

Introducción a los **datos**, **bases de datos** y **conjuntos de** **datos**



GOBIERNO DE LA
CIUDAD DE MÉXICO

AGENCIA DIGITAL DE
INNOVACIÓN PÚBLICA

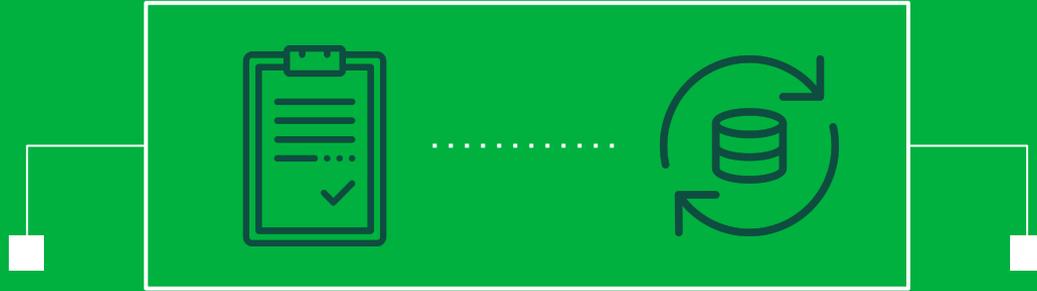


Recordemos qué es la gestión de datos

La gestión de datos se refiere al **conjunto de prácticas, procesos y procedimientos**, tanto tecnológicos y administrativos, que los Entes Públicos llevan a cabo con el objetivo **controlar, proteger, aprovechar e incrementar el valor** de los datos.

*-Política de Gestión de Datos de la
CDMX, Artículo noveno*

Es decir:



Son las reglas y procedimientos que se deben seguir a lo largo de todo el **ciclo de vida de los datos**



Parte 1. ¿Qué son los datos?

¿Qué son los datos?



Son mediciones, registros u observaciones, numéricas o textuales, de algún atributo o característica de un fenómeno, hecho o individuo, o incluso, sobre otros datos.

¿Qué son los datos públicos?

Datos públicos

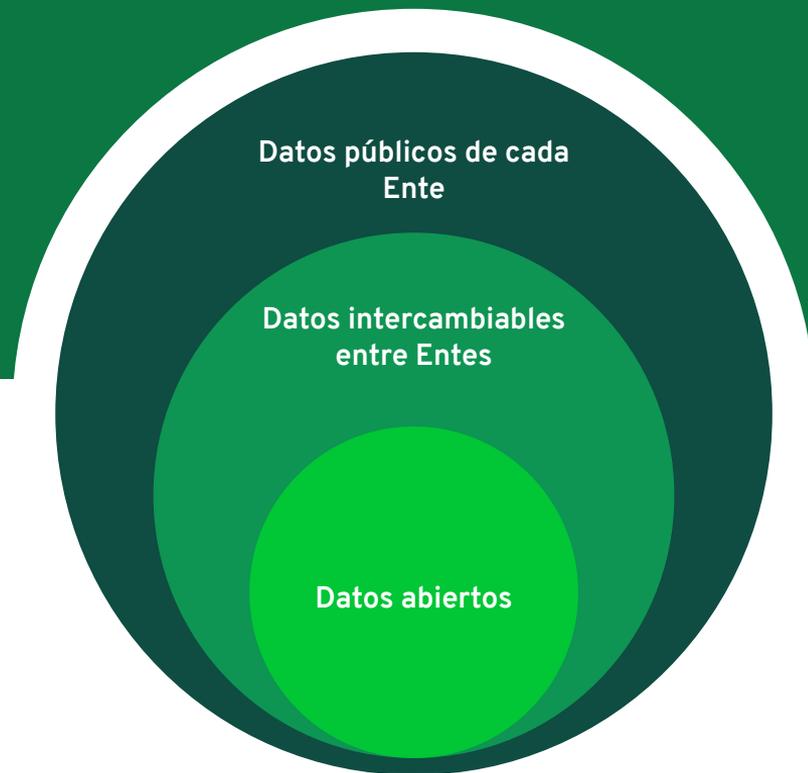


Todos los datos generados y/o recolectados por los Órganos de la Administración Pública de la CDMX en el cumplimiento de sus funciones*

