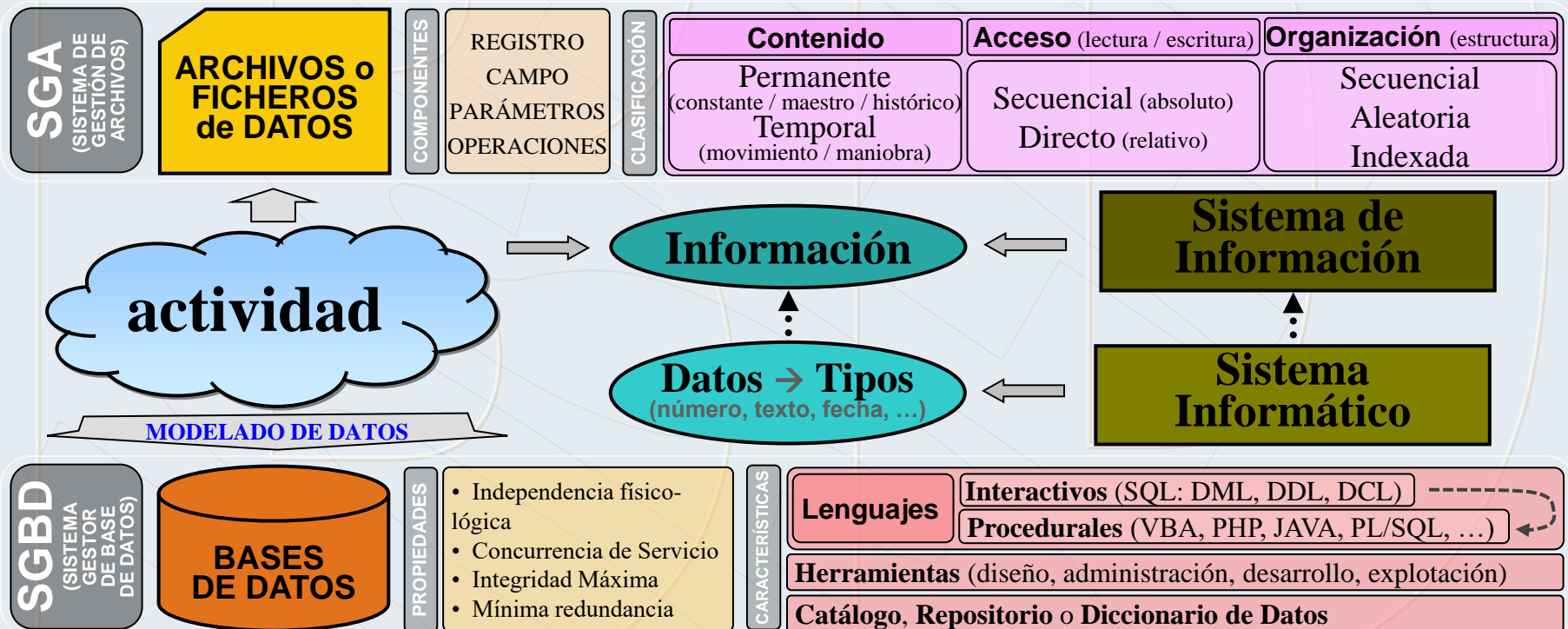


# Sistemas de Almacenamiento de Datos

## TEMA 1.1

# Sistemas de Almacenamiento de Datos.

Los **Archivos o Ficheros de Datos** y las **Bases de Datos** son sistemas de almacenamiento de datos en soportes de diverso tipo (magnético, óptico, u otros) que permiten disponer de información de forma permanente, masiva y prácticamente ilimitada.



## Archivos o Ficheros de Datos.

Un **Archivo** o **Fichero de Datos** es una estructura de datos formada por un número indeterminado de ocurrencias de registros lógicos del mismo tipo referentes a una entidad (persona, empresa, departamento, etc).

Un **Registro** es un conjunto de datos referentes a una entidad que puede ser de tamaño fijo o variable. El registro lógico se corresponde con la visión humana de los datos, y el registro físico se corresponde con la distribución de los datos en el soporte de almacenamiento. Una **ocurrencia** es el contenido de información de un registro lógico (un caso particular de valores de los datos del registro).

