Bienvenidos a la primera presentación dedicada a bases de datos distribuidas. Éstas son el resultado de combinar dos áreas de conocimiento de la ingeniería informática: las bases de datos y las redes de ordenadores.

Bases de datos distribuidas

- Sistemas distribuidos
- Bases de datos distribuidas

EIMT.UOC.EDU

Esta presentación se divide en dos secciones:

En la primera veremos qué es un sistema distribuido, sus ventajas y los retos que imponen.

Por su parte, en la segunda se presenta la definición de base de datos distribuida, las ventajas que de su uso esperamos obtener, así como las dificultades que existen en comparación a una base de datos centralizada.

Bases de datos distribuidas

- Sistemas distribuidos
- Bases de datos distribuidas

EIMT.UOC.EDU

Comenzamos, pues, la sección dedicada a sistemas distribuidos.

Sistemas distribuidos: definición

Un sistema distribuido es un conjunto de elementos de procesamiento autónomos (es decir, capaces de ejecutar programas), no necesariamente homogéneos, que están interconectados por una red de comunicaciones y que cooperan en la realización de las tareas que tienen asignadas.

- Qué se puede distribuir:
  - La ejecución del código
  - La funcionalidad
  - El almacenamiento de los datos
  - El control de la ejecución de las tareas asignadas

EIMT.UOC.EDU

Un sistema distribuido es un conjunto de elementos (o nodos), interconectados a través de una red, que son capaces de ejecutar programas. Estos elementos cooperan en la realización de una serie de tareas que tienen asignadas. Desde el punto de vista del usuario, el sistema se percibe como una sola unidad.

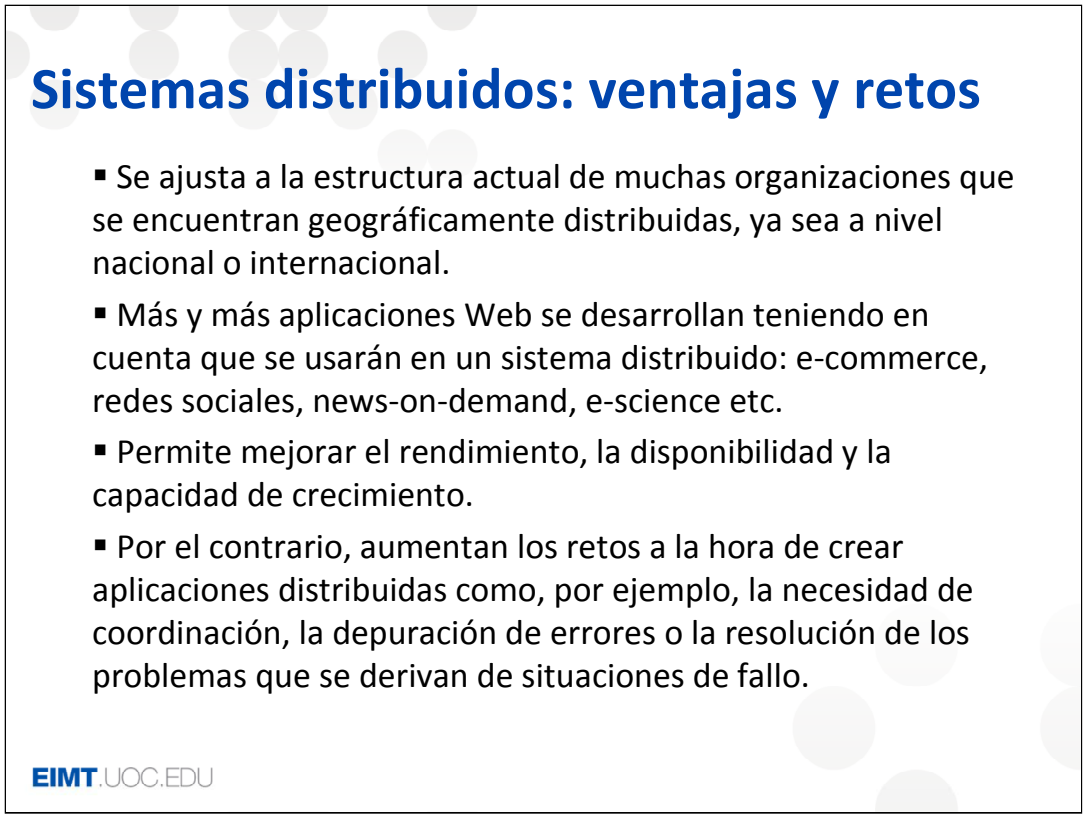
De la definición previa, es importante saber qué elementos son susceptibles de ser distribuidos. Básicamente son cuatro:

