UD 1. Sistemas de almacenamiento de la información

Introducción

Según el Diccionario de la Real Academia Española (DRAE), informática es el "Conjunto de conocimientos científicos y técnicas que hacen posible el tratamiento automático de la información por medio de ordenadores". El diccionario de Cambridge University Press define *information technology (IT)* como "la ciencia y la actividad de utilizar ordenadores y otras herramientas electrónicas para almacenar y enviar información". En ambos casos, el objeto de la disciplina es la información, y el objetivo su gestión.

Definimos sistema de información como el conjunto de procedimientos y funciones dirigidos a la recogida, elaboración, evaluación, almacenamiento, recuperación, condensación y distribución de informaciones dentro de una organización.

Antes de que surgieran las bases de datos el procesamiento automatizado de información se hacía mediante *ficheros*. Las aplicaciones eran orientadas al proceso (el esfuerzo se enfocaba al tratamiento que los datos recibían en una aplicación concreta). Los ficheros se diseñaban a medida para cada sistema de información, sin que existiera un formato común.

Introducción

Un mainframe da servicio a muchos usuarios de forma simultánea.

Esta aproximación no contemplaba la gestión de la información a medio o largo plazo. Una organización disponía de varias aplicaciones que, en algunos casos, trataban la misma información (ejemplo: el software utilizado por el departamento de recursos humanos debía gestionar un fichero con datos de empleados, mientras la aplicación de contabilidad mantenía otro fichero distinto con los mismos datos organizados de otra forma). Surgían los primeros problemas:

- Redundancia de datos (duplicidad innecesaria de la información).
- Mal aprovechamiento del espacio de almacenamiento.
- Aumento en el tiempo de proceso.

Introducción

- Inconsistencia de información debida a la redundancia (si un dato cambiaba en el fichero de una aplicación, no cambiaba en los demás).
- Aislamiento de la información (imposibilidad de transferirla a otros programas a no ser que se desarrollara un software de migración específico).

Había, en definitiva, una gran falta de flexibilidad originada en la dependencia total de la estructura física de los datos.

Ficheros

Las aplicaciones gestoras de bases de datos se encargan de configurar una estructura óptima de almacenamiento de información con mínima intervención por parte del usuario. No obstante, es interesante completar la perspectiva histórica con una breve descripción teórica sobre organización de ficheros.

Tipos de ficheros según su estructura de almacenamiento

En relación con su contenido, encontramos los siguientes tipos básicos de ficheros:

- **Texto plano:** Almacenan secuencias de caracteres correspondientes a una codificación determinada (ASCII, Unicode, EBCDIC, etc.). Son legibles mediante un software de edición de texto como el Bloc de Notas de Windows o el Vi de Linux.

  Ejemplos: .txt, .csv, .htm, .html, .xml o .rss.

- **Binarios**: Contienen información codificada en binario para su procesamiento por parte de aplicaciones. Su contenido resulta ilegible en un editor de texto.

  Ejemplos: .exe, .pdf, .docx, .xlsx, .pptx, .jpg, .gif, .mp3, .avi, .mkv, .dll.

Tipos de ficheros según su estructura de almacenamiento

Cuando se utilizan ficheros de texto plano para almacenar información se pueden clasificar de acuerdo a su organización interna:

- **Secuenciales:** La información se escribe en posiciones físicamente contiguas. Para acceder a un dato hay que recorrer todos los anteriores.

```
00567465A#Susana#Sanz#González#666777888#12332145T#Antonio#Alba#Moreno#61234567
8#23415178B#Paula#Gómez#Salas#689908765#60986754P#Luis#Raya#Delgado#678492011
```

- **De acceso directo o aleatorio:** Cada línea de contenido se organiza con unos tamaños fijos de dato. Se puede acceder directamente al principio de cada línea.

```
00567465ASusana    Sanz      González    666777888
12332145TAntonio   Alba      Moreno      612345678
23415178BPaula     Gómez     Salas       689908765
60986754PLuis      Raya      Delgado     678492011
```

Tipos de ficheros según su estructura de almacenamiento

En esta ocasión cada contacto ocupa una línea del fichero (al final de cada una el sistema operativo incluirá uno o dos caracteres de salto de línea invisibles para el usuario), y cada dato utiliza un número de caracteres fijo, aunque no lo ocupe totalmente.