Un **Campo** es una característica o propiedad almacenable de una entidad, que puede ser de longitud fija o variable. Cada campo es de un tipo determinado, según el valor del dato asociado (clave o indicativo, indicador o flag, cantidad, literal, fecha, código, puntero, ....). El campo **clave** es un campo que identifica unívocamente, con un valor diferente, cada ocurrencia de registro.

Registro 1

Campo 1	Campo 2	Campo 3		Campo K
<b>CLAVE</b>				
<b>DNI</b>	<b>Nombre</b>	<b>Edad</b>	<b>'''</b>	<b>Fecha_nac</b>
"22345643H"	"Juan"	18		12/07/1985

Registro 2

Campo 1	Campo 2	Campo 3		Campo K
<b>CLAVE</b>				
<b>DNI</b>	<b>Nombre</b>	<b>Edad</b>	<b>'''</b>	<b>Fecha_nac</b>
"974451239A"	"Paula"	18		21/11/1985

Los archivos o ficheros requieren algún soporte de almacenamiento, de un tipo u otro, para cuya elección se puede utilizar el valor de determinados **parámetros** como el volumen (tamaño actual), crecimiento (tamaño previsible), actividad (número de accesos de recuperación o modificación), u otros.

El manejo de archivos o ficheros se lleva a cabo por medio de **operaciones** para creación, eliminación, apertura, cierre, borrado, alta (inserción, anexión), baja (lógica, física), recuperación, modificación, mezcla, concatenación, partición, copia o duplicado, ordenación o clasificación, reorganización, restauración, ejecución, renombrado. Las operaciones con archivos o ficheros se pueden realizar en modo gráfico, generalmente desde el entorno del sistema operativo, o en modo texto, desde la línea de comandos del sistema operativo, o a través de instrucciones de programa.

# Archivos o Ficheros de Datos.

## Clasificación

### Tipo de Contenido

- **Archivo permanente:** contiene información sujeta a pocos cambios pero que debe estar continuamente actualizada durante un tiempo ilimitado (normalmente, el tiempo de vida de una aplicación informática).
  - **Archivo constante:** contienen información actualizada que no se modifica y sólo se consulta. Cuando este archivo ocupa poco espacio se suele cargar en tablas de memoria principal y recibe el nombre de archivo de tablas (constantes físicas, coeficientes, ...).
  - **Archivo maestro o de situación:** contiene información permanentemente al día que refleja el estado o situación actual de una entidad (clientes, empleados, ...).
  - **Archivo histórico:** contiene datos acumulados sobre la historia de las actualizaciones que han sufrido los datos (con fines estadísticos o de consulta por regla general).
- **Archivo temporal:** almacena datos transitorios durante un tiempo limitado (normalmente la ejecución de un proceso).
  - **Archivo de movimiento o de transacciones:** contiene información que se utiliza para actualizar el archivo maestro, destruyéndose cuando deja de ser útil, una vez que se realiza dicha operación. Los registros se denominan movimientos o transacciones y son de tres clases (altas, bajas y modificaciones).
  - **Archivo de maniobra o de trabajo o intermedio o transitorio:** almacena datos auxiliares.

**Tipo de Acceso** determina la forma en que se va a leer y escribir la información.

- **Acceso Secuencial:** los datos se acceden (de forma absoluta) secuencialmente en un orden preestablecido.
- **Acceso Directo:** los datos se acceden (de forma relativa) directamente sin orden predeterminado.

**Tipo de Organización** ligada a la estructura física del soporte de almacenamiento que contiene los archivos.

- **Organización Secuencial:** los registros ocupan posiciones consecutivas en el soporte de almacenamiento. Hay que recorrer todos los registros para acceder a uno concreto.
- **Organización Aleatoria o al Azar:** los registros ocupan posiciones aleatorias en el soporte de almacenamiento sin orden físico. Se requiere algún mecanismo de direccionamiento, que consume tiempo y espacio.
- **Organización Indexada:** Se basa en la utilización de índices en archivos de organización secuencial, con el fin de subsanar los inconvenientes de la organización secuencial y aleatoria.