En primer lugar, la ejecución de código. Por ejemplo, una consulta sobre los datos se podría dividir en subconsultas que se ejecutarían en paralelo en diferentes nodos del sistema.

En segundo lugar, se pueden distribuir (o repartir) las funcionalidades del sistema entre los diferentes nodos. Por ejemplo, algunos nodos pueden recoger y procesar el resultado de las consultas, mientras que otros almacenan datos.

El tercer aspecto susceptible de ser distribuido es el almacenamiento de los datos. Por ejemplo, un nodo del sistema podría guardar los datos sobre clientes y sus pedidos, mientras que otro nodo podría almacenar los datos sobre los catálogos de productos.

Finalmente, el cuarto elemento susceptible de ser distribuido es el control de la ejecución de las tareas asignadas. Por ejemplo, podría existir un nodo en el sistema que tuviese la misión de localizar los nodos que guardan los datos pedidos en las consultas, de distribuir las consultas entre dichos nodos y de asegurarse que éstos devuelven los datos solicitados.



El uso de sistemas distribuidos es útil por diversos motivos.

En primer lugar, el procesamiento distribuido se ajusta a la estructura de muchas organizaciones que, o bien se encuentran geográficamente dispersas, o bien tienen un ámbito de actuación supranacional. En consecuencia, más y más aplicaciones se desarrollan teniendo en cuenta que se usarán en un entorno distribuido. Éste es el caso, entre otras, de aplicaciones de e-commerce. Además, en algunos casos, puede ser la única alternativa viable. Un escenario típico, en el caso de España, se relaciona con la Ley Orgánica de Protección de datos. Dicha ley obliga, entre otros aspectos, a que todas las compañías (incluidas las extranjeras) que operen en España almacenen los datos relativos a los usuarios de España en servidores que físicamente estén en España.

En segundo lugar, un sistema distribuido permite mejorar el rendimiento, la disponibilidad y la capacidad de crecimiento. El rendimiento mejora debido al reparto de tareas entre los nodos. Por otro lado, si una situación de fallo causa que algún nodo no esté operativo, ello no significa que globalmente el sistema no lo esté. Una plataforma de e-commerce (por ejemplo, Amazon o eBay) no puede estar caída durante horas, dado que eso implicaría una pérdida importantísima (y posiblemente intolerable) de ingresos y prestigio. Finalmente, un sistema distribuido (en comparación a uno centralizado) facilita el crecimiento de una organización de acuerdo a sus necesidades. Esto puede ser interesante en el caso de *start-ups* y de compañías que esperan ratios de crecimiento futuros. También es especialmente relevante en el caso de empresas que tienen que gestionar una cantidad masiva de datos que están distribuidos en cientos e incluso miles de nodos.

A pesar de las ventajas descritas, un sistema distribuido puede imponer ciertos retos, especialmente en el desarrollo de aplicaciones como sería, por ejemplo, la necesidad de coordinación, la depuración de errores o la resolución de los problemas derivados de situaciones de fallo.