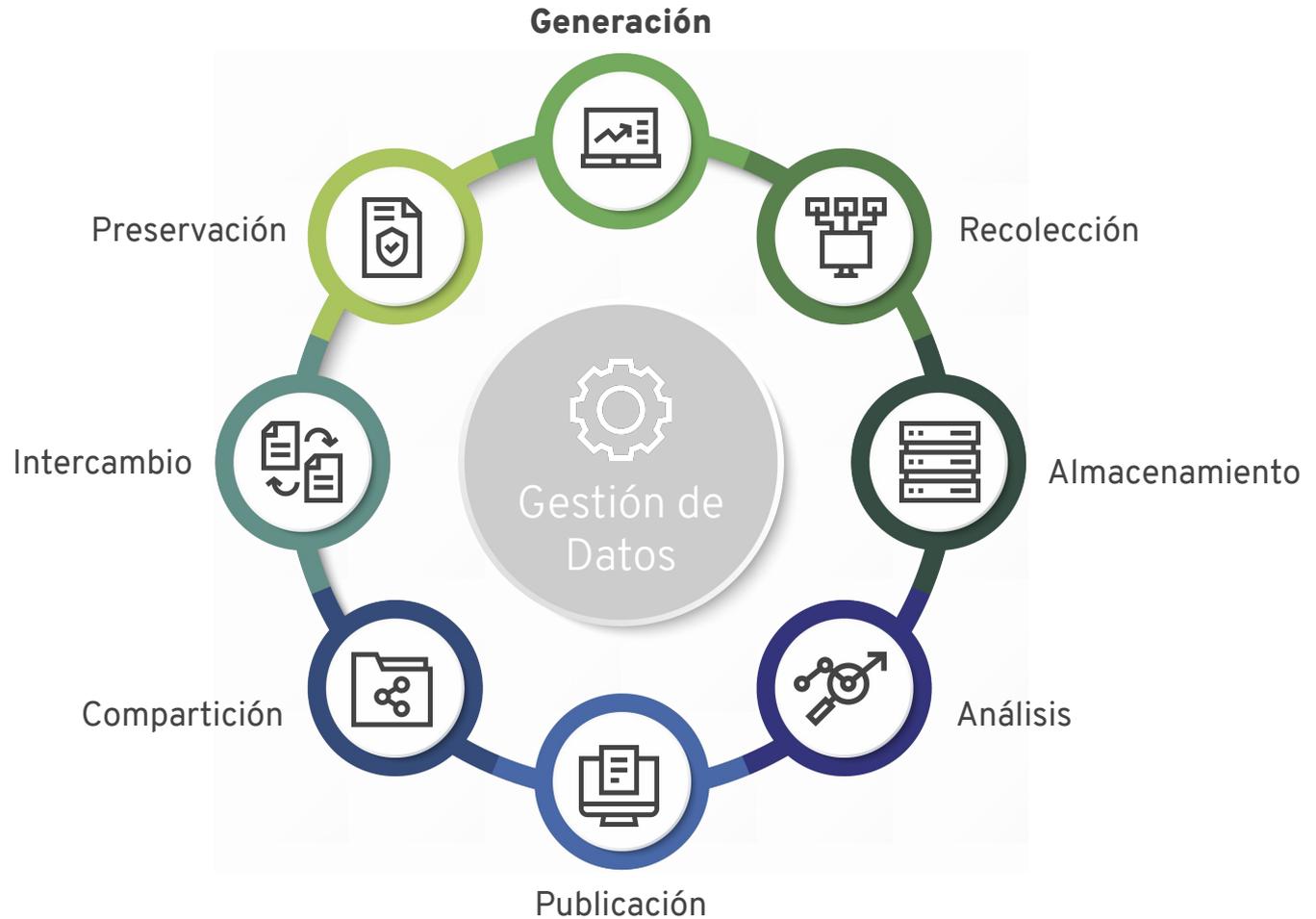
*Según los términos que se establecen en la Ley de Transparencia, Acceso a la Información Pública y Rendición de Cuentas de la Ciudad de México.

