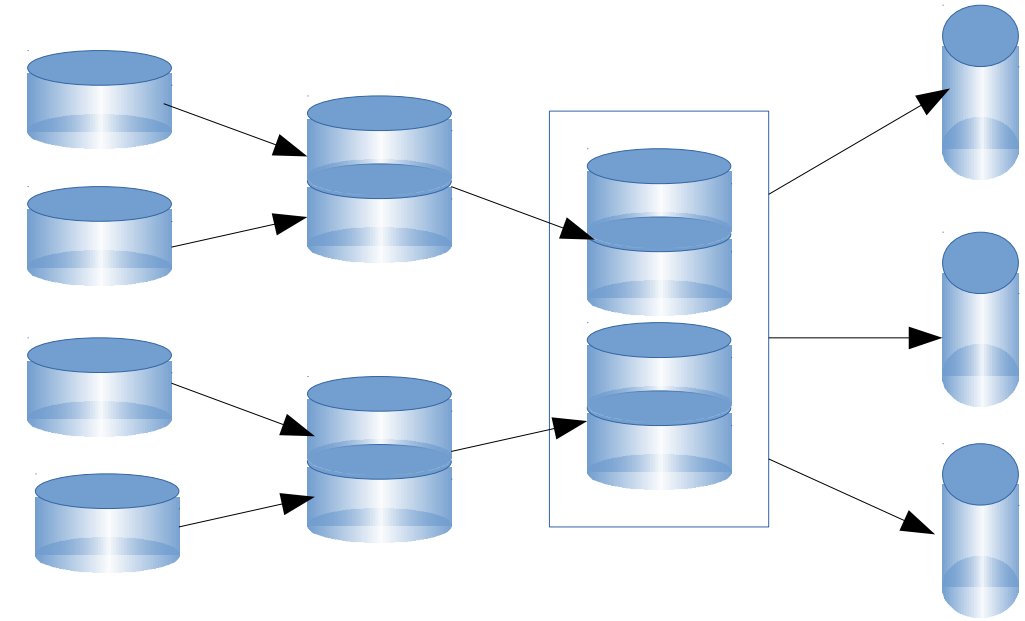
Servidores Intel

- Servidores con procesadores Intel o AMD.
- Arquitectura de 32 bit (i686) o 64 bit (x86_64), en la actualidad casi exclusivamente 64 bit.
- Sistemas operativos Windows o Linux.
- Para cargas grandes sólo Linux es viable.
- Caso real: Procesador Intel 1CPUx6core, 64GB RAM soporta 2000 usuarios en interactivo con uso del 15% de la CPU. Coste aprox.: 5K€.
- En la actualidad no hay ninguna carga transaccional que no pueda basarse en un servidor Linux con 2 socket.
- Para cargas pequeñas muchos servidores están virtualizados.

Cabinas de discos

- Caja para discos que permite cambio en caliente con dos controladoras.
- Las controladoras forman un cluster activo/pasivo (gama baja) o activo/activo (gama media y alta).
- Las controladoras tienen memoria caché protegida por baterías, esto permite que el cache write-back sea seguro si se pierde la alimentación eléctrica.
- Cada controladora puede tener conexiones Fibre Channel (FC), Serial Attached SCSI (SAS) o iSCSI (SCSI sobre TCP/IP). En muchos casos tienen las tres.
- Las cabinas juntan los discos en RAID, varios RAID se unen en un grupo de volúmenes y cada grupo de volúmenes se puede partir en volúmenes lógicos, cada volumen lógico se presenta a cada servidor como una LUN SCSI.

Cabinas de discos



Cabinas de discos: Lo que hacen

- Cache: entre 2 y 64GB de RAM por controladora
- Snapshots: La cabina hace una copia instantánea de un volumen lógico y presenta la copia a los servidores como una nueva LUN, muy útil para backup.
- Replicación: Una cabina puede replicar las operaciones sobre un volumen lógico en otro volumen lógico de la misma cabina o de otra distinta que puede estar a miles de km.
- Virtualización: Una cabina puede controlar otras cabinas esclavas y se presenta a los servidores como una sóla cabina unificada: útil para migración de datos y ampliaciones.
- Dynamic provisioning: Crea copias de volúmenes lógicos virtuales bajo demanda a partir de plantillas. Útil para escritorios virtuales.
- Multi-tier: Discos de distintas velocidades (SSD, mecánicos) y distribuye los datos dependiendo de lo calientes que están.