Como todos los clientes ocupan el mismo espacio en el fichero, podemos acceder fácilmente a cualquiera de ellos multiplicando la posición en la que se encuentra menos una por el número de caracteres que mide cada línea.

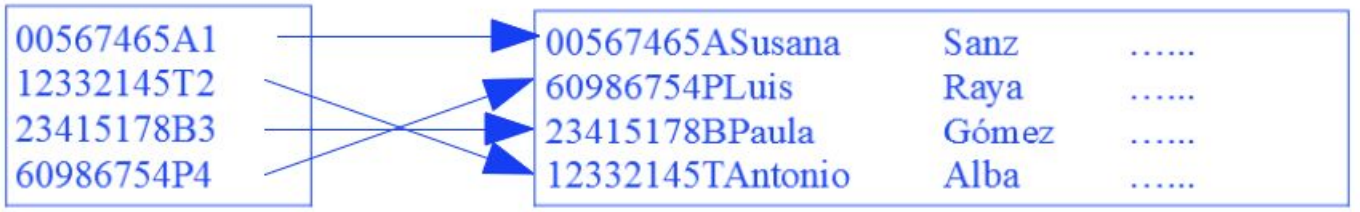
La contrapartida a esta facilidad de posicionamiento es que el tamaño del fichero crece considerablemente respecto a su versión secuencial.

Tipos de ficheros según su estructura de almacenamiento

- **Indexados:** Generalmente en un fichero de acceso aleatorio la información se almacena en el orden en que se da de alta. Incluso aunque se consiguiera introducir dicha información de acuerdo a algún criterio de ordenación concreto, en algunas ocasiones es útil poder ordenarla por varios criterios distintos. información de acuerdo a algún criterio de ordenación concreto, en algunas ocasiones es útil poder ordenarla por varios criterios distintos.

En el ejemplo anterior es posible que necesitemos un listado de clientes ordenado por NIF. Para dar solución a este problema se creó la organización indexada, que consiste en la existencia de uno o varios archivos adjuntos que ordenan el dato (llamado clave) por el que se desea ordenar el fichero y lo relacionan con la localización de la línea correspondiente.

Tipos de ficheros según su estructura de almacenamiento



Aunque se utilice en este caso para simplificar el ejemplo, generalmente el acceso a cada posición no lo marca el número de línea, sino un puntero a la celda de memoria correspondiente.

Crea manualmente el fichero índice que ordene
el fichero clientes original por apellido.

Bases de datos

La evolución lógica de los problemas derivados del uso de ficheros fue estandarizar el acceso a la información, de modo que un diseño físico concreto sirviera para todas las aplicaciones de una organización. Este nuevo enfoque se centraba en los datos y no en el proceso, es decir, se estructuraba el almacenamiento de dichos datos con independencia de las aplicaciones que los fueran a utilizar. Se eliminaba su redundancia y se favorecía la transferencia de información de aplicaciones. Aparecía el concepto de bases de datos.

Definición

Adoración de Miguel y Mario Piattini ofrecen la siguiente definición de **base de datos**:

Colección o depósito de datos integrados, almacenados en soporte secundario (no volátil) y con redundancia controlada. Los datos, que han de ser compartidos por diferentes usuariosy aplicaciones, deben mantenerse independientes de ellos y su definición (estructura de la BD), única y almacenada junto con los datos, se ha de apoyar en un modelo de datos, el cual ha de permitir captar las interrelaciones y restricciones existentes en el mundo real. Los procedimientos de actualización y recuperación, comunes y bien determinados, facilitarán la seguridad del conjunto de datos.