Datos públicos =
Propiedad del pueblo

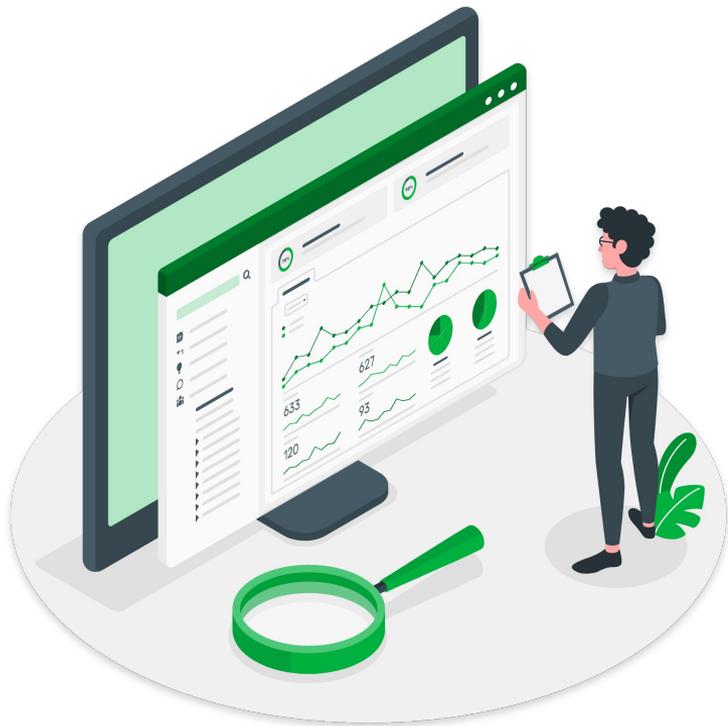
Datos públicos
≠ datos abiertos



¿Cuál es el **ciclo de vida** de los datos?



¿Qué es la **generación y recolección** de los datos?



Generación



Generación es la **acción de producir datos de forma deliberada** mediante procesos metodológicos o técnicos diseñados específicamente para su generación.

Recolección



Recolección es la **acción de recopilar datos ya existentes** producidos como resultado indirecto de otras actividades.

Generación y recolección de los datos

Captured

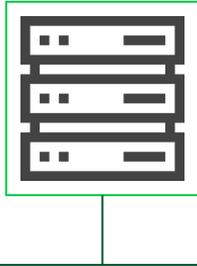
- Se generan de forma deliberada, mediante un proceso diseñado específicamente para su generación o recolección.
- Ejemplo: ENOE, RNIE, Registro de carpetas de investigación, datos de calidad del aire.

Exhaust

- Se recolectan como un producto colateral de otras actividades.
- No son el fin último de la actividad realizada.
- Ejemplo: datos de actividad en redes sociales, datos de uso de un smartphone, datos de inicio de sesión, etc.



¿Qué es el **almacenamiento** de los datos?



Es la acción de **albergar** datos en medios digitales o virtuales.



Almacenamiento de los datos



Digital

- Disco duro interno
- Disco duro externo
- Tarjeta de memoria
- USB
- CD o DVD
- Servidor interno
- Servidor externo



Virtual (Nube)

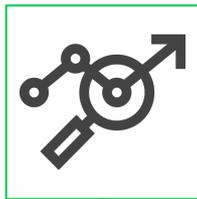
- AWS (Amazon Web Services)
- Microsoft Azure
- Dropbox
- OneDrive
- Google Drive
- Google Cloud
- iCloud
- IBM Cloud
- Oracle Cloud



Red (NAS)

- Network attached storage (almacenamiento conectado en RED)
- Servidor de archivos optimizado para el almacenamiento de datos en redes locales que accede a la red local a través del protocolo de red

Análisis de datos



Es el proceso iterativo de **inspeccionar, limpiar, transformar y modelar** los datos para descubrir información útil que permite hacer conclusiones y tomar decisiones.

Inspección		¿Cómo se ven y cuál es la estructura de los datos?
Limpieza		¿Qué errores o anomalías hay en los datos?
Transformación		¿Cómo convertir el formato o estructura de mis datos a otro para que me sea útil?
Modelado y algoritmos		¿Qué relación hay entre mis datos? ¿Qué patrones encuentro? ¿Qué puedo inferir? ¿Qué puedo predecir?

¿Qué es **compartición** y el **intercambio** de los datos?



