1. Introducción

Las organizaciones de todo tipo, hace mucho tiempo que han reconocido la necesidad de almacenar datos y transformarlos en información. Esta información debe ser administrada, planificada, controlada y tratada como un activo. Este activo debe ser manipulado en forma efectiva y eficiente.

La tarea de las disciplinas de Inteligencia de Negocios (Business Intelligence), Análisis de Datos (data Analytics) y Ciencia de Datos (Data Science) es tomar unos ciertos datos y transformarlos en información para describir, pronosticar y generar conocimiento a partir de ellos. Para finalmente tomar decisiones basados en esos datos.

Sin embargo, para lograr estas metas se deben tener las capacidades de diseñar en forma correcta los datos a capturar para esa generación de conocimiento. ¿Se deben colectar todos los datos?, ¿cómo discriminar aquellos relevantes? ¿cómo muestrear adecuadamente si no dispongo del universo de datos? ¿cuándo se debe efectuar métodos de imputación de datos?

1.1 Datos

Información concreta sobre hechos, elementos, etc., que permite estudiarlos, analizarlos o conocerlos.

"los datos del censo; el análisis aportó datos de gran interés respecto a la génesis de esta fobia; cada ficha contiene los datos comerciales, fiscales y estadísticos de cada proveedor; estos datos configuran una densidad de población débil, aunque ello no descarta que haya núcleos muy poblados y muchas regiones vacías"

Cifra, letra o palabra que se suministra a la computadora como entrada y la máquina almacena en un determinado formato.

"al introducir palabras o números en una hoja de cálculo, la computadora los procesa y los almacena como datos en código binario"

Es una descripción o imagen relacionados con un hecho, evento, personas, objetos u otras entidades del mundo real. El significado del dato cambia dependiendo dentro del contexto en que se encuentre.

- Considere el número **25…**
- Ahora… **25 "Kilos"**
- Y ahora… **25 "kilos" de "papas"**
- Finalmente… **25 "kilos" de "papas" en "mercado" de "Concepción"**

1.2 Información

La información son datos que han sido organizados o preparados en una forma adecuada para apoyar la toma de decisiones.

Por ejemplo, una lista de productos y su stock sin ningún orden son datos, pero una lista de productos ordenados por stock (de menor a mayor) representa información para el encargado de compras de un supermercado.

1.2 Perfiles para trabajar con datos

1.2.1 Director de Datos (CDO)

El CDO (**Chief Data Officer**) es un nuevo rol dentro de aquellas organizaciones con una alta especialización y valoración de los datos. Es un puente entre el área comercial estratégica y el área TI que combina capacidades tecnológicas, estadísticas y gerenciales entre otras:

- Entiende los datos y las necesidades de la empresa respecto a los datos.
- Decide qué datos deben almacenarse en la base de datos.
- Establece políticas para mantener y gestionar los datos almacenados.
- Gestiona los datos como valor estratégico de la organización.
- Establece las bases para el aseguramiento de la calidad de los datos.

El conocimiento a cabalidad del área de negocios de la organización es fundamental para este perfil, ya que es quién guía a través de todo el proceso de generación de información. Define los objetivos para la generación de valor del negocio respecto de la información y hace parte la analítica dentro del objetivo de negocio.

El rol fundamental del CDO se enfoca en sustentar la “visión del negocio” con información.

También dentro de su área de acción se encuentra la gobernanza de los datos y el establecimiento de las políticas de uso de la información. Pasan de un rol de

administrador a uno más estratégico e innovador que permite responder a los cambios tecnológicos cambiando su ambiente de datos en las nuevas áreas de Big Data, automatización y aprendizaje de máquinas.

El CDO dentro de la gobernanza de datos asesora en la implementación de políticas y coordina tanto los requisitos como el control de la información sobre los restantes actores.

1.2.2 Ingeniero de Datos (Data Engineer)

El trabajo del ingeniero de datos es "la representación y el movimiento de datos para que sean consumibles y utilizables". Si eres un ingeniero de datos, debes tomar los datos sin procesar, limpiarlos, moverlos a una base de datos, etiquetarlos y, en general, asegurarte de que estén listos para la siguiente etapa del proceso

Herramientas como Apache Spark, Scala, Docker, Java, Hadoop Podríamos decir que el perfil de ingeniero de datos es el más técnico en el ámbito del Big Data.

Los ingenieros de datos se encuentran entre los desarrolladores de aplicaciones y los científicos de datos (Data Scientists).

- Se encargan de diseñar, construir y gestionar los datos y la infraestructura necesaria para almacenarlos y procesarlos.
- Construyen la base tecnológica para que los científicos de datos y analistas puedan realizar sus tareas. Por lo tanto, son los responsables de mantener sistemas escalables, con alta disponibilidad y rendimiento, integrando nuevas tecnologías y desarrollando el software necesario.
- Deben conocer el stack de tecnologías Big Data, entender cómo se integran sus tecnologías y las formas de procesar, transformar y tratar los datos con herramientas de ingesta y ETL.
- Además, deben saber cómo mover datos hacia y desde el ecosistema Hadoop, implementar y configurar herramientas y bases de datos como Hive o HBase.
- Entre sus funciones también se encuentra dar apoyo y facilitar el trabajo a analistas y científicos de datos, así como a negocio. Esta es la razón de que las habilidades de comunicación tengan una gran importancia.
- Entre los conocimientos básicos debe estar Linux. La mayoría de las cargas y despliegues Cloud y Big Data se realizan sobre este sistema operativo. Al menos debes sentirte cómodo usando la terminal para editar ficheros, ejecutar comandos y navegar por el sistema.
- Automatización y scripting con algún lenguaje de programación como Python. Este punto incluye la capacidad de interaccionar con APIs y otras fuentes de datos de manera simple y directa.

1.2.3 Analista de Datos (Data Analyst)

Un analista de datos es un profesional especializado en interpretar conjuntos masivos de datos con el objetivo de extraer información valiosa y significativa para respaldar la toma de decisiones empresariales. Este rol implica habilidades en el manejo de

herramientas y software especializado para recopilar, limpiar y analizar datos provenientes de diversas fuentes, como bases de datos, redes sociales o sistemas empresariales.

El analista de datos utiliza técnicas estadísticas y de visualización para identificar patrones, tendencias y relaciones dentro de los datos, transformando la información cruda en conocimiento accionable. Su función principal radica en convertir datos complejos en perspectivas comprensibles que ayuden a las organizaciones a optimizar procesos, identificar oportunidades de crecimiento, mitigar riesgos y mejorar la toma de decisiones estratégicas.

