

Exploración de los aspectos básicos de los datos



Agenda



- Conceptos básicos de datos
- Roles de datos y servicios

Objetivos de aprendizaje

Tras finalizar este módulo, podrá:

- 1 Identificar formatos de datos comunes.
- 2 Describir las opciones para almacenar datos en archivos y bases de datos.
- 3 Describir las características de las soluciones de procesamiento de datos transaccionales y analíticos.
- 4 Identificar los roles comunes de los profesionales de datos.
- 5 Identificar los servicios en la nube comunes que usan los profesionales de datos.

1: Conceptos básicos de los datos



¿Qué son los datos?

Valores usados para registrar información, a menudo representando *entidades* que tienen uno o varios *atributos*

Estructurado

Cliente				
ID	FirstName	LastName	Correo electrónico	Dirección
1	Joe	Jones	joe@litware.com	1 Main St.
2	Samir	Nadoy	samir@northwind.com	123 Elm Pl.

Producto		
ID	Nombre	Price
123	Martillo	2,99
162	Screwdriver	3,49
201	Llave	4,25

Semiestructurados

```
{
  "firstName": "Joe",
  "lastName": "Jones",
  "address": {
    "streetAddress": "1 Main
St.",
    "city": "New York",
    "state": "NY",
    "postalCode": "10099"
  },
  "contact": [
    {
      "type": "home",
      "number": "555 123-
",
      "address": {
        "streetAddress": "123 Elm
Pl.",
        "unit": "500",
        "city": "Seattle",
        "state": "WA",
        "postalCode": "98999"
      }
    },
    {
      "type": "email",
      "address": "joe@litware.com"
    }
  ]
}
```

No estructurado


Estimado Joe:

Gracias por pedir sus suministros de hardware desde nuestra tienda en línea (número de pedido 1000) el 1/1/2022.


Su pedido se ha enviado y debe llegar en 3-5 días laborables.

Hardware de Contoso

Nuestros productos son de la más alta calidad y los utilizan los profesionales. Tenemos destornilladores increíbles, que son realmente útiles para apretar y aflojar tornillos.



También tenemos llaves inglesas (o, si lo prefiere, llaves)...



¿Cómo se almacenan los datos?

Archivos

Texto delimitado

```
FirstName, LastName, Email  
Joe, Jones, joe@litware.com  
Samir, Nadoy, samir@northwind.com
```

Notación de objetos JavaScript (JSON)

```
{  
  "customers":  
  [  
    { "firstName": "Joe", "lastName": "Jones"},  
    { "firstName": "Samir", "lastName": "Nadoy"}  
  ]  
}
```

Lenguaje de marcado extensible (XML)

```
<Customer firstName="Joe" lastName="Jones"/>
```

Objeto binario grande (BLOB)

```
10110101101010110010...
```

Formatos optimizados:

- Avro, ORC y Parquet

Bases de datos

Relacional



Cliente		
ID	Correo electrónico	Dirección
1	joe@litware.com	1 Main St.
2	samir@northwind.com	123 Elm Pl.

Producto		
ID	Nombre	Price
123	Martillo	2,99
162	Screwdriver	3,49
201	Llave	4,25

Pedido		
OrderNo	OrderDate	Cliente
1000	1/1/2022	1
1001	1/1/2022	2

Linitem			
OrderNo	ItemNo	ProductID	Quantity
1000	1	123	1
1000	2	201	2
1001	1	123	2

No relacional

Products	
Key	Value
123	"Hammer (\$2.99)"
162	"Screwdriver (\$3.49)"
201	"Wrench (\$4.25)"

Key-Value

Customers	
Key	Document
1	{ "name": "Joe Jones" }
2	{ "name": "Samir Nadoy" }

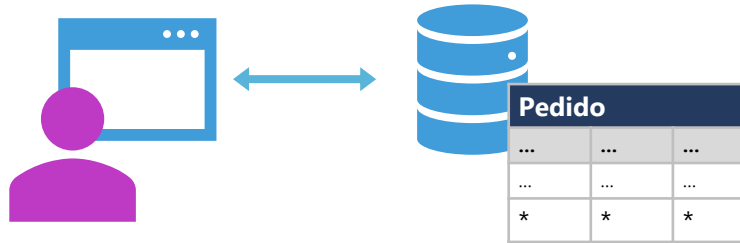
Document

Orders				
Key	Customer		Product	
	Name	Address	Name	Price
1000	Joe Jones	1 Main St.	Hammer	2.99
1001	Samir Nadoy	123 Elm Pl.	Wrench	4.25

Column Family



Cargas de trabajo de datos operativos



Los datos se almacenan en una base de datos optimizada para operaciones de *procesamiento de transacciones en línea* (OLTP) que admite aplicaciones