Bases de datos distribuidas

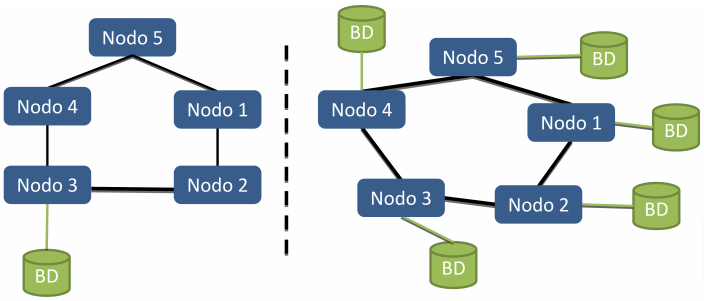
- Sistemas distribuidos
- Bases de datos distribuidas

EIMT.UOC.EDU

Una vez comprendido qué es un sistema distribuido, a continuación presentamos el concepto de base de datos distribuida, las situaciones que pueden conducir a su existencia, así como las ventajas que se derivan de su uso y los retos que imponen.

BD distribuidas: definición

Una base de datos distribuida es un conjunto de múltiples bases de datos, lógicamente interrelacionadas, que están distribuidas en una red de ordenadores. Las BD están gestionadas por un software específico que, entre otros, hace que la distribución sea transparente para los usuarios.



Base de datos centralizada Base de datos distribuida

EIMT.UOC.EDU

Como ya se ha comentado, las bases de datos distribuidas son el resultado de combinar dos conceptos principales: las bases de datos y las redes de ordenadores. A primera vista, ambos conceptos pueden parecer incompatibles pero, en realidad, son complementarios. Por una parte, las bases de datos promulgan la integración de los datos, y no necesariamente su centralización en un único punto de almacenamiento, como en ocasiones se asume. Por otra parte, las redes de ordenadores promueven el reparto de atribuciones entre los diferentes participantes que están conectados mediante una red de comunicaciones.

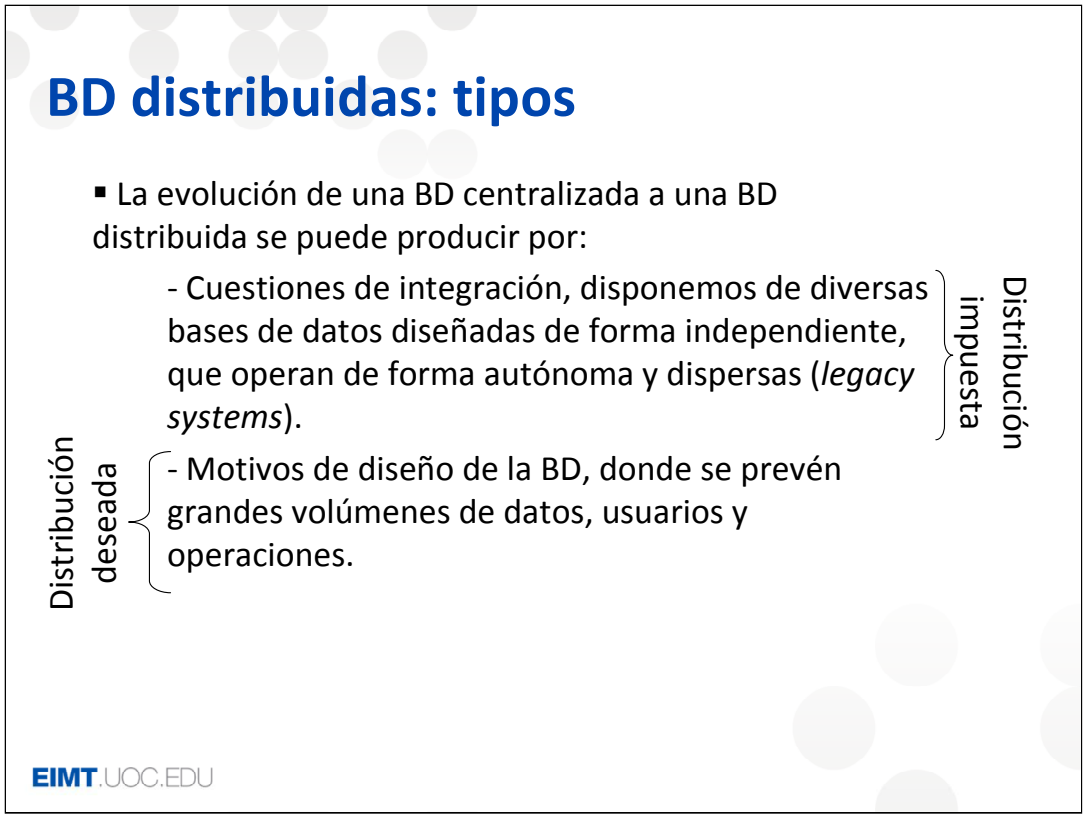
Una base de datos distribuida no es más que un conjunto de múltiples bases de datos (lógicamente interrelacionadas) que están distribuidas en varios ordenadores. Estas bases de datos están gestionadas por un software específico que hace la distribución de los datos transparente a los usuarios. Es decir, los usuarios tienen la ilusión de que la base de datos es única, o sea, que se comporta como si fuese centralizada.