1.2.4 Científico de Datos (Data Scientist)

Un científico de datos es un profesional que emplea técnicas analíticas avanzadas para extraer información significativa de conjuntos masivos de datos. Esta disciplina combina habilidades en programación, estadística, matemáticas y conocimientos en el campo de estudio para descubrir patrones, tendencias y relaciones dentro de los datos. Utilizan herramientas y lenguajes de programación como Python, R o SQL para limpiar, procesar y analizar datos con el fin de obtener perspectivas que puedan respaldar la toma de decisiones estratégicas. Además, los científicos de datos desarrollan modelos predictivos y descriptivos, diseñan algoritmos y aplican técnicas de aprendizaje automático para generar pronósticos, identificar oportunidades o resolver problemas complejos en diversas industrias, desde la medicina y las finanzas hasta el marketing y la tecnología.

Su función principal radica en transformar datos crudos en información accionable. Esto implica comprender las necesidades del negocio o campo en el que trabajan, identificar fuentes de datos relevantes, limpiar y procesar esos datos, aplicar técnicas estadísticas y algorítmicas para revelar patrones y tendencias, y comunicar de manera efectiva los hallazgos a las partes interesadas. Los científicos de datos juegan un papel fundamental en la toma de decisiones estratégicas al proporcionar información basada en evidencia, lo que permite a las organizaciones optimizar procesos, anticipar tendencias y mejorar su desempeño general.

1.2.5 Administrador de Base de datos (DBA)

El DBA (Data Base Admininistrator) es el profesional informático encargado de la administración de una o varias bases de datos gestionando su uso y funcionamiento. Es responsable por el diseño de la base de datos y la gestión de ella, fijando normas que resguardan tanto la seguridad como la integridad de ellas.

Funciones

- Crea la base de datos.
- Implementa los controles necesarios para que se respeten las políticas establecidas por el administrador de datos.
- Es el responsable de garantizar que el sistema obtenga las prestaciones deseadas. Presta servicios técnicos.
- Mantener la base de datos disponible y actualizada.
- Realizar los respaldos de seguridad. Define políticas de seguridad y de respaldo.

- Disponer del acceso a los datos desde las aplicaciones.
- Mantener la seguridad de los datos.
- Diseñar y administrar la estructura de los datos.
- Monitorear la actividad de los datos.
- Se asegura de que la comunicación del sistema con la base de datos sea expedita.

Los Administradores de Bases de Datos son responsables del manejo, mantenimiento, desempeño y de la confiabilidad de bases de datos. Asimismo, están a cargo de la mejora y diseño de nuevos modelos de estas. Manejar una base de datos implica recolectar, clasificar y resguardar la información de manera organizada, por ello, estos profesionales velan por garantizar que la misma esté debidamente almacenada y segura, además de que sea de fácil acceso cuando sea necesario.

1.2.6 Desarrollador de Base de Datos

Personas como analistas de sistemas y programadores que diseñan nuevos programas de aplicación para los usuarios finales.

Los programadores de sistemas informáticos escriben programas para controlar el funcionamiento interno de los ordenadores, lo que implica diseñar programas que sean eficientes, rápidos y versátiles. Dedican mucho tiempo a probar los programas, y también puede instalar, personalizar y dar soporte a estos sistemas operativos.

1.2.7 Usuario

Personas que utilizan datos de la base de datos para su trabajo cotidiano no necesariamente del área de la informática. Normalmente no utilizan la base de datos directamente, sino aplicaciones creadas para ellos a fin de facilitar la manipulación de los datos. Estos usuarios sólo acceden a ciertos datos.

2. Bases de Datos

El almacenamiento de datos es y ha sido relevante desde siempre, antes de la era informática los datos eran capturados y almacenados en documentos y planillas de todo tipo. Luego, se almacenan por excelencia en planillas digitales como Excel o Lotus donde el problema siempre ha sido el mismo, cómo asegurar el ingreso de información correcta y cómo estructurar la información.

Siempre se ha intentado recopilar los datos en forma ordenada y sistemática de forma que este almacenamiento contribuya a la extracción de información relevante.

2.1 Archivos

En el enfoque de archivos tradicionales (documentos) los datos se almacenan en archivos individuales, exclusivos para cada aplicación particular.

Hoy en día los documentos tradicionales para el almacenamiento de datos es la planilla de Microsoft Excel y su competidor más cercano Google Sheet. Muchos departamentos dentro de organizaciones usan estos archivos para la carga de datos, la razón sigue siendo su simplicidad y rapidez para ingresar datos, para usuarios no necesariamente especialistas en informática.

La información en estas planillas se ingresa en forma tabular con conceptos de fila y columnas que se han agregado también a su uso en las bases de datos.

Es muy fácil ingresar datos en estas planillas de datos y luego cargar esta información para análisis en programas de análisis o en bases de datos especializadas. Generalmente debe exportarse estos datos en un formato más simple denominado CSV (“comma separated value”).

En estas planillas los datos pueden ser una alta fuente de error y la actualización de los archivos es más lenta que en una base de datos.

Algunos de los problemas de utilización los ficheros se resumen en:

- Redundancia

Al no existir algún tipo de control sobre el ingreso más que el del usuario, es muy normal que existan este tipo de errores de duplicidad en los registros.

- Error de ingreso

Errores comunes en el ingreso manual de datos, errores de tipo ortográfico, números mal ingresados, etc.

- Estandarización

Es el tipo de error más común y se ejemplifica en el ingreso de fechas donde a pesar de poder regir el formato de entrada, no impide que se ingrese otros formatos que si bien pueden ser correctos, interfieren en la forma de incluirse en una base de datos. 21-12-2021 o bien 21/02/2021, o 21/2/2021).

- Seguridad

No hay un control de uso y acceso por usuarios a los datos, más que el control al archivo físico en el computador local o servidor.

2.2 Bases de Datos

Gran parte de los errores mencionados anteriormente pueden ser evitados utilizando una base de datos para la captura y almacenamiento de datos. Para lograr un efectivo tratamiento del recurso dato las organizaciones han optado por trabajar con Bases de Datos que permiten tener un conjunto de datos relacionados y almacenados de forma permanente y usados con diferentes propósitos por múltiples usuarios y que permiten:

- **Integrar**: significa que los diferentes archivos de datos han sido lógicamente organizados para reducir la redundancia de datos y facilitar el acceso a ellos.
- **Compartir**: significa que todos los usuarios calificados tienen acceso a los mismos datos, para usarlos en diferentes actividades.

Se puede definir una base de datos como un conjunto organizado de información almacenada de forma estructurada en un sistema computarizado que permite el almacenamiento, gestión y recuperación de datos de manera eficiente para su uso y análisis posterior.