Evolución y tipos de bases de datos

Un **modelo de bases de datos** es la arquitectura mediante la que se almacena e interrelaciona la información que se va a gestionar. Para la clasificación habitual de bases de datos se tiene en cuenta la siguiente evolución histórica:

- **Sistemas de archivos (1950-1960):** Antes de la aparición de los SGBD, los datos se almacenaban en ficheros almacenados.
- **1ª generación de bases de datos (1960):**
  - *Modelo de datos jerárquico:* Los campos de los registros están organizados en niveles, con estructura en árbol donde cada nodo del mismo nivel corresponden a campos y las ramas a registros.
  - *Modelo de datos en red:* Corresponde a una estructura de grafo, desapareciendo el concepto de jerarquía.

Evolución y tipos de bases de datos

- **2ª generación de bases de datos (1970-actualidad):**
  - *Bases de datos relacionales:* Pueden resolver, mejor que otras organizaciones, las dificultades de redundancia y no integración de datos. Suprimen la jerarquía de campos, pudiéndose utilizar cualquier de ellos como clave de acceso.
  - La organización relacional se caracteriza porque las tablas de la BD tienen estructura de matriz o tabla bidimensional, donde las filas son los registros y las columnas los campos.
  - En 1979 se introdujo un nueva versión extendida denominada RM/T y posteriormente en 1990 RM/V2. La importancia del modelo relacional condujo a dos grandes desarrollos:
  - El desarrollo de un lenguaje de consultas estructurado estándar *SQL*.
  - SGBD relacionales durante los años 80, de éxito, como DB2 y ORACLE.
  - Hoy en día, la mayoría de los SGBD soportan el modelo relacional.

Evolución y tipos de bases de datos

- **3ª generación de los SGBD:** Como respuesta a la creciente complejidad de las aplicaciones que requieren bases de datos, han surgido nuevos modelos de datos y SGBD que los soportan:
  - *Modelo orientado a objetos:* Aplica a los datos el paradigma de la orientación a objetos. Irrumpió con fuerza en los años noventa debido a las nuevas necesidades de almacenamiento de las bases de datos relacionales (imágenes, documentos, ficheros de audio y vídeo). Por ejemplo: Versant, Objectivity, etc.
  - *Modelo de datos objeto-relacional:* En los últimos años los fabricantes de bases de datos relacionales han incorporado a su software diversas capacidades de las bases de datos orientadas a objetos, creando modelos híbridos con base relacional (Oracle, Microsoft SQL Server, PostgreSQL, etc.).
  - *Modelo orientado al documento:* Gestionan datos provenientes de documentos previamente estructurados, generalmente de lenguajes de marcas (XML, JSON).

Sistemas gestores de bases de datos

Un **sistema gestor de bases de datos (SGBD)** (DBMS DataBase Management System) es el software que el fabricante pone a disposición del usuario para manejar sus bases de datos. Nuevamente, De Miguel y Piattini (1993) nos definen el término con más detalle:

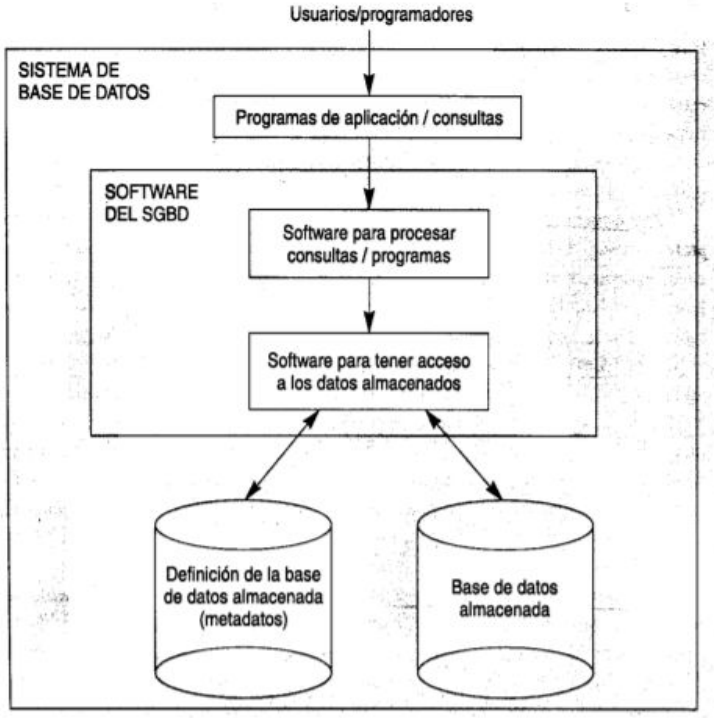
Un conjunto coordinado de programas, procedimientos, lenguajes, etc., que suministra, tanto a los uarios no informáticos como a los analistas, programadores, o al administrador, los medios necesarios para describir, recuperar y manipular los datos almacenados en la base, manteniendo su seguridad.