La figura muestra dos ejemplos de sistemas distribuidos. En la parte de la izquierda existe un único punto de almacenamiento. Por lo tanto, a efectos prácticos, se trata de una base de datos centralizada. En cambio, a la derecha, podemos observar que existen diferentes nodos que almacenan datos. En consecuencia, podría tratarse de una base de datos distribuida.

De la definición dada, es necesario destacar los siguientes aspectos:

En primer lugar *las bases de datos están lógicamente interrelacionadas*. Este aspecto refleja que, desde un punto de vista semántico, las bases de datos representan un mismo dominio de aplicación. Es decir, la base de datos distribuida es algo más que una mera colección de ficheros inconexos, de tal manera que, al menos, es necesario proveer alguna interfaz común de acceso a los datos.

En segundo lugar, la noción de que *la distribución es transparente a los usuario*, tiene múltiples e importantes implicaciones. La transparencia se refiere a la capacidad del software que gestiona la base de datos distribuida de ocultar a los usuarios (entre los que están los desarrolladores de las aplicaciones) los detalles de implementación de la base de datos distribuida. Entre esos detalles, podemos incluir dónde se encuentran físicamente almacenados los datos, cómo se han distribuido y si hay datos replicados. Dependiendo de la capacidad de ocultar esos detalles, el desarrollo de aplicaciones será más o menos complejo.



Llegados a este punto, es importante que nos formulemos la siguiente pregunta: *¿Qué situaciones pueden conducir a la existencia de una base de datos distribuida?*

Principalmente son dos, en función de si la distribución es impuesta o deseada.

La primera situación tiene que ver con la existencia de bases de datos preexistentes (los denominados *legacy systems*). Pensemos, por ejemplo, en el caso de bases de datos desarrolladas por diferentes departamentos de una misma empresa, o en situaciones de fusión de empresas.

Por su parte, la segunda se corresponde a una situación en donde diseñamos una base de datos desde cero, y se ha decidido que ésta tiene que ser distribuida. Por ejemplo, esto puede ser consecuencia del volumen de datos a almacenar, por la ubicación de los usuarios que van a usar la base de datos y por el volumen (y tipología) de las operaciones que éstos van a realizar.

Los métodos y técnicas empleadas para intentar proporcionar una visión integrada de la base de datos en cada situación son muy diferentes. Desde un punto de vista de diseño, la primera se corresponde con un enfoque *botton-up* (en otras palabras, la construcción de la base de datos distribuida se realiza a partir de las bases de datos preexistentes). La segunda situación se corresponde con un enfoque *top-down*. En este caso, podemos seguir las mismas estrategias de diseño de una base de datos centralizada. A estas estrategias se deberían añadir estrategias que ayuden a tomar decisiones relativas a cómo dividir la base de datos (y los datos que ésta contenga), dónde almacenar los datos y si es necesario replicar datos.