2.3 Sistemas Gestores de Bases de Datos

Las principales funciones del Sistema Gestor de Bases de datos SGBD (en inglés Database Management System, "DBMS" o también Relational Database management System, "RDBMS") son:

- **Definición de Datos**: (se puede realizar a través del lenguaje de definición de datos o DDL) que provee el DBMS.
- **Manipulación de Datos**: permite almacenar, modificar y recuperar los datos de la Base de Datos. Esto se logra a través del lenguaje de manipulación de datos o DML provisto por el DBMS
- **Seguridad de Datos**: el DBMS provee de mecanismos para controlar el acceso y para definir qué operaciones puede realizar cada usuario. Además, debe proveer de mecanismos de respaldo y recuperación de la Base de Datos, También debe manejar el acceso concurrente a la Base de Datos.

2.3.1 Funciones Generales

- Permitir a los usuarios almacenar datos, acceder a ellos y actualizarlos, ocultando su estructura física.
- Proporcionar un catálogo (diccionario de datos) accesible por los usuarios.
- Proporcionar un mecanismo que garantice el procesamiento de las transacciones.
- Proporcionar un mecanismo que realice el control de la concurrencia.
- Proporcionar un mecanismo para recuperación ante fallos.
- Proporcionar un mecanismo de seguridad.
- Integrarse con algún software de comunicación.
- Encargarse de mantener las reglas de integridad.
- Encargarse de mantener la independencia entre los programas y la estructura de la base de datos.
- Proporcionar herramientas para administrar la base de datos.

2.4 Ventajas del uso de bases de datos

- **Mínima Redundancia de Datos**: Al integrar los datos en una sola estructura lógica y almacenando cada ocurrencia de un ítem de dato en un solo lugar de la Base de Datos, se reduce la redundancia.
- **Consistencia de Datos**: Al controlar la redundancia de datos, se reduce enormemente la inconsistencia, dado que, al almacenarse un dato en un solo lugar, las actualizaciones no producen inconsistencia.
- **Integración de Datos**: En una Base de Datos, los datos son organizados de una manera lógica que permite definir los relacionamientos entre ellos.
- **Compartir Datos**: Una Base de Datos es creada para ser compartida por todos los usuarios que requieran de sus datos; muchos sistemas de Base de Datos permiten a múltiples usuarios compartir la Base de Datos en forma concurrente, aunque bajo ciertas restricciones.
- **Esfuerzo por Estandarización**: Establecer la función de Administración de Datos es una parte importante de este enfoque, su objetivo es tener la autoridad

para definir y fijar los estándares de los datos, así como también posteriores cambios de estándares.

- **Facilitar el Desarrollo de Aplicaciones**: Este enfoque reduce el costo y tiempo para desarrollar nuevas aplicaciones.
- **Controles de Seguridad, Privacidad e Integridad**: El control centralizado que se ejerce bajo este enfoque, a través de la función Administración de Base de Datos, puede mejorar la protección de datos en comparación con archivos tradicionales Flexibilidad en el acceso: Este enfoque provee múltiples formas de recuperación de cada ítem de dato, permitiendo a un usuario mayor flexibilidad para ubicar datos que en archivos tradicionales.
- **Independencia de los Datos**: Permite cambiar la organización de los datos sin necesidad de alterar los programas. Es uno de los objetivos principales del enfoque de Base de Datos.
- **Reducción de la Mantención de Programas**: Como los datos son independientes de los programas se reduce la necesidad de modificar (mantener) los programas aun cuando existan una modificación constantes de éstos.

2.5 Desventajas

- **Personal Especializado**: Generalmente se necesita contratar o capacitar a personas para convertir sistemas existentes, desarrollar y estimar nuevos estándares de programación, diseñar Bases de Datos y administrarlas.
- **Necesidad de Respaldos**: Al tener mínima redundancia se requiere contar con respaldos independientes que ayuden a recuperar archivos dañados, los DBMS generalmente proveen de herramientas que permiten respaldar y recuperar archivos.
- **Problemas al compartir Datos**: El acceso concurrente a los datos puede causar datos no consistentes o bloqueo de datos (deadlock). Los DBMS deben ser diseñados para prevenir o detectar tales interferencias, de una forma que sea transparente para el usuario.
- **Conflicto Organizacional**: El mantener los datos en una Base de Datos para ser compartidos, requiere de un consenso en la definición y propiedad de los datos como también en la responsabilidad por la exactitud de ellos.

2.6 ACID

En bases de datos se utiliza el termino **ACID** para definir a aquellas que cumplen ciertas características en sus transacciones. Una **transacción** es una serie de procesos que se aplican dentro de una base de datos en forma secuencial u ordinal y que debe realizarse de una vez y sin alterar la estructura de los datos.

Una base de datos es complementaria ACID si presenta estas cuatro propiedades dentro de las transacciones de una BD:

- **Atomicidad**
  Referida a la propiedad que determina que la operación se haya realizado o no, pero nunca a medias. Se ejecuta la operación completa con todos sus pasos o no se ejecuta del todo.

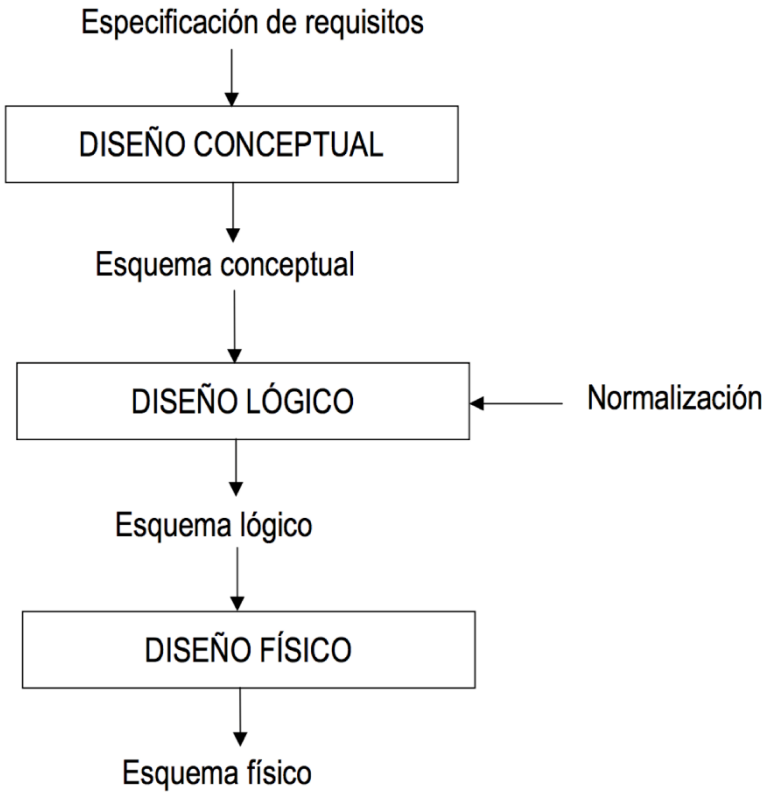
- **Consistencia**
  Solo se ejecutan las operaciones que no afectan la integridad de la base de datos. Cualquier operación que se lleva a cabo será de un estado válido a otro con datos consistentes.
- **Aislamiento**
  (Isolation) Cada operación es única y no afecta a otras aunque se realicen sobre la misma información.
- **Durabilidad**
  Asegura que la operación una vez realizada, es persistente y no se podrá deshacer a pesar de fallos en el sistema.