≠



Muchas veces se usa indistintamente, sin embargo, no son lo mismo...

Compartición: la acción de transferir datos **al interior** de una misma agrupación.

Intercambio: la acción de transferir datos a miembros u organizaciones ajenas a la propia, **hacia el exterior**.



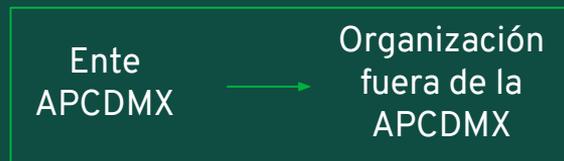
Compartición

Entre Entes de la
APCDMX



Intercambio

Entre Entes y organizaciones
fuera de la APCDMX



Publicación de datos



Es la acción de difundir los datos para que estén disponibles para toda la Ciudadanía.*

En la Ciudad de México, los datos oficiales se publican de manera proactiva a través del **Portal de Datos Abiertos**.

*Según los términos que se establecen en la Ley de Transparencia, Acceso a la Información Pública y Rendición de Cuentas de la Ciudad de México.

Datos públicos

Todos los datos generados o recolectados por los Entes de la APCDMX en el ejercicio de sus funciones.



¿Tiene información que es reservada?

Si la respuesta es sí, ésta debe removerse de los datos para la elaboración de **versiones públicas**.



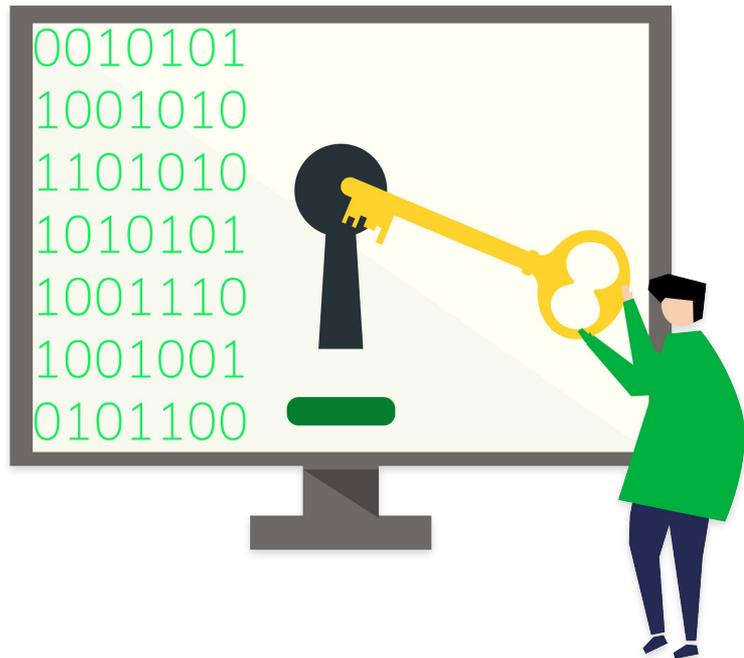
¿Tiene datos personales?

Si la respuesta es sí, los datos deben pasar por un proceso de **ANONIMIZACIÓN**.



Datos elegibles para ser publicados en el portal de datos abiertos

Ley de Transparencia, acceso a la información pública y rendición de cuentas de la Ciudad de México



Artículo 27

La aplicación de esta ley, deberá de interpretarse bajo el **principio de máxima publicidad** y en caso de duda razonable entre la publicidad y la reserva de la información, deberá favorecerse el principio de máxima publicidad.

Siempre que sea posible, se elaborarán **versiones públicas** de los documentos que contengan información clasificada como reservada o confidencial.

Preservación de datos



La acción de salvaguardar la información a lo largo del tiempo para que pueda seguir usándose en el futuro

Por
ejemplo:



- Archivo General de la Nación



- Registro Público



- Ley de Archivos



Parte 2. Conjuntos de datos vs Bases de datos

Ya tengo mis datos ¿ahora qué?



1. Generación y recopilación de datos

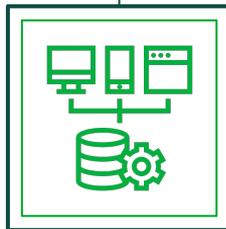
2. Organización de los datos en un conjunto de datos

Ya tengo mis datos.
¿Cuál es la mejor forma de organizarlos y almacenarlos?



