Datos, información y conocimiento

Autores: Marc Álvarez Brotons
Carlos Calonge Cebrián

Índice

Marc Álvarez Brotons

Ingeniero en Informática por la Universitat Oberta de Catalunya (UOC) y máster en Project Management por La Salle Business Engineering School. Gerente en el Departamento de Sistemas de Gestión de la Información de una entidad financiera, liderando las funciones de *data content management* y *data quality*. Profesor colaborador de la UOC.

Carlos Calonge Cebrián

Ingeniero en Informática por la Universitat Autònoma de Barcelona (UAB) y postgrado en Dirección de Sistemas de Información por la Universitat Politècnica de Catalunya (UPC). Arquitecto de soluciones empresariales en una multinacional tecnológica y especialista en soluciones de *business intelligence*, CRM y Cloud. Profesor colaborador de la UOC.

Introducción

Los datos nos han acompañado siempre a lo largo de la historia, siendo un elemento fundamental para la construcción de nuestro conocimiento, ya sea personal o colectivo. Ese conocimiento adquirido a través del dato nos permite, tras procesarlo, reflexionar y tomar decisiones.

La palabra ***dato*** proviene del latín *datum*, que significa 'lo que es dado' o simplemente 'lo dado'. Es el participio pasado neutro del verbo *dare*, que significa 'dar'. En este sentido, *datum* se refiere a algo que ha sido dado o presentado y, en contextos modernos, se utiliza para referirse a una pieza individual de información.

Un dato se puede definir como una representación simbólica o numérica de un hecho o evento. Es una unidad básica de información que se utiliza para describir, medir o representar algo en un entorno determinado.

Los datos pueden ser de diferentes tipos (por ejemplo, números, texto, sonido, imágenes o vídeos) y pueden obtenerse a partir de diversas fuentes, como mediciones, registros o transacciones. Estos por sí solos carecen de significado, pero cuando se procesan, estructuran y analizan dentro de un contexto pueden transformarse en información útil.

Tomando como ejemplo una vivienda actual, sus distintos elementos y dispositivos generan una gran cantidad de datos, tales como consumos de agua, electricidad y gas, así como el funcionamiento de las instalaciones de aire acondicionado y calefacción y de determinados electrodomésticos (frigorífico, lavadora, TV, etc.) o de sistemas de seguridad y alarmas. Una adecuada organización de esta información puede facilitar la toma de decisiones para optimizar el consumo y reducir el gasto mensual.

Figura 1. Ejemplo de datos generados por una casa inteligente

Fuente: https://universoabierto.org/2018/05/28/la-conquista-del-hogar-digital-hogar-inteligente-negocio-inteligente/

La importancia de los datos radica en que son la materia prima para el análisis, la toma de decisiones y la generación de conocimiento. A medida que la tecnología y la digitalización avanzan, la cantidad de datos disponibles ha crecido exponencialmente, lo que ha llevado al surgimiento del término *big data* (macrodatos) y a nuevos roles de trabajo, como *científico de datos*, *ingeniero de datos* o *analista de datos*. El análisis de grandes volúmenes de datos permite descubrir patrones, tendencias y correlaciones que pueden aprovecharse en diferentes campos, como la ciencia, la investigación, los negocios, la medicina y muchas otras áreas.

Sin embargo, es crucial tener en cuenta que los datos por sí solos no son infalibles ni objetivos. La calidad de los datos, así como su veracidad, precisión y relevancia, son aspectos fundamentales para su correcta interpretación y uso. Además, la privacidad y la seguridad de los datos también deben considerarse para proteger la información sensible de las personas.