3. Modelamiento de Datos

Al igual que los arquitectos realizan sus planos para construir casas, los diseñadores de base de datos necesitan realizar modelos para construir sus bases de datos.

Los modelos facilitan la comunicación entre el diseñador de base de datos y los usuarios finales. Los modelos son fáciles de utilizar y cambiar, ya que son sólo una imagen muy simplificada del sistema de información que se desea desarrollar.

Actualmente en la participación de la construcción del modelo de datos se involucran varias etapas.



- **Conceptual**: esta fase incluye la identificación de las entidades clave del sistema y empresariales de nivel superior y sus relaciones, que definen el ámbito del problema que tratará el sistema. Estas entidades clave del sistema y empresariales se definen mediante la utilización de elementos de modelado del perfil UML para el modelado empresarial, incluidos los elementos del modelo de análisis empresarial y el modelo de clase de análisis del modelo de análisis.

- **Lógica**: esta fase incluye el perfeccionamiento de las entidades del sistema y empresariales de alto nivel de la fase conceptual en entidades lógicas más detalladas. Estas entidades lógicas y sus relaciones se pueden definir, opcionalmente, en un modelo lógico de datos mediante la utilización de los elementos de modelado del perfil UML para el diseño de bases de datos, como se describe en la Directriz: Modelo de datos. Este modelo lógico de datos forma parte del Producto de trabajo: Modelo de datos.

- **Física**: esta fase incluye la transformación de los diseños de la clase lógica en diseños de tablas de bases de datos físicas detalladas y optimizadas. La fase física también incluye la correlación de los diseños de tablas de base de datos con espacios de tablas y con el componente de base de datos en el diseño de almacenamiento de bases de datos.

3.1 Modelo Entidad-relación (MER)

El Modelo Entidad Relación (MER) es una herramienta de modelado que fue introducido originalmente por Peter Chen[8] en 1976 y aunque ha sufrido variaciones en cuanto a los elementos de diagramas utilizados para representar sus elementos, su operación y utilidad siguen vigentes. La base del MER está en identificar los elementos o entidades importantes del sistema, los datos (atributos) que componen cada uno de ellos y la interacción relación) entre dichos elementos.

Es una metodología de diseño de Bases de Datos que consiste en representar a nivel conceptual los datos que soportan el funcionamiento de un sistema. Los componentes básicos de un MER son: Entidades, Atributos y Relaciones. Las entidades representan abstracciones con atributos que almacenan datos; las relaciones son las asociaciones que existen entre entidades y permiten generar información al combinar diferentes entidades.

ENTIDAD: Se denomina entidad a todo ente (conceptual o físico) del cual se desea establecer su participación dentro de un sistema de información. Una entidad concreta o física es aquella con existencia física, representa un objeto del mundo real (persona o elemento). Unaentidad abstracta no tiene una representación física concreta (posición laboral, asignatura).

ATRIBUTO: El atributo es elementos de información que caracteriza a una entidad, identificándola, calificándola, cuantificándola, o declarando su estado. Por lo general una entidad se compone de uno o más atributos (edad, genero, estatura, nombre, etc.). Los atributos permiten diferenciar elementos dentro de un conjunto de entidades.

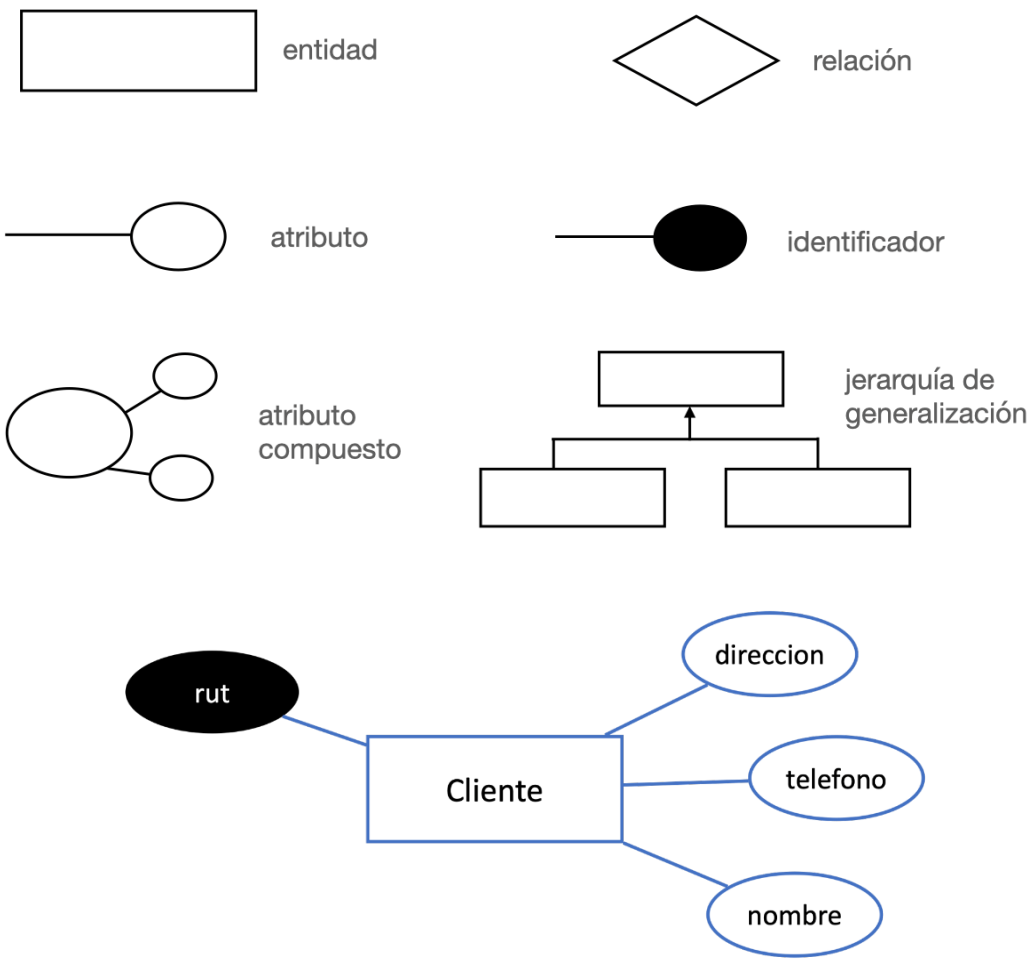
Dentro de una entidad de tipo persona es muy raro el caso que existan dos con exactamente los mismos atributos.

