UNIDAD 5

DATOS DE OTRAS FUENTES

DATOS DE OTRAS FUENTES

Contenido de la Unidad

- **Datos Semiestructurados**
- Bases de Datos Documentales
- Estrategias para Integración de Datos Desde Múltiples Fuentes
- Resumen

DATOS DE OTRAS FUENTES

Datos estructurados

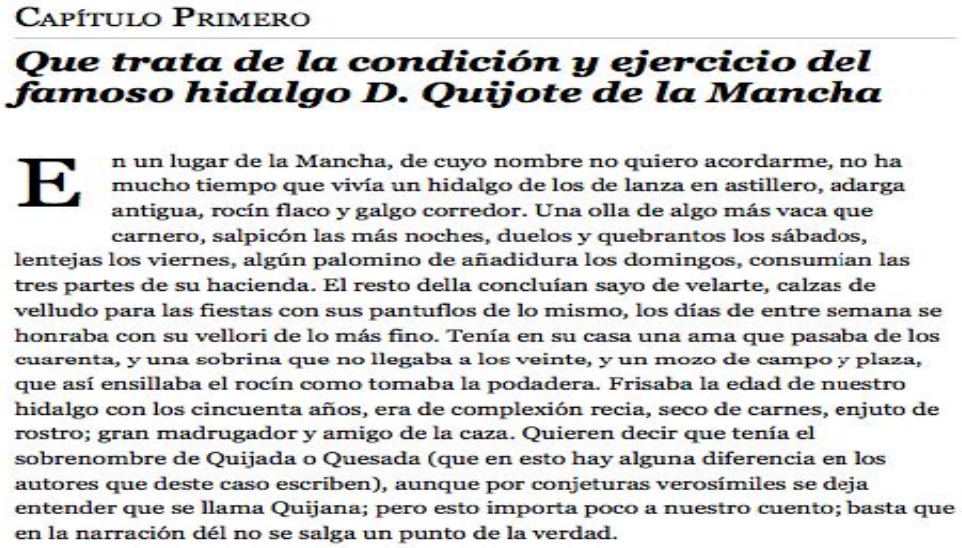
Los datos estructurados tienen perfectamente definido la longitud, el formato y el tamaño de sus datos. Se almacenan en formato tabla, hojas de cálculo o en bases de datos relacionales .

	nombre	color	edad	altura	peso	puntuacion
1:	Paco	Rojo	24	182	74.8	83
2:	Juan	Green	30	170	70.1	500
3:	Andres	Amarillo	41	169	60.0	20
4:	Natalia	Green	22	183	75.0	865
5:	Vanesa	Verde	31	178	83.9	221
6:	Miriam	Rojo	35	172	76.2	413
7:	Juan	Amarillo	22	164	68.0	902

DATOS DE OTRAS FUENTES

Datos no estructurados

Los datos no estructurados se caracterizan por no tener un formato específico. Se almacenan en múltiples formatos como documentos PDF o Word, correos electrónicos, archivos multimedia de imagen, audio o video,...



CAPÍTULO PRIMERO

Que trata de la condición y ejercicio del famoso hidalgo D. Quijote de la Mancha

En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigua, rocín flaco y galgo corredor. Una olla de algo más vaca que carnero, salpicón las más noches, duelos y quebrantos los sábados, lentejas los viernes, algún palomino de añadidura los domingos, consumian las tres partes de su hacienda. El resto della concluían sayo de velarte, calzas de velludo para las fiestas con sus pantuflos de lo mismo, los días de entre semana se honraba con su vellori de lo más fino. Tenía en su casa una ama que pasaba de los cuarenta, y una sobrina que no llegaba a los veinte, y un mozo de campo y plaza, que así ensillaba el rocín como tomaba la podadera. Frisaba la edad de nuestro hidalgo con los cincuenta años, era de complexión recia, seco de carnes, enjuto de rostro; gran madrugador y amigo de la caza. Quieren decir que tenía el sobrenombre de Quijada o Quesada (que en esto hay alguna diferencia en los autores que deste caso escriben), aunque por conjeturas verosímiles se deja entender que se llama Quijana; pero esto importa poco a nuestro cuento; basta que en la narración dél no se salga un punto de la verdad.

DATOS DE OTRAS FUENTES

Datos semiestructurados

Los datos semi estructurados son una mezcla de los dos anteriores, no presentan una estructura perfectamente definida como los datos estructurados, pero sí presentan una organización definida en sus metadatos donde describen los objetos y sus relaciones, y que en algunos casos están aceptados por convención, como por ejemplo los formatos HTML, XML o JSON.

```
{
    "marcadores": [
      {
        "latitude": 40.416875,
        "longitude": -3.703308,
      },
      {
        "latitude": 40.417438,
        "longitude": -3.693363,
        "description": "Paseo del Prado"
      },
      {
        "latitude": 40.407015,
        "longitude": -3.691163,
        "city": "Madrid",
        "description": "Estación de
          Atocha"
      }
    ]
}
```

DATOS DE OTRAS FUENTES

Datos semiestructurados - XML

El lenguaje de marcado extensible (XML) es un lenguaje de marcado que proporciona reglas para definir cualquier dato. A diferencia de otros lenguajes de programación, XML no puede realizar operaciones de computación por sí mismo.
En cambio, se puede implementar cualquier software o lenguaje de programación para la administración de datos.