# Bases de Datos.

Una **base de datos** es un conjunto de datos referentes a determinadas entidades recopilados y agrupados con un fin determinado (banco de datos) y organizado según una determinada estructura orientada a su manipulación en un entorno concreto.

Una **base de datos informatizada** es una estructura de datos a modo de colección de ocurrencias de múltiples tipos de registros lógicos pertenecientes a archivos de datos integrados en una sola estructura global e interrelacionados entre sí a través de los campos de datos comunes.

Las bases de datos informatizadas surgieron en la década de los 80 dependientes en un principio de organismos públicos, y posteriormente privados. Son normalmente de gran tamaño por lo que requieren hardware y software potentes para su implementación y para su explotación.

## Características

- **Independencia de datos lógica y física** en cuanto a que un cambio en la estructura lógica (visión lógica de datos o imagen que tiene el usuario de los datos) o física (visión física de datos o distribución real de los datos en el soporte de almacenamiento) no implique la modificación de las aplicaciones que los utilizan.
- **Concurrencia de servicio** a una o más aplicaciones y uno o varios usuarios, incluso al mismo tiempo, permitiendo establecer distintos niveles de privilegio correspondientes a distintos niveles de autoridad con el fin de velar por la seguridad.
- **Integridad de la información** ante incidencias para que no sea deteriorada por ningún fallo de equipos o procesos mal diseñados o mal utilizados o por problemas derivados de la propia concurrencia (actualización incorrecta, bloqueo mutuo, ...).
- **Mínima redundancia** o repetición de información incontrolada, ocasionada por error durante el diseño de la base de datos, ya que origina varios inconvenientes (coste adicional del soporte de almacenamiento, problemas de múltiples operaciones de actualización, riesgo de posibles estados diferentes de actualización que conducen a posibles incoherencias, ...) aunque permitiéndose la redundancia o repetición de información controlada o mínima por motivos de optimización.





# Bases de Datos.

## Centralizada / Distribuida

- **Centralizada:** concentra todos los componentes y funciones en un único ordenador central al que derivan todos los accesos.
- **Distribuida:** implantada en una red distribuida, con gestión compartida por un entre nodos distintos en diferentes ubicaciones.

Las bases de datos distribuidas admiten distintas posibilidades de configuración:

- Repitiendo toda la información en todos los nodos a modo de base de datos multiplicada.
- Repartiendo la información entre los nodos sin repetirla a modo de base de datos partida.
- Combinando repetición y/o reparto de información en ciertos nodos solamente, teniendo en cuenta estadísticas de acceso.

Las bases de datos distribuidas requieren una gestión global, frente a una gestión local, que por un lado, implica mayor complejidad, aumentando las posibilidades de error, lo que dificulta la coordinación, pudiéndose producir retrasos adicionales en los accesos a la información ocasionados por la comunicación entre nodos, y por otro lado, requiere mayor robustez que proporcione independencia a los nodos en mayor o menor grado (autonomía local) ocultando los detalles de distribución de la información a la que se accede de forma transparente, compartida, e incluso en paralelo, mediante transacciones locales a los datos del propio nodo y, en algún caso, transacciones globales a los datos de otros nodos

## Transaccional / No Transaccional

- **Transaccional** (compatible ACID): admite transacciones o conjunto de operaciones ejecutadas como un bloque unitario.
- **No Transaccional** (sin soporte ACID total, y por tanto no compatible ACID): no admite transacciones.