El registro de datos en la historia se remonta a tiempos antiguos y va ligado a la evolución de la escritura. A medida que la sociedad humana evolucionó, se desarrollaron diferentes métodos para registrar información y datos relevantes. A continuación se definen algunos hitos importantes en la historia del registro de datos:

- **Tablillas cuneiformes** (c. 3400 a. C.). Las tablillas de arcilla cuneiformes, utilizadas por las antiguas civilizaciones mesopotámicas, son consideradas uno de los primeros sistemas de registro de datos. Estas tablillas contenían inscripciones en forma de cuñas realizadas con un estilete.
- **Papiro y pergaminos** (c. 3000 a. C.). El uso del papiro, una planta que se procesaba para obtener una superficie de escritura, y los pergaminos, hechos de piel de animales, permitieron la escritura y el registro de información en el antiguo Egipto y otras culturas mediterráneas.
- **Ábacos y quipus** (varios períodos históricos). Los ábacos, dispositivos mecánicos utilizados para realizar cálculos y contar, se utilizaron en diferentes culturas, como en la antigua Mesopotamia, China, Grecia y Roma. Por otro lado, los quipus eran cuerdas y nudos utilizados por los incas para el registro y la contabilidad.
- **Invención de la imprenta** (1440). La invención de la imprenta por Johannes Gutenberg marcó un hito importante en la historia de la difusión y el registro de información. La posibilidad de imprimir textos en masa facilitó la distribución de datos y conocimientos en Europa y más allá.
- **Revolución digital** (siglo XX). Con el advenimiento de la computadora y la tecnología digital, se produjo una transformación significativa en la forma en que se registraban, almacenaban y procesaban los datos. El desarrollo de bases de datos, software especializado y herramientas de análisis ha permitido una gestión más eficiente y avanzada de los datos en diversas industrias y disciplinas.

Figura 2. Evolución del registro de datos

Fuente: Wikipedia

Estos son solo algunos ejemplos destacados en la historia del registro de datos. A medida que la sociedad ha avanzado, se han desarrollado una variedad de métodos y tecnologías para el registro y el tratamiento de datos, que continúan evolucionando hasta la actualidad.

El concepto de ***dato*** es fundamental en el campo de las tecnologías de la información y ha estado presente desde los primeros días de esta disciplina. A medida que estas tecnologías han ido evolucionando, ha sido necesario almacenar y procesar información de manera estructurada, lo que llevó a la aparición del concepto de *dato*. Sin embargo, es importante destacar que el uso y la comprensión de los datos han evolucionado significativamente con el tiempo.

Históricamente, el concepto de dato en las tecnologías de la información se ha asociado con la representación de información en forma de valores numéricos, alfanuméricos o binarios que pueden manipularse y procesarse mediante algoritmos y programas de software. Los datos son la materia prima con la que se trabaja en sistemas informáticos y se utilizan para almacenar información, realizar cálculos, tomar decisiones y brindar resultados.

A medida que estas tecnologías han ido avanzando, se han ido desarrollando diferentes estructuras y modelos de datos para organizar y representar la información de manera más eficiente, como bases de datos relacionales, estructuras de datos jerárquicas y modelos de datos orientados a objetos. Además, con la llegada de internet, los dispositivos móviles, las redes sociales y, en consecuencia, la explosión de datos digitales, el concepto de datos ha adquirido una importancia aún mayor.

Objetivos

Mediante el estudio de este módulo, el estudiantado debe ser capaz de:

1. Entender qué es el dato y su importancia.
2. Conocer cómo pueden estructurarse los datos.
3. Conocer diversas formas de clasificar los datos.
4. Conocer el ciclo de vida del dato, desde su generación hasta su historificación o borrado.
5. Entender el concepto de sistema de base de datos, sus características y sus funcionalidades.
6. Entender la importancia de la analítica de datos y conocer diferentes métodos y aplicaciones.

1. Tipos de datos, estructura y representación

Para obtener, categorizar y analizar debidamente los datos en un contexto determinado, es fundamental conocer los diferentes tipos de datos que pueden existir así como su estructura. De esta forma, se puede trabajar en un modelo que permita explorar e interpretar estos datos de una forma óptima.

Los tipos de datos pueden clasificarse según su estructura en tres categorías principales: estructurados, semiestructurados y no estructurados. A continuación, se define cada categoría:

- **Datos estructurados**. Los datos estructurados están organizados en una estructura predefinida y siguen un formato consistente. Suelen almacenarse en bases de datos relacionales y se representan mediante tablas con filas y columnas (similar a una hoja de cálculo de Excel). Cada columna tiene un tipo de datos específico y se espera que los valores se ajusten a esa estructura. Los datos estructurados se pueden consultar, analizar y procesar fácilmente utilizando consultas y algoritmos diseñados para ese formato.