Respaldo para las transacciones interempresariales
Cuando una empresa vende un bien o servicio a otra empresa, las dos empresas necesitan intercambiar información como el costo, las especificaciones y los plazos de entrega.
Con el lenguaje de marcado extensible (XML), pueden compartir toda la información necesaria electrónicamente y cerrar negocios complejos de forma automática, sin intervención humana.

DATOS DE OTRAS FUENTES

Datos semiestructurados - XML

Etiquetas XML

Los símbolos de marcado, denominados etiquetas en XML, se utilizan para definir los datos. Por ejemplo, para representar los datos de una librería, puede crear etiquetas como <libro>, <título> y <autor>.
El documento XML de un solo libro tendría el siguiente contenido:

```
<libro>
    <título>Introducción a Amazon Web Services</título>
    <autor>Mark Wilkins</autor>
</libro>
```

Las etiquetas ofrecen una sofisticada codificación de datos para integrar los flujos de información en diferentes sistemas.

DATOS DE OTRAS FUENTES

Datos semiestructurados - JSON

JSON es un formato estándar para datos que destaca por ser ligero y rápido (por tanto muy útil para desarrollos web). Los datos en formato JSON pueden ser utilizados por prácticamente todos los lenguajes de programación (como Java, C#, C, C++, PHP, JavaScript, Python, etc.).

Los archivos JSON son simples archivos de texto con extensión json. Por ejemplo, un nombre de archivo podría ser estudiantes.json.

El contenido del mismo son datos representados en un conjunto de pares clave:valor, igualmente que un objeto JavaScript tradicional.

DATOS DE OTRAS FUENTES

Datos semiestructurados - JSON

Sintaxis básica de JSON

```
{
  "llave1": "valor1",
  "llave2": "valor2",
  "llave3": "valor3",
  "llave4": 7,
  "llave5": null,
  "favAmigos": ["Kolade", "Nithya", "Dammy", "Jack"],
  "favJugadores": {"uno": "Kante", "dos": "Hazard", "tres": "Didier"}
}
```

DATOS DE OTRAS FUENTES

Datos semiestructurados

Dudas / Preguntas

DATOS DE OTRAS FUENTES

Contenido de la Unidad

- Datos Semiestructurados.
- **Bases de Datos Documentales.**
- Estrategias para Integración de Datos Desde Múltiples Fuentes.
- Resumen.

DATOS DE OTRAS FUENTES

Bases de Datos Documentales

Una **base de datos documental** es una de las principales variantes de las bases de datos no relacionales o NoSQL. Se caracterizan por almacenar la información en registros, cada uno de los cuales funciona como una unidad autónoma de información.

DATOS DE OTRAS FUENTES

Bases de Datos Documentales

Como su propio nombre indica, las **bases de datos orientadas a documentos** utilizan documentos para el almacenamiento de todos los registros y los datos asociados a ellos. Cada uno de estos registros puede almacenar distintos tipos de datos. A su vez, los documentos que contienen los registros pueden tener diferentes formatos, desde archivos JSON o XML hasta documentos de texto.

DATOS DE OTRAS FUENTES

Bases de Datos Documentales

Ventajas

Las principales **ventajas de las bases de datos documentales** se resumen en las siguientes:

- Permiten almacenar y consultar información semiestructurada sin una estructura definida.
- Son un modelo muy flexible que puede albergar numerosos tipos de datos.
- Simplifican las tareas de adición o actualización de datos. La mayoría de aplicaciones web o móviles están sometidas a cambios constantes. Gracias a las bases de datos documentales se pueden añadir nuevos datos o modelos de análisis de manera mucho más flexible.
- Tienen una gran escalabilidad y son uno de los mejores métodos para el almacenamiento de grandes volúmenes de información.

DATOS DE OTRAS FUENTES

Bases de Datos Documentales

Ventajas

Las principales **ventajas de las bases de datos documentales** se resumen en las siguientes:

- Aseguran una escritura rápida, dando prioridad a la disponibilidad de la escritura sobre la consistencia de los datos. Esto permite asegurar la rapidez incluso en casos de fallos en el hardware o en la red, que en otras bases de datos supondría retrasos en la modificación de los datos y repercutirá negativamente en su coherencia.
- Garantizan un buen rendimiento. La mayoría de bases de datos documentales cuentan con potentes motores de búsqueda y avanzadas propiedades de indexación, lo que asegura una mayor rapidez a la hora de consultar la información.