RELACIONES: Las relaciones identifican la interacción que existe entre dos o más entidades. Establecen el comportamiento del sistema de información.

3.1.1 Elementos básicos

Los elementos básicos de MER se presentan en un diagrama simple que permite establecer en forma general un modelo de datos.

Los elementos fundamentales son:



3.2 Modelamiento Conceptual

Un modelo conceptual de datos identifica las relaciones de más alto nivel entre las diferentes entidades.

Características

- Incluye las entidades importantes y las relaciones entre ellas.
- No se especifica ningún atributo.
- No se especifica ninguna clave principal.

El requisito para un modelamiento exitoso pasa necesariamente por el "conocimiento del negocio", esto es, para lograr la meta de representar y organizar los datos para obtener la información que requiere el problema a resolver, se necesita un conocimiento cabal del problema.

No es igual modelar un sistema de inventario para un negocio de local único que para una cadena de tiendas, o para una clínica de salud. Siempe el modelo final va a estar supeditado a los requerimientos específicos del negocio.

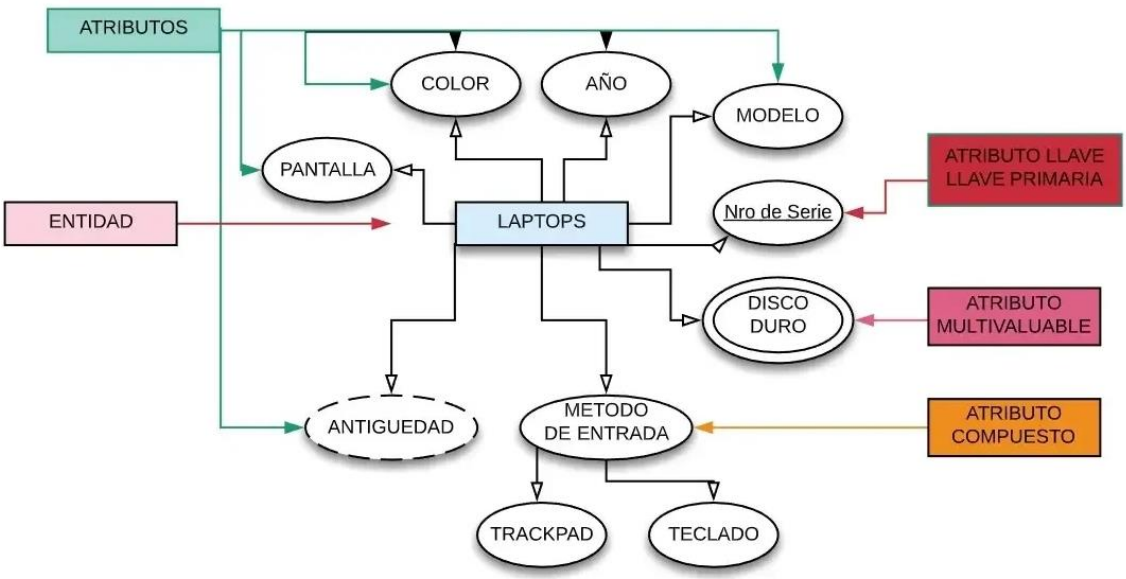
Modelar significa en un modo amplio simplificar la realidad del negocio, pero sin perder significancia de sus datos. Modelar implica **organizar** y **clasificar** la información en componentes simples que representen la información del negocio.

4.2.1 Entidades

Una entidad es algo similar a un objeto (programación orientada a objetos) y representa algo en el mundo real, incluso algo abstracto. Tienen atributos que son las cosas que los hacen ser una entidad y por convención se ponen en plural.

Ejemplo de entidad en bases de datos

En la imagen puedes observar como ejemplo que la entidad Laptops posee diferentes atributos como color, pantalla, año, modelo, etc.



Los atributos son las características o propiedades que describen a la entidad (se encierra en un óvalo). Los atributos se componen de:

Los atributos compuestos son aquellos que tienen atributos ellos mismos.

Los atributos llave son aquellos que identifican a la entidad y no pueden ser repetidos. Existen:

- Naturales: son inherentes al objeto como el número de serie
- Clave artificial: no es inherente al objeto y se asigna de manera arbitraria.

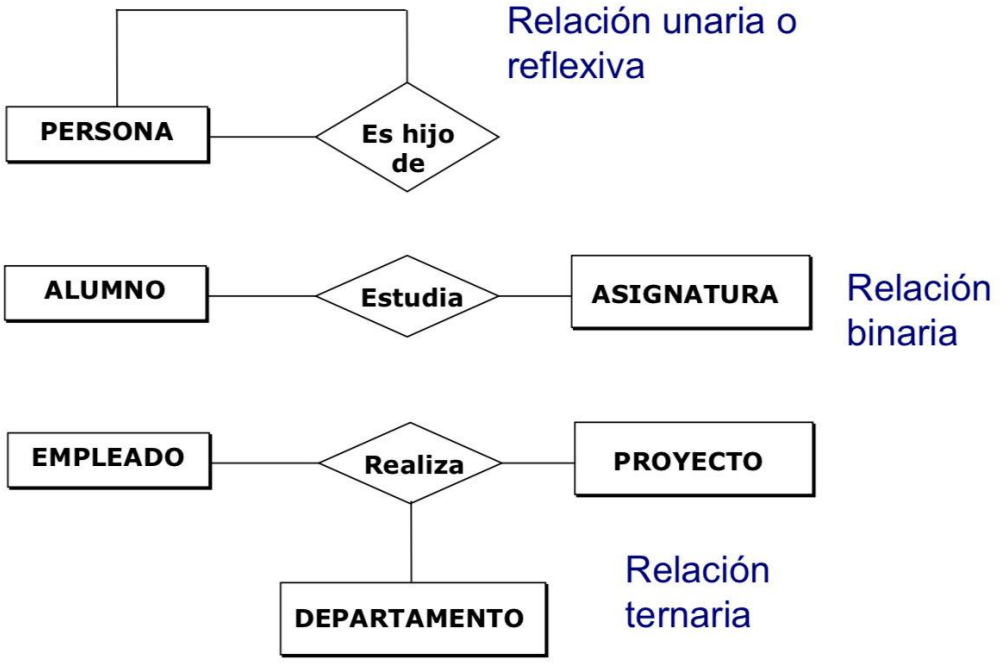
4.2.2 Relaciones

La relación de una base de datos es el vínculo que se establece entre distintos elementos de las tablas que la conforman. En este tipo de relaciones es fundamental el uso de los campos de llave primaria (primary key) que son los que se relacionan con otros registros de otras tablas.