Cabinas: el lado oscuro

- Las cabinas modernas hacen maravillas, pero no hacen milagros.
- Para escribir un bloque de 4K en un RAID-5 hace falta leer un bloque de 64K de TODOS los discos, recalcular todos y escribirlos. Resultado: La velocidad en escritura aleatoria de un RAID-5 es peor que la de un sólo disco.
- No hacen distribución óptima de los datos calientes entre los discos físicos. Un buen DBA sí sabe hacerlo. Oracle sabe hacerlo.
- De los dos anteriores: Los discos físicos agrupados de dos en dos en RAID y presentados de forma independiente a los servidores da rendimientos mucho mayores que un RAID-5 con todos los discos presentado al servidor como una sóla LUN.
- El tamaño de la caché es irrelevante: Un dato que se pide a la cabina no está en la caché de la base de datos, y la caché de la base de datos siempre es mayor que la de la cabina. Resultado: el acierto de caché de la cabina siempre será muy bajo.
- La escritura del REDO LOG siempre debe ser preferente y sólo la base de datos sabe qué escritura es de REDO y qué escritura no.
- Si sólo utilizas RAID-1, y el tamaño de caché es irrelevante, entonces **una cabina de gama media-baja da el mismo rendimiento que una de gama alta y cuesta uno o dos órdenes de magnitud menos.**

Storage: Rendimiento

- Hay dos cosas que matan el rendimiento de una base de datos: RAID-5 o 6 e iSCSI: Sencillamente contesta „no“.
- Los discos que se utilicen para la base de datos han de ser de uso exclusivo. Si tu base de datos comparte los discos físicos con otra el rendimiento será pésimo.
- El almacenamiento para REDO LOG tiene que estar separado de todo lo demás y en discos de uso exclusivo para el REDO LOG.
- Paralelizar las tareas de backup para conseguir el tiempo de backup más corto posible no siempre es una buena idea.
- El número mágico: entre 15 y 20 discos por CPU (core) en cargas transaccionales.

Librerías de cintas

- Los sistemas grandes siguen haciendo el backup en cinta.
- Para poder gestionar el contenido de las cintas es necesario un sistema de backup adecuado.
- Los sistemas de cintas de uso corporativo consisten en librerías de cintas: una librería tiene uno o más lectores, capacidad para varias cintas (entre 10 y varios miles) y un brazo robot para trasladar las cintas.
- Coste: una TS-3100 de IBM con 1 lector, brazo robot y capacidad para 24 LTO-5, aprox. 2,5K€. Una cinta LTO-5: 30-100€.
- LTO-5: Capacidad por cinta: 1.5TB (3.0TB comprimidos), velocidad de transferencia: 130MB/s (sin comprimir).

Librerías de cintas

El trabajo de DBA

- „Un DBA es un hombre que viste, a la vez, cinturón y tirantes“
- Los DBA son los trabajadores más felices del mundo (http://www.businessnewsdaily.com/6130-happiest-unhappiest-jobs.html): Si tienes mucho stress no estás haciendo bien tu trabajo y duras poco.
- Ser DBA es una cosa muy polifacética:
  - Mantener la base de datos en buen estado.
  - Ser mejor desarrollador que tu equipo de desarrollo.
  - Ser mejor en administración de sistemas que tu equipo de sistemas.
  - Saber más de almacenamiento que tu equipo de almacenamiento.
  - Saber más de red que tu equipo de redes.
  - Recomendar inversiones.
  - Etc, etc, etc....

Rendimiento de la base de datos

- Al DBA se le exige hacer magia, y no siempre es posible.
- Problemas de rendimiento:
  - Hardware inadecuado.
  - Storage inadecuado.
  - Mala parametrización.
  - Distribución de I/O no óptima.
  - Consultas no óptimas.
- El DBA sólo puede arreglar la mala parametrización y, a veces, la distribución de I/O, pero un buen DBA puede ayudar a la organización a arreglar las demás.
- Cuando el usuario tiene problemas con el tiempo de respuesta de la aplicación la culpa será siempre del DBA hasta que demuestre lo contrario.