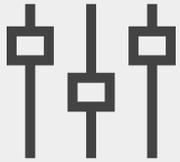
3. Modelado de una base de datos

Si el caso lo amerita (estructuras más complejas que una tabla), el conjunto de datos se pueden organizar y almacenar en una base de datos.



¿Por qué es importante la **distinción** entre conjuntos de datos y bases de datos?

Grado de complejidad:



Un conjunto de datos puede ser muy sencillo, como una tabla de excel, o muy complejo, como una base de datos.

Es decir, **una base de datos es un conjunto de datos con un modelado lógico.**



Todas las bases de datos son conjuntos de datos, pero **no todos los conjuntos de datos son bases de datos.**

Dataset (conjunto de datos)



- Los datos se organizan mediante tablas.
- Se pueden registrar manualmente en un archivo (como excel).
- Se pueden registrar automáticamente en una base de datos.

Nombre	Apellido	Edad
Juan	Morales	47
Roberta	Aguilera	25

Datos ordenados en **conjuntos de datos**

Un conjunto de datos es una colección de **valores**, usualmente números o caracteres. Los valores se organizan en:



- **Atributo o campo**



- **Registro u observación**

Atributo o campo: Contiene todos los valores que miden el mismo atributo a través de unidades. Cada atributo debe tener su propia fila.

Registro u observación: Contiene todos los valores medidos en la misma unidad a través de atributos. Cada registro es una fila.

Valor: Cada una de las características de los atributos registrada o cuantificada. Cada valor, una celda.

pais	anio	casos	poblacion
Afganistán	1999	745	19987071
Afganistán	2000	2666	20595360
Brasil	1999	37737	172006362
Brasil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

■ Atributos

pais	anio	casos	poblacion
Afganistán	1999	745	19987071
Afganistán	2000	2666	20595360
Brasil	1999	37737	172006362
Brasil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

■ Registros

pais	anio	casos	poblacion
Afganistán	1999	745	19987071
Afganistán	2000	2666	20595360
Brasil	1999	37737	172006362
Brasil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

■ Valores

Así se ve en una **hoja de cálculo** de excel

The image shows the Microsoft Excel interface. The ribbon at the top includes 'ARCHIVO', 'INICIO', 'INSERTAR', 'DISEÑO DE PÁGINA', 'FÓRMULAS', 'DATOS', 'REVISAR', and 'VISTA'. The 'INICIO' ribbon is active, showing options for font (Fuente), alignment (Alineación), and numbers (Número). The cell 'E8' is selected, and its address is shown in the Name Box. The spreadsheet grid shows columns A through G and rows 1 through 15. The column header 'E' is highlighted with a red box and labeled 'Columna'. The row header '8' is highlighted with a red box and labeled 'Fila'. The cell 'E8' is highlighted with a red box and labeled 'Celda'. The status bar at the bottom shows 'Hoja1' as the active sheet.

¿Quieres saber más?



Consulta las *Recomendaciones básicas para tener datos de calidad en hojas de cálculo* en nuestro **Micrositio**

Database (base de datos)

- Son un conjunto de datos albergado y administrado por un sistema informático.
- Requieren de un sistema gestor de base de datos para acceder y usar los datos (SGBD).
- Tienen un modelo de datos, de distintas estructuras como:

Relacional
(SQL)

Orientada a
objetos (NoSQL)

Jerárquica

Red

Database (base de datos)

¿Qué son los modelos de datos?

Son distintos lenguajes para definir y manipular una base de datos.

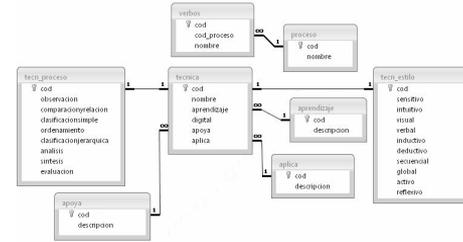
Definir: estructura (tipo de datos que hay y su forma de relacionarse entre ellos).

Manipular: operaciones de manipulación de datos (agregar, borrar, modificar, etc).