El acrónimo ACID (atomicidad, consistencia, aislamiento, durabilidad) hace referencia a los siguientes parámetros:

- **Atomicidad:** propiedad que asegura la ejecución completa de operaciones. Cuando una operación consiste en una serie de pasos, se ejecutan todos ellos o no se ejecuta ninguno
- **Consistencia:** propiedad que asegura que sólo se empieza aquello que se puede acabar. Por lo tanto se ejecutan aquellas operaciones que no van a romper las reglas y directrices de integridad de la base de datos. La propiedad de consistencia sostiene que cualquier transacción llevará a la base de datos desde un estado válido a otro también válido.
- **Aislamiento:** propiedad que asegura que una operación no puede afectar a otras.. Esta propiedad define cómo y cuándo los cambios producidos por una operación se hacen visibles para las demás operaciones concurrentes, procurando que la realización de transacciones sobre la misma información sean independientes para no causar ningún tipo de error. El aislamiento puede alcanzarse en distinto grado, siendo un parámetro crucial para elegir un SGBD.
- **Durabilidad:** propiedad que asegura la persistencia de las operaciones realizadas. Cuando se realiza una operación, no se podrá deshacer aunque falle el sistema, garantizando que los datos sobrevivan de alguna manera.

# Bases de Datos.

## Clasificación

### Estructurada / No Estructurada

- **Estructurada:** base de datos sujetas a reglas de diseño formales preconcebidas, con estructuras de datos fijas, que originan esquemas en cierto modo estáticos e inflexibles. A este grupo pertenecen las bases de datos tradicionales o convencionales, como las bases de datos SQL (basadas en el lenguaje SQL).
- **No Estructurada:** base de datos sujetas a reglas de diseño informales no predefinidas, sin estructuras de datos fijas, que generan esquemas en cierto modo libres, dinámicos y flexibles, y hasta cierto punto caóticos, e incluso ausentes. A este grupo pertenecen las últimas generaciones de bases de datos, como las bases de datos NOSQL (Not Only SQL).

Los esquemas de datos estructurados facilitan el procesamiento de la información desde las aplicaciones, aunque dificulta las posibilidades de escalabilidad (aumento de los recursos computacionales o de almacenamiento), pudiéndose producir cuellos de botella cuando se procesan grandes cantidades de datos.

Los esquemas de datos estructurados dificultan el procesamiento de la información desde las aplicaciones, aunque favorece las posibilidades de distribución y escalabilidad (ampliación de recursos computacionales o de almacenamiento), permitiendo un alto rendimiento en el procesamiento de grandes cantidades de datos, salvo en las búsquedas de datos combinados que no son clave, por posible sobrecarga, que se puede solucionar en la capa de aplicación y/o desnormalizando el esquema.



# Bases de Datos.

## ¿QUÉ SON LAS BASES DE DATOS?

- 1** Conjunto organizado de información sobre un tema.  
Ej. La lista de estudiantes de un curso y sus calificaciones.

- 2** Que se almacenan usando un sistema.  
Ej. Crear tablas para guardar los nombres de los estudiantes y sus calificaciones.

- 3** Que permite consultarlos cuando se necesitan.  
Ej. Necesito la lista de los mejores estudiantes.

**Motor de bases de datos** es el software para organizar y consultar la información. **La base de datos** es solo la información.

**Hay dos tipos principales de motores:**

### RELACIONALES (SQL)

Organizan la información en tablas y evitan que los datos se repitan.