A la hora de definir las relaciones entre los campos de distintas tablas en una base de datos, los nombres de los mismos no tienen por qué ser iguales. Sin embargo, sí es necesario a la hora de establecer estas relaciones, que el tipo de datos de los campos enlazados sea el mismo.

Las relaciones en las bases de datos son claves para establecer concordancias en las asignaciones y garantizar la integridad referencial de la información (que los datos no se modifiquen o varíen durante el proceso).

Gracias a las relaciones se mantiene una lógica y consistencia entre todos los datos que almacena. Además, las relaciones evitan que se dupliquen los registros dentro de una base de datos.



3.3 Modelamiento Lógico

Un modelo de datos lógicos describe los datos con el mayor detalle posible, independientemente de cómo se implementarán físicamente en la base de datos.

Características

- Incluye todas las entidades y relaciones entre ellos.
- Todos los atributos para cada entidad están especificados.
- La clave principal para cada entidad está especificada.
- Se especifican las claves externas (claves que identifican la relación entre diferentes entidades).
- La normalización ocurre en este nivel.

Etapas

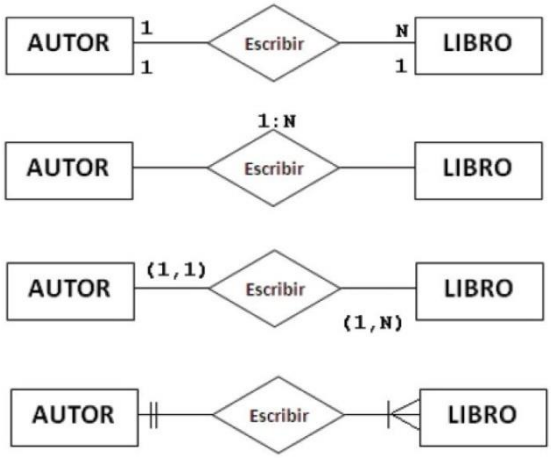
- Especificar claves primarias para todas las entidades.
- Encontrar las relaciones entre diferentes entidades.
- Encontrar todos los atributos para cada entidad.
- Resolver las relaciones de muchos a muchos.
- Normalización.

4.3.1 Cardinalidad

La **cardinalidad** está definida como la cantidad de elementos en términos de proporción que participan en la **relación** entre dos o más **entidades**. Esta puede ser entre elementos únicos (unitarios) o múltiples.

Generalmente se utiliza la denominación "1" para elemntos unitarios, y "N" para varios elementos participantes.

Los diagramas siguientes expresan que "un autor escribe varios libros" o "un libro es escrito por un autor".



Las expresiones que pueden indicarse a partir de estos diagramas son:

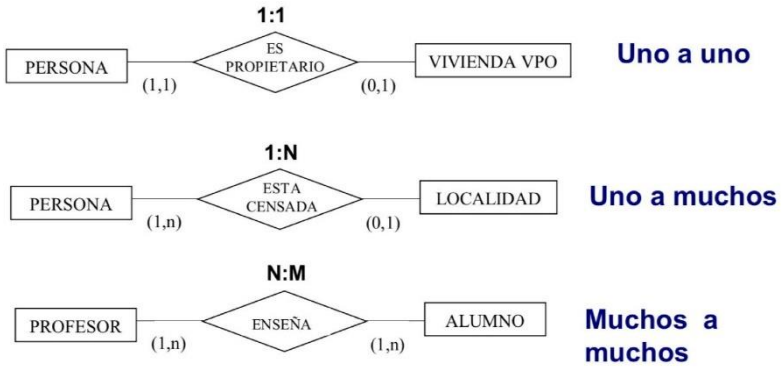
- Un AUTOR escribe AL MENOS un LIBRO" (Cardinalidad mínima=1),

- Un AUTOR escribe VARIOS LIBROS” (Cardinalidad máxima = N),
- Un AUTOR escribe uno o varios LIBROS” (Cardinalidad mínima= 1, máxima =N)
- Un LIBRO es escrito por un y sólo un AUTOR” (Cardinalidad mínima y máxima = 1)

La FORMA de representar la CARDINALIDAD en un diagrama puede variar.

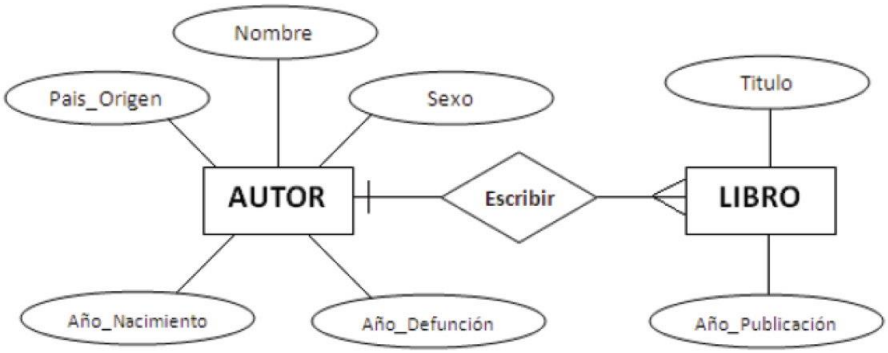
Algunos prefieren dejar escrito en el diagrama el mínimo y el máximo con números entre paréntesis, otros recomiendan escribir la Cardinalidad sobre la relación, mientras que otros recomiendan usar la notación de PATAS DE GALLO (CROW’S FEET), la cual es muy aceptada y usada por desarrolladores.

Por tanto, la cardinalidad representa el número máximo de ocurrencias de una entidad asociadas al número máximo de ocurrencias del resto de las entidades relacionadas.



Lo importante es ADOPTAR una nomenclatura y ser consecuente con ella. INDEPENDIENTE de la nomenclatura que se escoja, es imprescindible que en el diagrama se refleje la Cardinalidad mínima y máxima de las relaciones.

NOTA: Aunque son muy parecidos en la segunda versión del diagrama, la línea que une a las entidades cumple la función de denotar la cardinalidad mínima del modelo. En este caso al ser línea continua se entiende que la cardinalidad mínima es 1, si fuera una línea segmentada se entendería que la cardinalidad mínima es 0, es decir, una OPCIONALIDAD.