En el mercado hay una amplia tipología de SGBD que corresponde con el modelo de base de datos subyacente.

Sistemas gestores de bases de datos

Se denomina **sistema de base de datos** al conjunto formado por:

- la base de datos.
- el SGBD
- y otros posibles programas de aplicación /consulta del usuario.

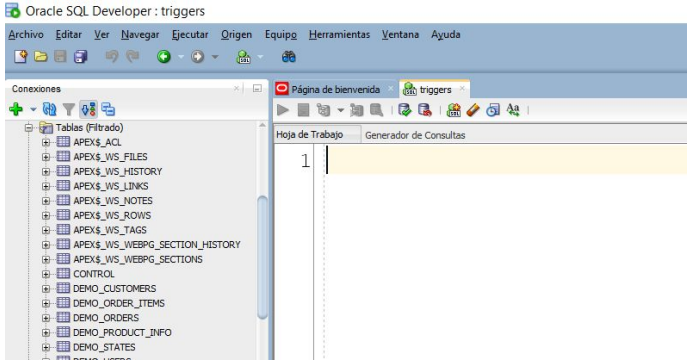


Componentes del SGBD

Generalizando, podemos encontrar la siguiente enumeración de **componentes** en la mayoría de los SGBD:

- **Datos:** Almacenados de forma eficiente en ficheros del sistema operativo.
- **Herramientas de acceso a los datos:** Un lenguaje de programación mediante el que los usuarios técnicos puedan crear, leer y modificar la información, así como un diccionario de datos que albergue los metadatos, es decir, la información sobre el diseño de cada base de datos. Como mínimo, se ofrecerá una interfaz de línea de comandos mediante la que acceder a estas herramientas.
- **Utilidades:** Herramientas adicionales para gestión de backups, estadísticas, tareas programadas, mantenimiento de usuarios, grupos y permisos, etc.
- **Entornos gráficos:** Simplifican la gestión del SGBD

y sirven como alternativa a la línea de comandos.



Funciones del SGBD

A pesar de la gran variedad de modelos y soluciones comerciales, podemos enumerar una serie de **funciones** comunes a un gran número de SGBD:

- Recuperar y modificar la información de los ficheros que conforman la base de datos de forma transparente para el usuario.
- Garantizar la integridad de los datos, impidiendo inconsistencias semánticas.
- Ofrecer un lenguaje de programación mediante el que interaccionar con la información.
- Proveer el diccionario de datos.
- Solucionar los conflictos derivados de accesos concurrentes a la información.
- Gestionar transacciones, garantizando la unidad de varias instrucciones de escritura relacionadas entre sí.
- Incluir utilidades de backup.
- Proporcionar mecanismos de seguridad para evitar accesos y operaciones indebidos.

Estructura del SGBD

Un SGBD cuenta con una arquitectura a través de la que se simplifica a los diferentes usuarios de la base de datos su labor. El objetivo fundamental es separar los programas de aplicación de la base de datos física.

Encontrar un estándar para esta arquitectura no es una tarea sencilla, aunque los tres estándares que más importancia han cobrado en el campo de las bases de datos son *ANSI/SPARC/X3*, CODASYL y ODMG (éste sólo para las bases de datos orientadas a objetos). Tanto ANSI (EEUU), como ISO (Resto del mundo), son el referente en cuanto a estandarización de bases de datos, conformando un único modelo de bases de datos.