- MySQL
- Postgres
- SQL Server
- Oracle



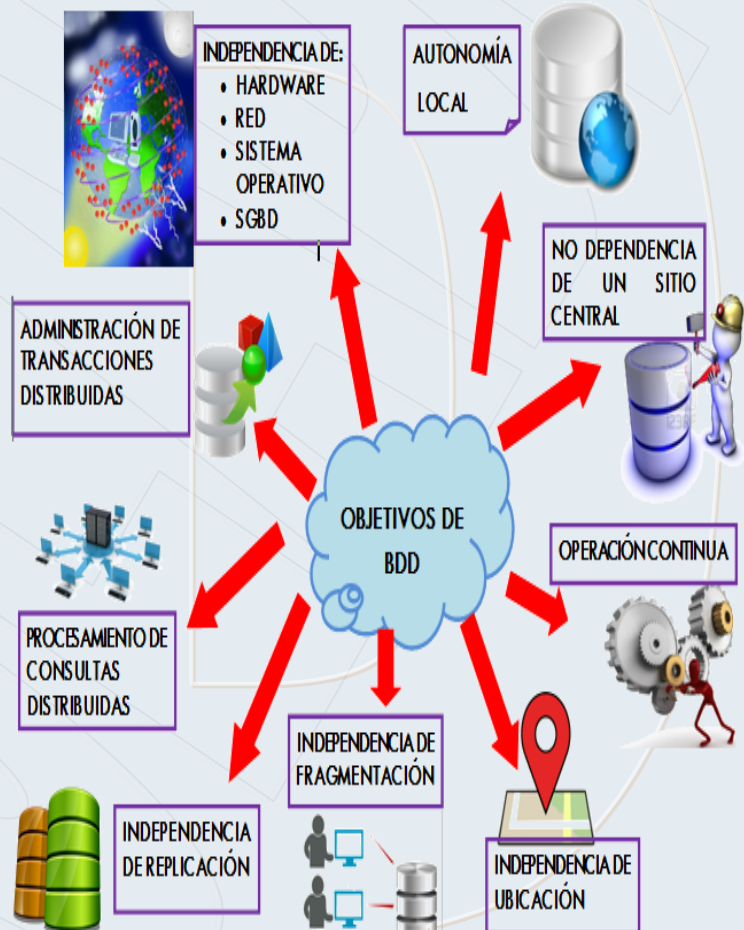
### NO RELACIONALES (NOSQL)

No usan tablas y los datos pueden repetirse en diferentes lugares para acceder mas rápido.

- Cassandra
- Mongo DB
- Firebase
- Redis



El profesional que diseña y administra las bases de datos es el **DBA (Data Base Administrator)**





# Arquitectura de Base de Datos.

La **Arquitectura de Bases de Datos** hace referencia a la organización y estructura de la base de datos (mecanismos de acceso interno, medios de almacenamiento, ...), que se concreta en diferentes esquemas (externo, conceptual, lógico, e interno), según distintos niveles de abstracción (externo, lógico-conceptual, interno), conectados entre sí ocultándose entre ellos los detalles internos.

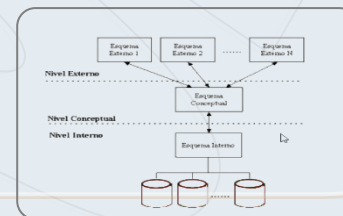
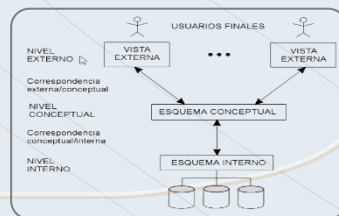
## Niveles

- **Nivel externo:** nivel asociado al esquema externo, referente a la realidad existente, correspondiente a diferentes vistas de la información o visiones de los datos (conjunto de datos y sus características y agrupaciones, cálculos a partir de los datos, ...).
- **Nivel lógico-conceptual:** nivel asociado al esquema de la base de datos (esquema conceptual y esquema lógico), referente a la propia información, que describe la definición de los datos (normas y reglas de validación, relaciones entre datos, rangos y limitaciones, ...).
- **Nivel interno:** nivel asociado al esquema interno, referente a un sistema concreto, que describe los datos tal y como están almacenados (soportes, archivos, modos de organización, métodos de acceso, estructura de registros y campos, ...).

El **nivel externo** se obtiene mediante el análisis previo a partir de la recopilación de información (entrevistas, cuestionarios, observación, ...) en un documento de especificación de requisitos, buscando la máxima fidelidad (rigurosidad en la reproducción de los datos) como objetivo principal.

El **nivel lógico-conceptual** se obtiene, por una parte, mediante el diseño conceptual a partir del esquema externo en base a la independencia respecto de un modelo de base de datos concreto y, por otra parte, mediante el diseño lógico a partir del esquema conceptual en base a la dependencia respecto de un modelo de base de datos concreto y buscando en ambos casos la flexibilidad (facilidad de modificación y ampliación ante cambios del sistema de información) como objetivo principal.