Los desarrolladores

- Los desarrolladores suelen tener pocos conocimientos de optimización de consultas.
- Los desarrolladores están muy presionados para añadir nuevas características a un sistema y no tienen tiempo de optimizar las consultas.
- Al final las consultas mal escritas acaban en el entorno de producción y todos pagan las consecuencias.
- El DBA puede a) Localizar las consultas problemáticas y b) ayudar a los desarrolladores a mejorarlas.
- Si los desarrolladores cuentan con un entorno de desarrollo parecido a producción se facilita su labor de localizar las consultas problemáticas ANTES de llegar al entorno productivo.

Los sistemas

- Una base de datos no se ejecuta en el vacío: se ejecuta sobre un sistema operativo, utiliza la red, almacenamiento.
- Cualquier problema en uno de los componentes provocará un problema de rendimiento en la base de datos, pero la responsabilidad de arreglarlo siempre recae en el DBA.
- Es trabajo del DBA localizar el problema y, cuando es un elemento externo a la base de datos, convencer al responsable de que tiene un problema y ayudarle a arreglarlo.
- Denunciar ante la organización que el responsable de red o de almacenamiento está haciendo un mal trabajo no suele mejorar su ánimo de colaboración, y los usuarios seguirán quejándose.
- Los responsables de almacenamiento, cuando los hay, son especialmente talibanes en sus creencias y refractarios a cualquier razonamiento en contra de ellas, la única forma de razonar con ellos son los informes de rendimiento.

El hardware

- En las organizaciones las decisiones se toman con criterios de negocio y financiero.
- Las personas que van a tomar la decisión sobre comprar un nuevo hardware no tienen capacidad ni conocimiento para entender los problemas técnicos.
- Las personas que van a tomar la decisión sobre comprar un nuevo hardware planifican los gastos de la empresa con 12-18 meses de antelación como mínimo y no les gustan las sorpresas.
- El DBA debe tener siempre bajo control el nivel de uso del hardware y una previsión del nivel de uso a 18-24 meses basados en la evolución de uso del sistema en el pasado y las previsiones de evolución del negocio en el futuro.
- Cuando se detecta que el hardware va a necesitar ampliación en 18-24 meses, es necesario transmitir la información a gerencia de forma técnicamente sencilla y en un lenguaje que puedan entender. Es imprescindible justificar la inversión desde un punto de vista de negocio y financiero.
- Cuando se toman decisiones de negocio que van a impactar en el grado de uso de las bases de datos es preciso implicarse en el proceso y advertir de las inversiones en hardware y licencias que serán necesarias ANTES de que la organización tome la decisión.

Backup

- La tarea más importante de un DBA es el backup... y la segunda... y la tercera.
- No sirve de nada el rendimiento de la base de datos, ni la calidad de estos ni lo bonita que es la aplicación si cuando hace falta recurrir al backup éste falla.
- El punto anterior dice „CUANDO hace falta“, no dice „SI hace falta“.... tarde o temprano hará falta.
- Los fallos en la recuperación de un backup del sistema financiero suelen suponer la quiebra de la compañía en 6 meses.
- Los procedimientos de backup y recuperación deben estar probados y automatizados, recuperar un backup es una tarea realizada bajo mucha presión y cualquier posibilidad de error será aprovechada por el humano para liarla.
- Las pruebas de recuperación de backup periódicas son imprescindibles para garantizar la calidad del backup.

Backup

- Backup significa „backup físico“, los backup lógicos (export, dump...) no sirven para nada (salvo en DB2).
- Los DBA de verdad hacen los backup en caliente, siempre.
- Las bases de datos SIEMPRE deben estar en modo archivado del log.
- El archivado se debe evacuar al sistema de backup con la mayor frecuencia posible, eso reduce la pérdida de datos en caso de catástrofe completa.
- El backup y los datos nunca se deben guardar en los mismos discos.
- Cuanto más lejos estén físicamente el backup y los datos, mejor.
- Las catástrofes ocurren.
- „Sólo los paranoicos sobreviven“ (Andrew S. Groove)