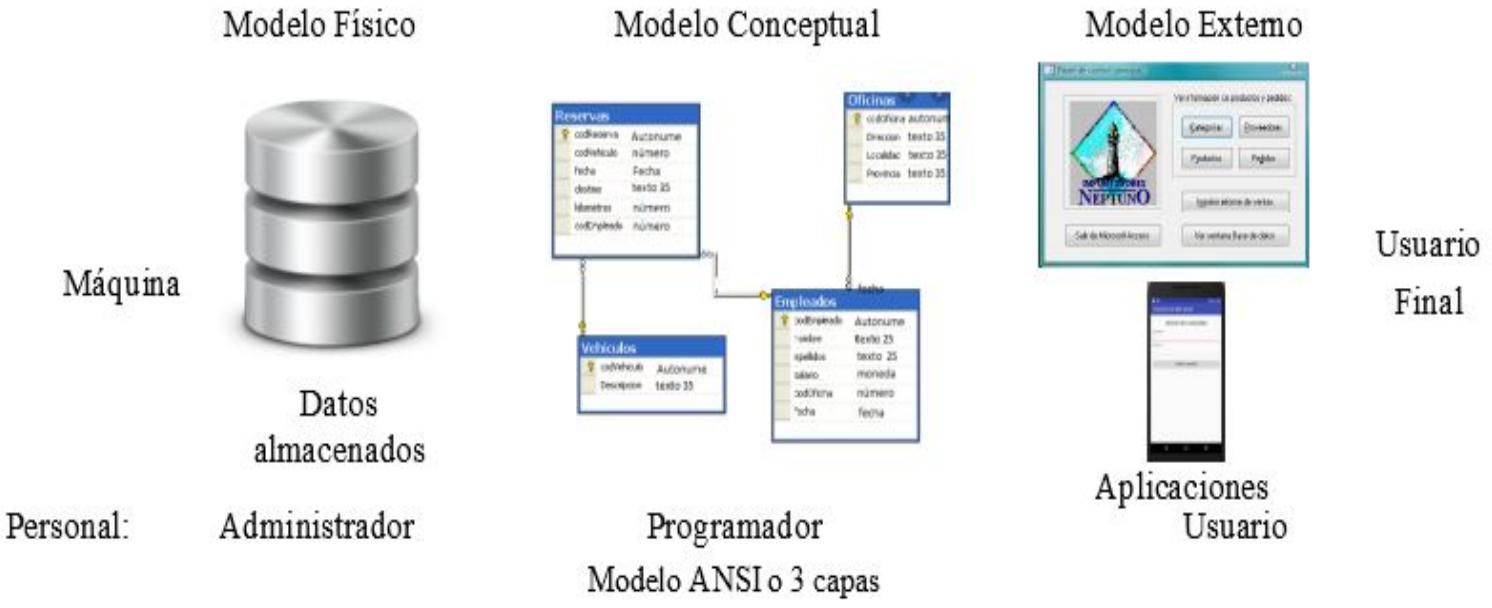
Estructura del SGBD

La arquitectura propuesta proporciona tres niveles de abstracción:

- **Nivel interno o físico:** Es la estructura de los datos tal cual se almacenan en las unidades de disco. La correspondencia entre la estructura lógica y la física se almacena en la base de datos en los metadatos (tipos de registros, longitud,…).
- **Nivel lógico o conceptual:** Describe la estructura completa de la base de datos a través de un esquema que detalla las entidades, atributos, relaciones, operaciones de los usuarios y restricciones. Los detalles relacionados con las estructuras de almacenamiento se ocultan, permitiendo realizar una abstracción a más alto nivel.
- **Modelo externo:** Son las vistas a las que acceden los usuarios mediante aplicaciones a la base de datos. Cada tipo de usuario o grupo de ellos verá sólo la parte de la base de datos que le interesa, ocultando el resto.

Estructura del SGBD

Para una base de datos, sólo existirá un único esquema interno, un único esquema conceptual y podrían existir varios esquemas externos definidos para uno o varios usuarios.



El lenguaje SQL

SQL es un Lenguajes de Consulta Estructurado que engloba instrucciones tanto de definición como de manipulación de datos, e incluso instrucciones de administración.