La base de datos distribuida que se deriva de la primera situación puede no ajustarse a la definición que hemos dado de base de datos distribuida. Por ello, en la literatura, este tipo de bases de datos se conocen con otras denominaciones (por ejemplo, como bases de datos federadas, multibases de datos etc.).

A nosotros nos interesan las bases de datos distribuidas que surgen como consecuencia de una decisión de diseño, es decir, nuestro foco de atención es la situación que hemos descrito en segundo lugar.

BD distribuidas: ventajas y retos

- Una BD distribuida hace transparente al usuario conceptos tales como la red y el hecho que el esquema de la BD (y los datos) esté disperso (y quizá parcialmente replicado), facilitando de esta manera el desarrollo de aplicaciones.
- Se consigue un aumento del rendimiento, fiabilidad y disponibilidad de los datos, gracias a la reducción del acceso remoto a los datos y a la ejecución de operaciones distribuidas.
- Facilita la extensión de la BD gracias a su descomposición en diferentes componentes.
- A pesar de las ventajas, hay que tener el en cuenta dificultades como el diseño de la BD, el mantenimiento de la integridad de los datos (por ejemplo, la coherencia de datos replicados), la coordinación para la ejecución de las operaciones, la resolución de situaciones de fallo, etc.

EIMT.UOC.EDU

El uso de bases de datos distribuidas presenta múltiples ventajas no exentas de dificultades. En esencia, son similares a las de los sistemas distribuidos, pero en este caso, la discusión se aborda desde la perspectiva de la gestión de los datos.

La base de datos distribuida (más concretamente el software que la gestiona) hace transparente al usuario conceptos tales como la red y el hecho que la base de datos esté dispersa (y quizá parcialmente replicada). Esto facilita el desarrollo de aplicaciones. A pesar de esta declaración de buenas intenciones, no siempre es posible proporcionar un nivel de transparencia total.

El uso de bases de datos distribuidas, con respecto a una centralizada, permite aumentar el rendimiento, la fiabilidad y la disponibilidad de los datos, gracias a la ejecución de operaciones distribuidas y a la reducción del acceso remoto a los datos.

También facilita la extensión de la base de datos gracias a su descomposición en diferentes componentes. Esto es especialmente cierto en el caso de bases de datos NoSQL. En bases de datos relacionales, la extensión no es tan sencilla, y si acontece, se debe planificar cuidadosamente.

A pesar de las ventajas, el desarrollo de base de datos distribuidas debe abordar ciertas dificultades. Básicamente, garantizar las funcionalidades que clásicamente se asocian a un sistema gestor de bases de datos es más complicado, debido a la distribución de los datos, de las tareas y del control de la ejecución de dichas tareas entre los nodos que forman la base de datos distribuida.

Bases de datos distribuidas

- Sistemas distribuidos
- Bases de datos distribuidas

EIMT.UOC.EDU

Aquí finaliza la primera presentación dedicada a bases de datos distribuidas donde hemos examinado las características más relevantes asociadas a los sistemas distribuidos y a las bases de datos distribuidas.

Referencias

C. Coronel, S. Morris & P. Rob (2013). *Database Systems: Design, Implementation and Management 10e*. Course Technology, Cengage Learning.

L. Liu & M.T. Özsu (Eds.) (2009). *Encyclopedia of Database Systems*. Springer.

M.T. Özsu & P. Valduriez (2011). *Principles of Distributed Systems*. 3rd edition. Springer.

O. Romero & M. Oliva (2012). Distributed Databases. Material docente UOC, asignatura Arquitectura de bases de datos.

P.J. Sadalage & M. Fowler. (2013). *NoSQL Distilled. A brief Guide to the Emerging World of Polyglot Persistence*, Pearson Education.

EIMT.UOC.EDU

Esperamos que hayáis disfrutado y aprendido con este vídeo. A continuación encontraréis algunas referencias que os permitirán profundizar más en los conceptos que hemos tratado.

Que tengáis un buen día.