El **nivel interno** se obtiene mediante el diseño interno o diseño físico a partir del modelo lógico en base a la dependencia respecto de un SGBD concreto y buscando el máximo rendimiento (máximo número de transacciones por unidad de tiempo y mínimo tiempo de proceso) como objetivo principal.

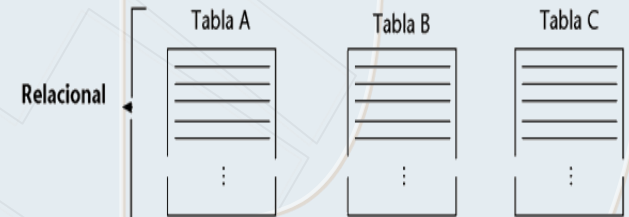
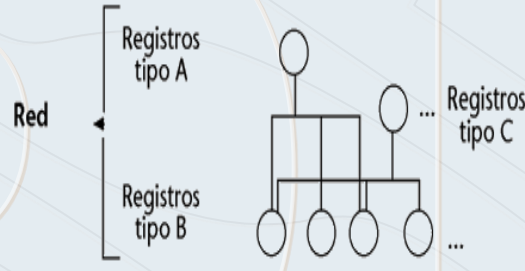
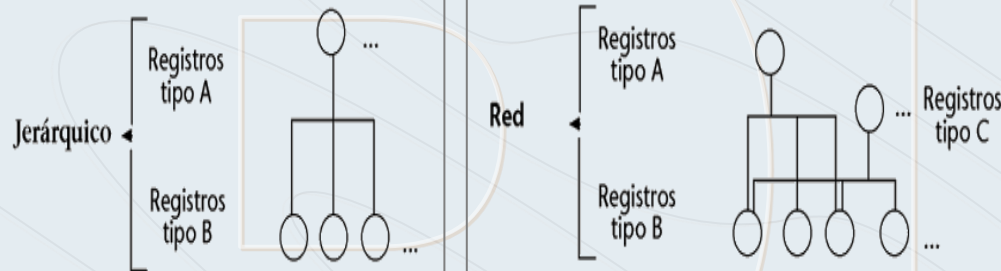


# Modelo de Base de Datos.

El **Modelo de Bases de Datos** hace referencia al conjunto de métodos y procedimientos empleados para definir la arquitectura de la base de datos, desarrollando los diferentes niveles de abstracción para la descripción del esquema de la base de datos, mediante técnicas de modelado de datos.

## Tipos

- Modelos basados en registros.
  - Modelo relacional:** modelo basado en estructuras de tablas, relacionadas entre ellas a través de campos comunes.
  - Modelo jerárquico:** modelo basado en estructuras de árbol, con registros interconectados mediante enlaces.
  - Modelo en red:** modelo basado en estructuras de grafo, con registros interconectados mediante enlaces.
- Modelos orientados a objetos (derivados en gran medida del modelo relacional): basados en tipos abstractos de datos (tipos de datos junto con las operaciones de manipulación de los datos).



# Modelado de Base de Datos.

El **Modelado de Datos** es el proceso de confección de un sistema de información, desde la idea mental inicial correspondiente al acotamiento de una parcela del mundo real (universo del discurso), hasta alcanzar un conjunto de datos final transformado para su implementación real.