3.4 Modelamiento Físico

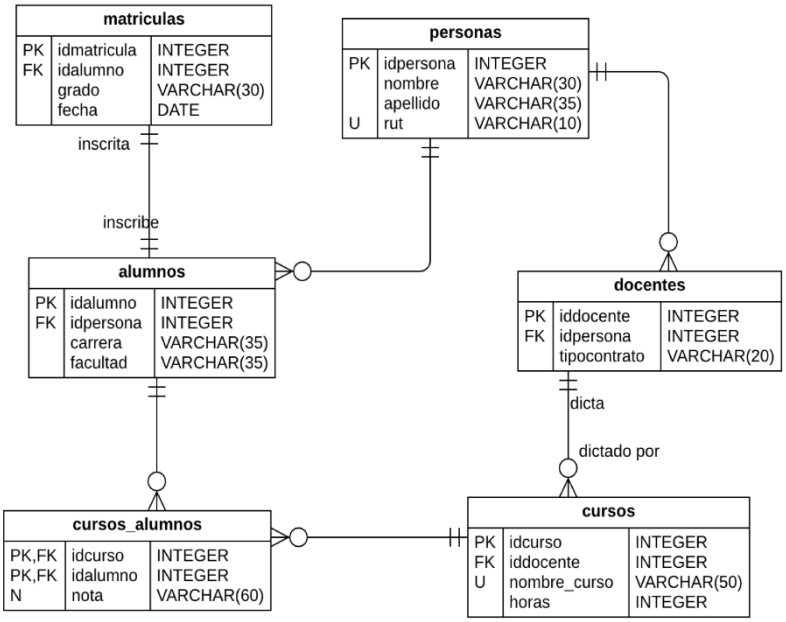
El modelo de datos físicos representa cómo se construirá el modelo en la base de datos. Un modelo de base de datos física muestra todas las estructuras de tabla, incluidos el nombre de columna, el tipo de datos de columna, las restricciones de columna, la clave principal, la clave externa y las relaciones entre las tablas.

Características

- Especificación de todas las tablas y columnas.
- Las claves externas se usan para identificar relaciones entre tablas.
- La desnormalización puede ocurrir según los requisitos del usuario.

Etapas

- Convertir entidades en tablas.
- Convertir relaciones en claves externas.
- Convertir atributos en columnas.
- Modificar el modelo de datos físicos en función de las restricciones / requisitos físicos.



4 El Modelo Relacional

El modelo relacional, para el modelado y la gestión de bases de datos, es un modelo de datos basado en la lógica de predicados y en la teoría de conjuntos.

Tras ser postuladas sus bases en 1970 por Edgar Frank Codd, de los laboratorios IBM en San José (California), no tardó en consolidarse como un nuevo paradigma en los modelos de base de datos.

Su idea fundamental es el uso de relaciones. Estas relaciones podrían considerarse en forma lógica como conjuntos de datos llamados tuplas. Pese a que esta es la teoría de las bases de datos relacionales creadas por Codd, la mayoría de las veces se conceptualiza de una manera más fácil de imaginar, pensando en cada relación como si fuese una tabla que está compuesta por registros (cada fila de la tabla sería un registro o "tupla") y columnas (también llamadas "campos").

Las 12 reglas de Codd son un sistema de 13 reglas —numeradas del 0 al 12— propuestas por el creador del modelo relacional de bases de datos, Edgar F. Codd, para definir los requerimientos que un sistema de administración de base de datos ha de cumplir para poder ser considerado relacional como lo son, por ejemplo, las bases de datos relacionales. Puedes ver más sobre las reglas en el siguiente enlace

Codd se percató de que existían bases de datos en el mercado que decían ser relacionales, pero lo único que hacían era guardar la información en tablas, sin estar estas tablas literalmente normalizadas; entonces publicó las 12 reglas que un verdadero sistema relacional debería cumplir, aunque en la práctica algunas de ellas son difíciles de realizar. Un sistema podrá considerarse «más relacional» cuanto más siga estas reglas.2

Es el modelo más utilizado en la actualidad para modelar problemas reales y administrar datos dinámicamente.

El modelo relacional desarrolla un esquema de base de datos (data base schema) a partir del cual se podrá realizar el modelo físico o de implementación en el DBMS.

Este modelo está basado en que todos los datos están almacenados en tablas (entidades/relaciones) y cada una de estas es un conjunto de datos, por tanto una base de datos es un conjunto de relaciones. La agrupación se origina en la tabla: tabla -> fila (tupla) -> campo (atributo)

El Modelo Relacional se ocupa de:

- La estructura de datos
- La manipulación de datos
- La integridad de los datos

Donde las relaciones están formadas por:

- Atributos (columnas)
- Tuplas (Conjunto de filas)

4.1 Objetivos

Los objetivos que este modelo persigue son:

- **Independencia Física**: La forma de almacenar los datos no debe influir en su manipulación. Si el almacenamiento físico cambia, los usuarios que acceden a esos datos no tienen que modificar sus aplicaciones.
- **Independencia Lógica**: Las aplicaciones que utilizan la base de datos no deben ser modificadas por que se inserten, actualicen y eliminen datos.
- **Flexibilidad**: En el sentido de poder presentar a cada usuario los datos de la forma en que éste prefiera
- **Uniformidad**: Las estructuras lógicas de los datos siempre tienen una única forma conceptual (las tablas), lo que facilita la creación y manipulación de la base de datos por parte de los usuarios.
- **Sencilles**: Las características anteriores hacen que este Modelo sea fácil de comprender y de utilizar por parte del usuario final.