Principales **modelos** de bases de datos

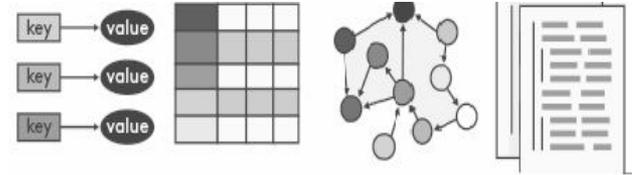
01

Relacional (SQL)



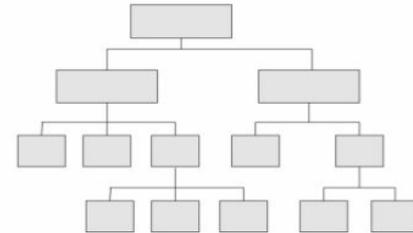
02

NoSQL y orientadas a objetos



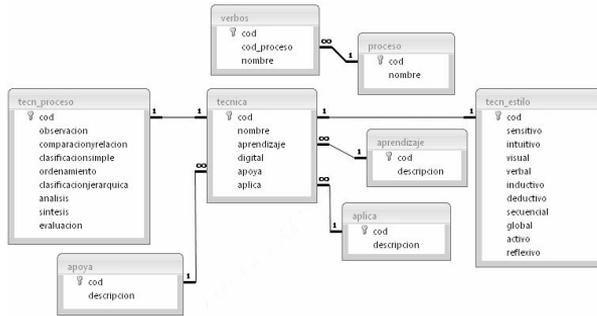
03

Jerárquico y de Red



01. Relacional (SQL)

- Uno de los modelos de base de datos más usado.
- Organiza los datos en tablas que se relacionan entre sí mediante valores clave.



- Una base de datos tiene varias tablas.
- Cada fila o registro tiene un identificador único que funciona como valor clave.
- Las columnas tienen los atributos o características de los datos.

01. Relacional (SQL)



- Generalmente se accede a los datos mediante SQL (Structured query language) es decir, mediante entradas o consultas que “devuelven” un resultado.
- Utilizan sistemas gestores de base de datos como:

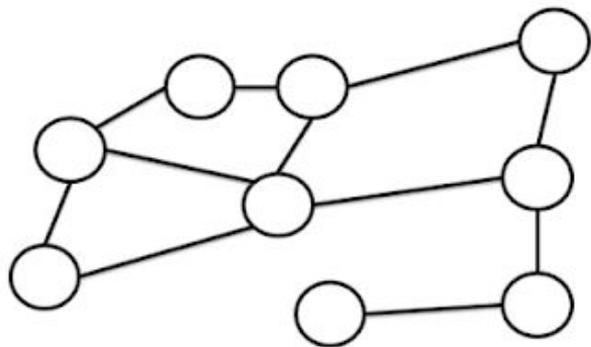
□ MySQL

□ PostgreSQL

□ Microsoft SQL Server

□ Oracle SQL

02. NoSQL y orientada a objetos



NoSQL = Not Only SQL = bases de datos que no sólo utilizan SQL

- Útiles para procesar datos no estructurados o semiestructurados.
- Permiten almacenaje, uso y acceso flexible a los datos.

Ejemplos de bases de datos NoSQL:

- Grafos
- Llave-valor
- Orientada a documentos

02. NoSQL y orientada a objetos

Base de datos de grafos



- Su principal característica es que representan la información en vértices y aristas.
- Un grafo está compuesto por dos elementos: nodos (vértices) y relaciones (aristas).
- Los nodos representan entidades, allí almacenaremos atributos de tipo clave-valor, y las relaciones representan cómo se conectan y se asocian dos nodos.

Ejemplos:

Neo4j

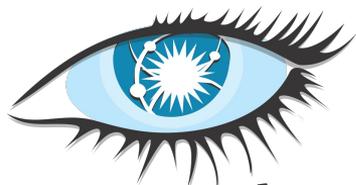
OrientDB

InfiniteGraph



02. NoSQL y orientada a objetos

Base de datos llave-valor



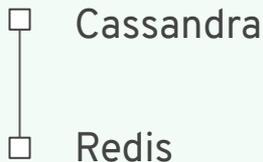
cassandra



redis

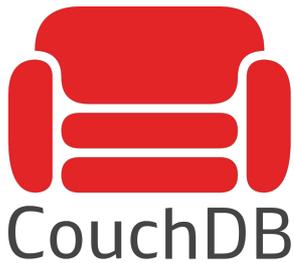
- Utiliza un método de clave o llave-valor para almacenar datos. Las llaves o clave representan una característica identificativa única.
- Gozan del aprecio de desarrolladores y programadores por ser muy efectivas en la consulta y fáciles de escalar.