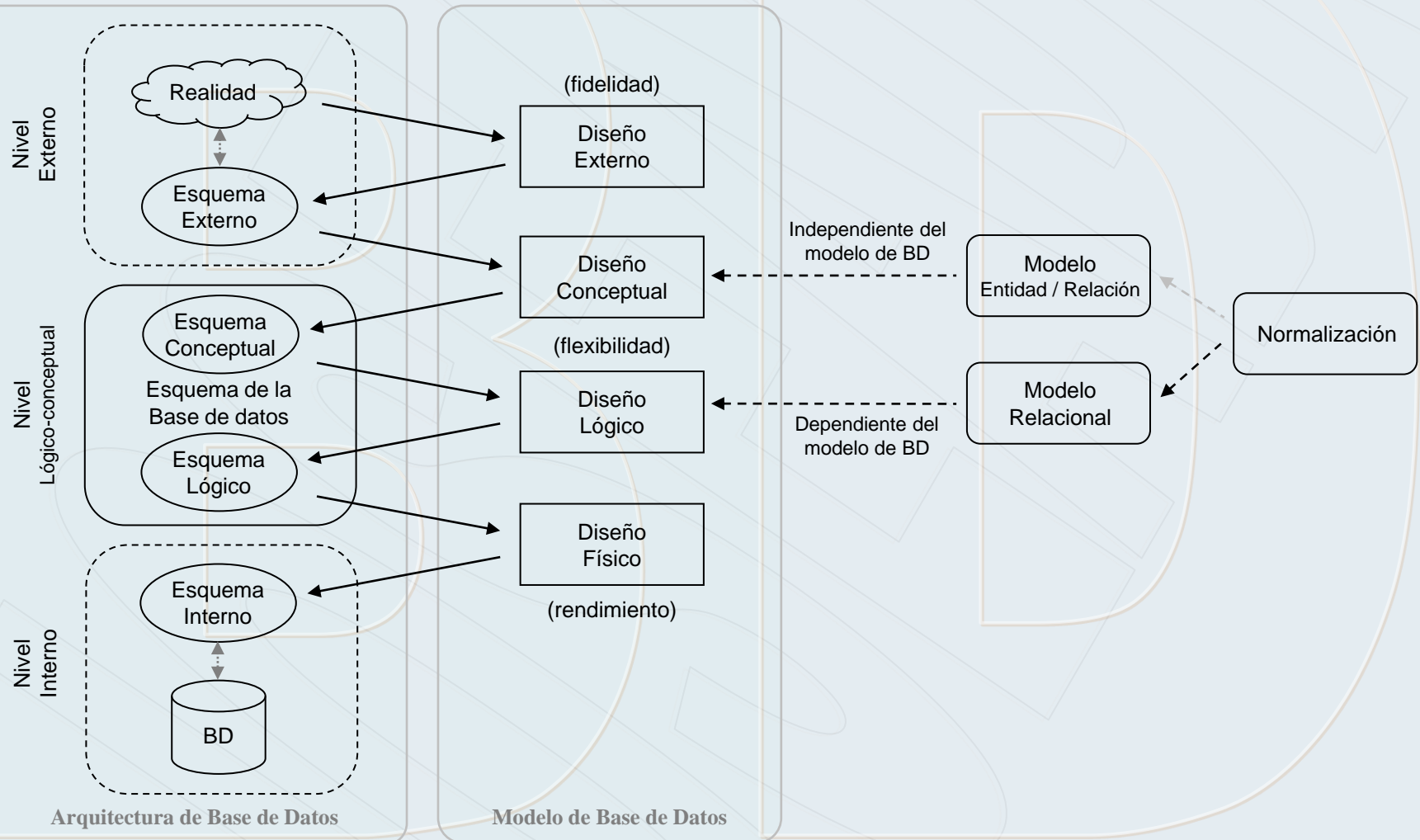
## Fases

El fases de modelado de una base de datos relacional, a partir del enunciado que describe el esquema externo de partida de la base de datos, son:

- 1. Diseño externo**, basado en la realidad existente, recogiendo las especificaciones del sistema de información en un documento de descripción del esquema externo.
- 2. Diseño conceptual**, basado en el modelo E/R (Entidad/Relación), aplicando transformaciones E/R, para obtener um diagrama E/R de representación gráfica de los datos del esquema conceptual.
- 3. Diseño lógico**, basado en el modelo relacional, aplicando transformaciones relacionales, para obtener un conjunto inicial de relaciones como primera aproximación al esquema lógico.
- 4. Diseño lógico normalizado**, basado en un modelo relacional mejorado, aplicando transformaciones de normalización, para obtener un conjunto de relaciones del esquema lógico optimizado.
- 5. Diseño físico**, basado en un modelo de SGBD, aplicando los mecanismos propios de dicho SGBD, para generar el conjunto de tablas que forman el esquema físico.



# Arquitectura y Modelo de Base de Datos.





# Bases de Datos.