Se trata de un lenguaje para administrar, almacenar y recuperar información, y es utilizado por la mayoría de los SGBD actuales. Habitualmente incluye las siguientes categorías de sentencias:

- **DDL** (Data Definition Language o Lenguaje de Definición de Datos): Crear, alterar o borrar tablas, columnas o vistas de la base de datos. Los comandos que utiliza son: *CREATE, DROP y ALTER.*
- **DML** (Data Modification Language o Lenguaje de Modificación de Datos): Manipulación de datos en esquemas ya existentes. Los comando que utiliza son: *SELECT, INSERT, UPDATE y DELETE.*
- **DCL** (Data Control Language o Lenguaje de Control de Datos): Control de acceso a datos en la base de datos. Algunos comandos son: *GRANT y REVOKE.*

Todas las sentencias SQL deben tener una sintaxis adecuada, para que se puedan ejecutar correctamente. Para describir dicha sintaxis se emplean diagramas sintácticos. Las palabras en mayúscula son palabras reservadas del lenguaje *(SELECT (seleccionar), ALL (todo), DISTINCT (distinto), FROM (desde), WHERE (donde).*

Usuarios del sistema de bases de datos

El personal implicado en el uso de un sistema de bases de datos que nos encontramos es:

- **Administradores de bases de datos (DBA):** Persona que supervisa y controla los recursos del sistema de base de datos, es decir, la BD en sí, el SGBD y el software relacionado. Algunas funciones típicas del DBA son:
  - Autorizar el acceso a la BD.
  - Coordinar y vigilar su empleo.
  - Instalación de la BD.
  - Optimizar el acceso a los recursos.
  - Sacar copias de seguridad de los datos.
- **Diseñadores de bases de datos:** Se encargan de identificar los datos que se almacenarán en la BD y de elegir las estructuras apropiadas para representar y almacenar dichos datos. Estas tareas se realizan antes de que se implemente la BD. Tal vez asuman la responsabilidad de DBA cuando se termine el diseño de la BD.
- **Analistas de sistemas y programadores de aplicaciones:** Determinan los requisitos de los usuarios finales y desarrollan especificaciones para programas de aplicación que satisfagan esos requisitos. Deben conocer a la perfección las capacidades del SGBD.
- **Usuarios finales:** Son las personas que necesitan tener acceso a la BD para utilizar su potencial.

Clasificación de SGBD

- Según modelo lógico:
  - Jerárquico
  - En red
  - Relacional
  - Objeto-relacional
  - Orientado a objetos
- Según número de sitios:
  - Centralizado
  - Distribuido
- Según tipo de datos:
  - Relacionales
  - XML
  - Objeto-relacionales
  - Orientados a objetos
- Según número de usuarios:
  - Monousuario
  - Multiusuario
- Según Ámbito de aplicación
  - Propósito general
  - Propósito específico
- Según lenguajes soportados:
  - SQL
  - NoSQL

SGBD Libres y comerciales

Sistemas Gestores de Bases de Datos Comerciales.		
SGBD	**Descripción**	**URL**
ORACLE	Reconocido como uno de los mejores a nivel mundial. Es multiplataforma, confiable y seguro. Es Cliente/Servidor. Basado en el modelo de datos Relacional. De gran potencia, aunque con un precio elevado hace que sólo se vea en empresas muy grandes y multinacionales. Ofrece una versión gratuita **Oracle Database Express Edition 11*g* Release 2.**	Oracle
MYSQL	Sistema muy extendido que se ofrece bajo dos tipos de licencia, comercial o libre. Para aquellas empresas que deseen incorporarlo en productos privativos, deben comprar una licencia específica. Es Relacional, Multihilo, Multiusuario y Multiplataforma. Su gran velocidad lo hace ideal para consulta de bases de datos y plataformas web.	MySQL
DB2	Multiplataforma, el motor de base de datos relacional integra XML de manera nativa, lo que IBM ha llamado pureXML, que permite almacenar documentos completos para realizar operaciones y búsquedas de manera jerárquica dentro de éste, e integrarlo con búsquedas relacionales.	DB2
Microsoft SQL SERVER	Sistema Gestor de Base de Datos producido por Microsoft. Es relacional, sólo funciona bajo Microsoft Windows, utiliza arquitectura Cliente/Servidor. Constituye la alternativa a otros potentes SGBD como son Oracle, PostgreSQL o MySQL.	Microsoft SQL Server 2008
SYBASE	Un DBMS con bastantes años en el mercado, tiene 3 versiones para ajustarse a las necesidades reales de cada empresa. Es un sistema relacional, altamente escalable, de alto rendimiento, con soporte a grandes volúmenes de datos, transacciones y usuarios, y de bajo costo.	Sybase

Otros SGBD comerciales importantes son: DBASE, ACCESS, INTERBASE y FOXPRO.