Ejemplos:



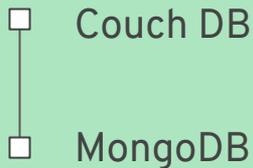
02. NoSQL y orientada a objetos

Base de datos orientada a documentos



- Se usan para almacenar y consultar datos. Los documentos puede ser de texto, archivos XML o JSON.
- Tiene una naturaleza flexible, semiestructurada y jerárquica de los documentos.
- Funciona bien con casos de uso como catálogos, perfiles de usuario y sistemas de administración de contenido.

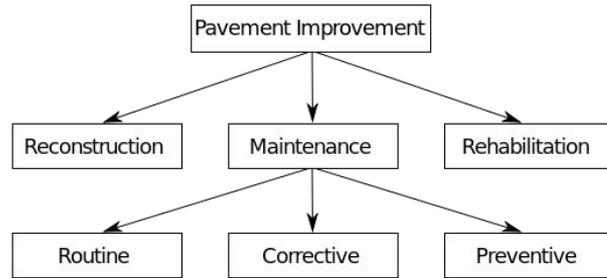
Ejemplos:



03. Jerárquico y de Red

Jerárquico

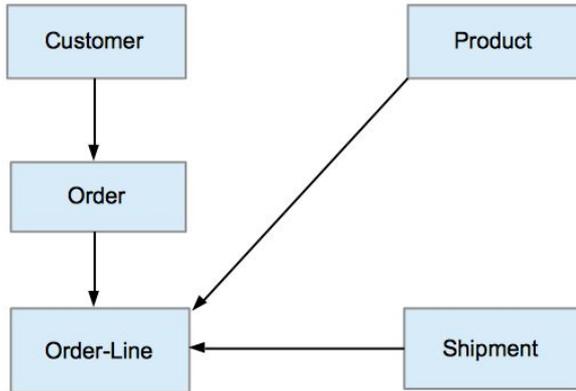
Hierarchical Model



- Representa objetos y sus relaciones, donde las dependencias responden a una jerarquía.
- Hay tipos de objetos “padres” que tienen “hijos”.
- Se ve como un árbol.

03. Jerárquico y de Red

Red



Representa objetos y sus relaciones, sin ningún tipo de jerarquía.

Su esquema consiste en:

- Tipos de objetos: nodos
- Relaciones: arcos



GOBIERNO DE LA
CIUDAD DE MÉXICO

AGENCIA DIGITAL DE
INNOVACIÓN PÚBLICA



Parte 3. Sistemas de Información

Sistema de Información



- Conjunto integrado de componentes de recursos de tecnologías de la información* que tiene como objeto recolectar, almacenar y/o procesar datos.
- Suelen consultar o escribir información en una base de datos.

Bases de datos del Sistema "x":

Tables

```
apps_api
catalogo_alcaldias_municipios
catalogo_colonias_cps
intentos_fallidos
migrations
modulos_atencion
password_resets
permiso_user
permisos
role_user
roles
```

Ejemplos de sistemas de información:

- Sistema de Información, Evaluación y Seguimiento (SIEV).
- Sistema Único de Información Delictiva (SUIDF).
- Sistema Integral de Administración de Personal (SIAP).

*Recursos de Tecnologías de Información: Los conjuntos de datos, bases de datos, las aplicaciones y/o sistemas de información digitales, la tecnología (hardware, software, sistemas operativos, sistemas manejadores de bases de datos, redes, y demás aplicables) e instalaciones (recursos para alojar y dar soporte a los sistemas de información).

Dirección de Coordinación de Proyectos de Datos



politicadedatos@cdmx.gob.mx

Contacto



GOBIERNO DE LA
CIUDAD DE MÉXICO

AGENCIA DIGITAL DE
INNOVACIÓN PÚBLICA