DATOS DE OTRAS FUENTES

Bases de Datos Documentales

Desventajas

No todo iban a ser puntos positivos. Estas son las principales **desventajas de las bases de datos documentales**:

- No utilizan el lenguaje SQL como lenguaje principal de consulta, aunque sí lo pueden usar de apoyo. Es decir, al contrario que las bases relacionales, no existe un lenguaje estandarizado para la creación de estas bases de datos.
- No siempre pueden garantizar las propiedades ACID de atomicidad, consistencia, integridad y durabilidad.

DATOS DE OTRAS FUENTES

Bases de Datos Documentales

Desventajas

No todo iban a ser puntos positivos. Estas son las principales **desventajas de las bases de datos documentales**:

- Al ser mas “nuevas” que las bases de datos relacionales, inicialmente no contaban con una gran comunidad y existe menos información y recursos para este tipo de bases de datos.
- Los índices pueden ocupar mucha memoria RAM, sobre todo en las bases documentales que manejan un gran volumen de datos.

DATOS DE OTRAS FUENTES

Bases de Datos Documentales

Diferencia entre base de datos documental y relacional

Por un lado, las bases de datos documentales **no emplean el lenguaje SQL**, o si lo utilizan es solo como apoyo. Por el contrario, emplean otro tipo de formatos para la información almacenada, tales como XML o JSON.

Relativo al punto anterior, cabe destacar que, a diferencia de las bases de datos relacionales, **no existe un estándar definido** para las bases de datos documentales (el SQL está definido como lenguaje estándar para las bases de datos por el ANSI).

DATOS DE OTRAS FUENTES

Bases de Datos Documentales

A continuación, veremos un pequeño ejemplo en formato JSON sobre cómo construir una librería de música muy sencilla:

```
{
  '_id' : 1,
  'artistName' : { 'Oasis' },
    'albums' : [
{
'albumname' : 'What´s the story morning glory',
'datereleased' : 1996,
'genre' : 'BritPop'
}, {
    'albumname' : 'Definately Maybe',
    'datereleased' : 1992,
    'genre' : 'Rock'
}, {
    'albumname' : 'Be Here now',
    'datereleased' : 1999,
    'genre' : BritPop
 }, {
 }
]
```

DATOS DE OTRAS FUENTES

Bases de Datos Documentales

Dudas / Preguntas

DATOS DE OTRAS FUENTES

Contenido de la Unidad

- Datos Semiestructurados.
- Bases de Datos Documentales.
- **Estrategias para Integración de Datos Desde Múltiples Fuentes.**
- Resumen.

DATOS DE OTRAS FUENTES

Estrategias para Integración de Datos Desde Múltiples Fuentes

Los métodos de integración de datos más comunes se pueden agrupar en cinco categorías:

- **Integración manual de datos**: todas las fases de la integración, desde la recopilación hasta la visualización, se realizan de forma manual.
- **Integración uniforme de acceso a datos**: una técnica que recupera y muestra datos de manera uniforme mientras los datos se mantienen en su fuente original.
- **Integración de datos de middleware**: integra datos de sistemas heredados en una nueva capa de middleware.

DATOS DE OTRAS FUENTES

Estrategias para Integración de Datos Desde Múltiples Fuentes

- **Integración de almacenamiento de datos común**: una copia de los datos de la fuente se almacena en una nueva base de datos o almacén de datos para presentar los datos de manera uniforme.
- **Integración basada en aplicaciones**: las aplicaciones de software extraen e integran datos armonizando datos de diferentes fuentes y sistemas.

DATOS DE OTRAS FUENTES

Dudas / Preguntas

DATOS DE OTRAS FUENTES

Contenido de la Unidad

- Datos Semiestructurados.
- Bases de Datos Documentales.
- Estrategias para Integración de Datos Desde Múltiples Fuentes.
- **Resumen.**

DATOS DE OTRAS FUENTES

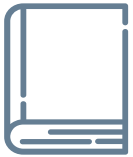
Resumen de la Unidad

Durante esta unidad hemos aprendido:

- Datos semiestructurados
- Bases de datos documentales
- Estrategias para integración de datos desde múltiples fuentes

DATOS DE OTRAS FUENTES

Bibliografía Recomendada



- Introducción a las bases de datos NoSQL usando MongoDB - Edición en Español de Antonio Sarasa Cabezuelo
- APRENDE MONGODB CON NO-SQL DESDE PRINCIPIANTE A EXPERTO - Edición en Español de CARMELO RAMOS SERRANO
- The Data Integration Guide: How to design, deliver, deploy, and sustain efficient data integration solutions in your information system de Ahmed Fessi