4.2 Características

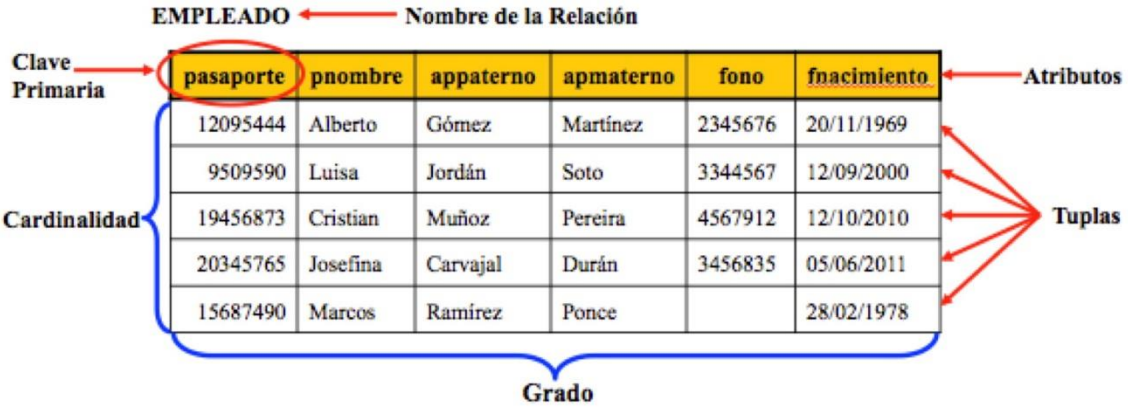
- Los datos son atómicos. Un dato se considera atómico cuando a cada información (cada asunto) se le reserva una celda propia.
- Los datos de cualquier columna son de un solo tipo.
- Cada columna posee un nombre único.
- El orden de las columnas no es de importancia para la tabla.
- Las columnas de una relación se conocen como atributos.
- Cada atributo tiene un dominio,
- No existen 2 filas en la tabla que sean idénticas.
- La información en las bases de datos es representada como datos explícitos.
- Cada relación tiene un nombre específico y diferente al resto de las relaciones.
- Los valores de los atributos son atómicos: en cada tupla, cada atributo (columna) toma un solo valor. Se dice que las relaciones están normalizadas.
- El orden de los atributos no importa: los atributos no están ordenados.
- Cada tupla es distinta de las demás: no hay tuplas duplicadas
- El orden de las tuplas no importa: las tuplas no están ordenadas.
- Los atributos son atómicos: en cada tupla, cada atributo (columna) toma un solo valor. Se dice que las relaciones están normalizadas.

4.3 Definiciones

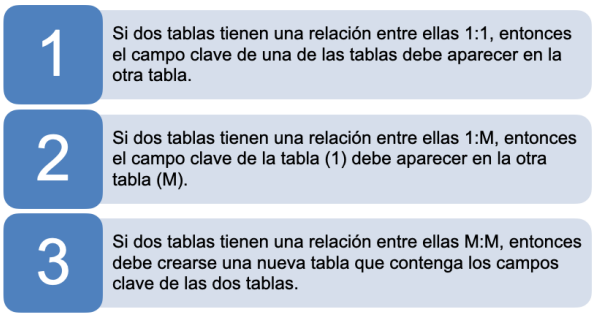
- **Relación**: Tabla bidimensional para la representación de datos. Ejemplo: Estudiantes.

- **Tuplas**: Filas de una relación que contiene valores para cada uno de los atributos (equivale a los registros). Ejemplo: 34563, José, Martinez, 19, Masculino. Representa un objeto único de datos implícitamente estructurados en una tabla. Un registro es un conjunto de campos que contienen los datos que pertenecen a una misma entidad.
- **Atributos**: Columnas de una relación y describe las características particulares de cada campo. Ejemplo: id estudiante
- **Esquemas**: Forma de representar una relación y su conjunto de atributos. Ejemplo: Estudiantes (id estudiante, nombre(s), apellido(s), edad, género)
- **Claves**: Campo cuyo valor es único para cada registro. Principal, identifica una tabla, y Foránea, clave principal de otra tabla relacionada. Ejemplo: id estudiante.
- **Clave Primaria**: identificador único de una tupla.
- **Cardinalidad**: número de tuplas(m).
- **Grado**: número de atributos(n).
- **Dominio**: colección de valores de los cuales el atributo obtiene su valor.

Terminología Relacional		Terminología de Tablas		Terminología de Archivo
Relación	=	Tabla	=	Archivo
Tupla	=	Fila	=	Registro
Atributo	=	Columna	=	Campo
Grado	=	Número de columnas	=	Número de campos
Cardinalidad	=	Número de filas	=	Número de registros



4.4 Reglas de Integridad



4.5 Atributo

Un Atributo en el Modelo Relacional representa una propiedad que posee esa Relación y equivale al atributo del Modelo E-R.

noficina	calle	area	telefono	fax
100	Lyon 2345	Las Condes	964201240	964201340
110	Alameda 234	Santiago Centro	964215760	964215670
120	Luis Thayer Ojeda	Providencia	964520250	964520255
130	Baldomero Lillo 2345	Puente Alto	964284440	
140	Calle Crucero 3456	La Dehesa	965678904	964252811

Se corresponde con la idea de campo o columna.

En el caso de que sean varios los atributos de una misma tabla, definidos sobre el mismo dominio, habrá que darles nombres distintos, ya que una tabla no puede tener dos atributos con el mismo nombre.

Por ejemplo, la información de las oficinas de una empresa inmobiliaria se representa mediante la relación OFICINA, que tiene columnas para los atributos noficina (número de oficina), calle, area, telefono y fax.

Los atributos pueden ser:

- Simples: no están divididos en subpartes (nombre, provincia …)
- Compuestos: se pueden dividir en otros atributos (dirección-> calle, número, ciudad…)
- Monovaluados: Solo puede tener un valor para una entidad (fecha nacimiento)
- Multivaluados: Pueden tener un conjunto de valores para una entidad (número de teléfono)

4.6 Dominio

El dominio dentro de la estructura del Modelo Relacional es el **conjunto de valores que puede tomar un atributo**.

Atributo	Nombre del Dominio	Descripción	Definición
noficina	NUM_OFICINA	Posibles valores de número de oficina	3 caracteres, rango 100 - 990
calle	NOM_CALLE	Nombres de calles y numero de Santiago donde se ubica la oficina	25 caracteres
area	NOM_AREA	Área de Santiago en la que se encuentra ubicada la oficina	20 caracteres
telefono	NUM_TEL_FAX	Números de teléfono de Santiago	9 caracteres
fax	NUM_TEL_FAX	Números de teléfono de Santiago	9 caracteres

- Un dominio contiene todos los posibles valores que puede tomar un determinado atributo. Dos atributos distintos pueden tener el mismo dominio.
- Un domino es un conjunto finito de valores del mismo tipo.
- Los dominios poseen un nombre para poder referirnos a él y así poder ser reutilizable en más de un atributo.

En el ejemplo, la tabla muestra los dominios de los atributos de la relación OFICINA. Nótese que en esta relación hay dos atributos que están definidos sobre el mismo dominio, teléfono y fax.

4.7 Tupla, Grado y Cardinalidad

- **Tupla**: es cada una de las filas de la relación. Representa por tanto el conjunto de cada elemento individual (ejemplar u ocurrencia) de esa tabla. En la relación OFICINA, cada tupla tiene cinco valores, uno para cada atributo. Las tuplas de una relación no siguen ningún orden.
- **Grado**: número de columnas de la relación (número de atributos). La relación OFICINA es de grado seis porque tiene seis atributos. Esto quiere decir que cada fila de la tabla es una tupla con seis valores.
- **Cardinalidad**: número de tuplas de una relación (número de filas). Ya que en las relaciones se van insertando y borrando tuplas a menudo, la cardinalidad de estas varía constantemente.