Una combinación de actividad de *lectura* o *escritura*

Por ejemplo:

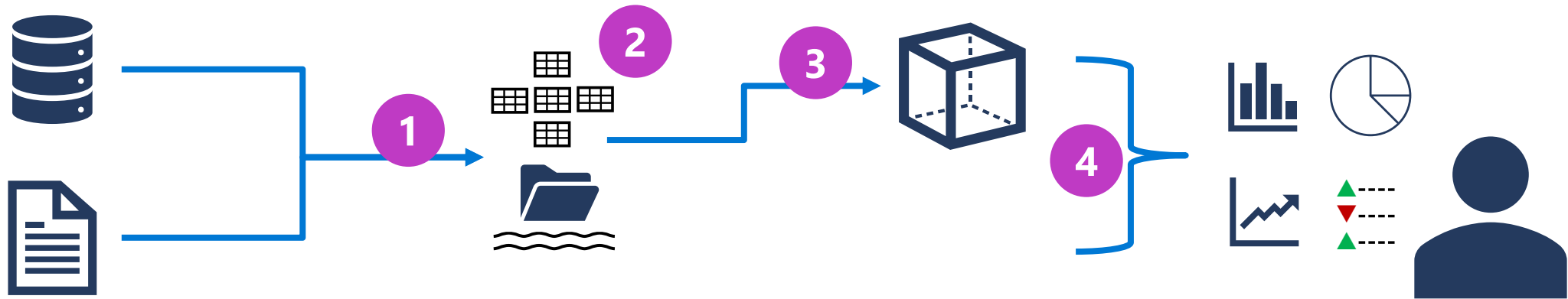
- Leer la tabla *Product* para mostrar un catálogo
- Escribir en la tabla *Order* para registrar una compra

Los datos se almacenan mediante *transacciones*

Las transacciones se basan en "ACID":

- **Atomicidad** : cada transacción se trata como una unidad única de trabajo, la cual se completa correctamente o produce un error general
- **Coherencia** : las transacciones solo pueden pasar los datos de la base de datos de un estado válido a otro
- **Aislamiento** : las transacciones simultáneas no pueden interferir entre sí
- **Durabilidad** : cuando una transacción se ha realizado con éxito, los cambios en los datos se conservan en la base de datos

Cargas de trabajo de datos analíticos



- 1 Los datos operativos se extraen, transforman y cargan (ETL) en un *lago de datos* para su análisis
- 2 Los datos se cargan en un esquema de tablas: normalmente en un *almacén de lago de datos* basado en Spark con abstracciones tabulares en los archivos del lago de datos o en un almacenamiento de datos con un motor SQL totalmente relacional
- 3 Los datos de las tablas se pueden agregar y cargar en un modelo de procesamiento analítico en línea (OLAP) o cubo
- 4 Los archivos del lago de datos, las tablas relacionales y el modelo analítico se pueden consultar para generar *informes y paneles*

2: Roles y servicios de datos



Roles profesionales de datos



Administrador de base de datos

- Aprovisionamiento, configuración y administración de bases de datos
- Seguridad de la base de datos y acceso de usuario
- Copias de seguridad y resistencia de la base de datos
- Supervisión y optimización del rendimiento de la base de datos



Ingeniero de datos

- Canalizaciones de integración de datos y procesos ETL
- Limpieza y transformación de datos
- Esquemas y cargas de datos del almacén de datos analíticos



Analista de datos

- Modelado analítico
- Informes y resumen de datos
- Visualización de datos

Servicios en la nube de Microsoft para datos

Cargas de trabajo de datos operativos



Detección de amenazas

- Familia de servicios de bases de datos relacionales basados en SQL Server



Azure Database para código abierto

- Maria DB, MySQL, PostgreSQL



Azure Cosmos DB

- Sistema de base de datos no relacional altamente escalable



Azure Storage

- Archivo, blob y almacenamiento de tabla
- Espacio de nombres jerárquico para almacenamiento en un lago de datos

Cargas de trabajo de datos analíticos

Software como servicio (SaaS)



Microsoft Fabric

Plataforma analítica unificada, basada en SaaS, basada en un almacén de lago abierto y gobernado:

- Ingesta de datos y ETL
- Almacén de lago de datos
- Data Warehouse
- Ciencia de datos y ML
- Análisis en tiempo real
- Visualización de datos
- Gobernanza y administración de datos

Plataforma como servicio (PaaS)



Azure Synapse Analytics

- Solución integrada para el análisis de datos en Azure
- Canalizaciones, Apache Spark, SQL, Data Explorer



Azure Databricks

- Análisis y procesamiento de datos de Apache Spark



Azure HDInsight

- Plataforma de código abierto de Apache

Otros...