SGBD Libres y comerciales

Sistemas Gestores de Bases de Datos Libres.		
SGBD	**Descripción**	**URL**
MariaDB	MariaDB es un sistema de gestión de bases de datos derivado de MySQL con licencia GPL (General Public License). Es desarrollado por Michael (Monty) Widenius —fundador de MySQL—, la fundación MariaDB y la comunidad de desarrolladores de software libre.	MariaDB
PostgreSQL	Sistema Relacional Orientado a Objetos. Considerado como la base de datos de código abierto más avanzada del mundo. Desarrollado por una comunidad de desarrolladores que trabajan de forma desinteresada, altruista, libre y/o apoyados por organizaciones comerciales. Es multiplataforma y accesible desde múltiples lenguajes de programación.	PostGreSQL
Firebird	Sistema Gestor de Base de Datos relacional, multiplataforma, con bajo consumo de recursos, excelente gestión de la concurrencia, alto rendimiento y potente soporte para diferentes lenguajes.	Firebird
Apache Derby	Sistema Gestor escrito en Java, de reducido tamaño, con soporte multilenguaje, multiplataforma, altamente portable, puede funcionar embebido o en modo cliente/servidor.	Apache Derby
SQLite	Sistema relacional, basado en una biblioteca escrita en C que interactua directamente con los programas, reduce los tiempos de acceso siendo más rápido que MySQL o PostGreSQL, es multiplataforma y con soporte para varios lenguajes de programación.	SQLite

Bases de datos centralizadas y distribuidas

Utilizando como criterio la ubicación física de la información, podemos diferenciar entre dos grandes tipos de bases de datos:

- **Centralizadas:** La base de datos reside en una sola máquina, típicamente el servidor de la base de datos.
- **Distribuidas:** La información se reparte por distintos servidores, generalmente alejados físicamente. Un ejemplo sería la base de datos de una compañía de seguros, concebida a partir de los datos de la oficina central y de los de todas su sucursales. Su implantación exige hacer un fuerte hincapié en aspectos de networking y seguridad.

Bases de datos distribuidas

La necesidad de integrar información de varias fuentes y la evolución de las tecnologías de comunicaciones, han producido cambios muy importantes en los sistemas de bases de datos. La respuesta a estas nuevas necesidades y evoluciones se materializa en los sistemas de bases de datos distribuidas.

- **Base de datos distribuida (BDD):** es un conjunto de múltiples bases de datos lógicamente relacionadas las cuales se encuentran distribuidas entre diferentes nodos interconectados por una red de comunicaciones.
- **Sistema de bases de datos distribuida (SBDD):** es un sistema en el cual múltiples sitios de bases de datos están ligados por un sistema de comunicaciones, de tal forma que, un usuario en cualquier sitio puede acceder los datos en cualquier parte de la red exactamente como si los datos estuvieran almacenados en su sitio propio.
- **Sistema gestor de bases de datos distribuida (SGBDD):** es aquel que se encarga del manejo de la BDD y proporciona un mecanismo de acceso que hace que la distribución sea transparente a los usuarios. El término transparente significa que la aplicación trabajaría, desde un punto de vista lógico, como si un solo SGBD ejecutado en una sola máquina, administrará esos datos.

Fragmentación, replicación y distribución de datos

Para responder a la pregunta de cómo distribuir los datos debemos conocer:

- **Fragmentación:** Cómo dividir los datos.
  - Horizontal: Separamos filas.
  - Vertical: Separamos campos.
- **Replicación:** Mantener una o varias copias de los datos.
  - Facilita la distribución de la carga.
  - Mejora la disponibilidad.
  - Sirve de copia de seguridad.