  Por ejemplo, una tabla de clientes suscritos a un servicio con columnas como ID de cliente, nombre, apellidos, correo electrónico, teléfono, fecha de alta e importe de suscripción.

Tabla 1. Clientes suscritos a un servicio

id_client	firstname	lastname	email	phone	subscription_date	amount
00001	Luisa	Martin	lmartin@correo.com	343555222	5/2/2023	30
00002	Alberto	Díaz	adiaz@correo.com	442123444	6/4/2023	50

En este ejemplo, cada fila de la tabla representa un cliente individual. Los datos están organizados en columnas (ID de cliente, nombre, apellidos, etc.). Cada columna tiene un tipo de dato específico, como cadenas de caracteres (nombre, apellidos y correo electrónico), fechas (fecha de compra) y valor numérico (importe suscripción). Estos datos son fáciles de organizar, clasificar y analizar utilizando bases de datos relacionales, herramientas de generación de informes, etc.

- **Datos semiestructurados**. Los datos semiestructurados tienen cierta estructura, pero no siguen un esquema estricto como los datos estructurados. Pueden estar organizados en formatos como JSON (JavaScript Object Notation, en castellano, 'notación de objetos JavaScript') o (eXtensible Markup Language, en castellano, 'lenguaje de marcado extensible'), donde se utiliza una jerarquía y etiquetas para representar la información. Aunque hay cierta estructura, los valores pueden ser opcionales o pueden variar entre diferentes elementos o documentos.

  Siguiendo el ejemplo anterior, se muestra un archivo JSON que contiene los datos de un cliente suscrito:

```
{
	"clientes": [
		{
		"Id_cliente": "00001",
		"nombre": "Luisa",
		"apellidos": "Martín",
		"email": "lmartin@correo.com",
		"telefono": "343555222",
		"fecha_alta": "5/2/2023",
		"importe_subscripcion": "30"
		},
		{
		"Id_cliente": "00002",
		"nombre": "Alberto",
		"apellidos": "Díaz",
		"email": "adiaz@correo.com",
		"fecha_alta": "6/4/2023",
		"importe_subscripcion": "50"
		}
	]
}
```

  En este ejemplo, se utiliza una estructura de datos llamada *clientes* que contiene una colección de objetos individuales representando a cada cliente. Cada objeto tiene propiedades como ID de cliente, nombre, apellidos, teléfono, etc. Según el cliente, pueden almacenarse más o menos propiedades.

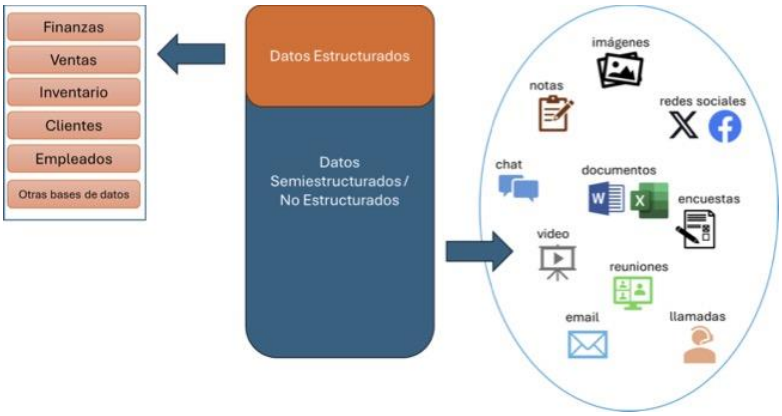
  El formato JSON, aunque no tiene una estructura rígida como una tabla de base de datos, permite representar datos semiestructurados de manera legible y fácilmente manipulable, puesto que los valores van acompañados de su correspondiente tipo de atributo.

- **Datos no estructurados**. Los datos no estructurados no tienen una estructura predefinida y no siguen un formato específico. Pueden ser datos en forma de texto libre, imágenes, vídeos, grabaciones de voz, correos electrónicos o publicaciones de redes sociales, entre otros. Estos datos no están organizados

en una forma tabular o jerárquica y no se puede realizar una clasificación fácilmente. Requieren técnicas de procesamiento especializadas, como el procesamiento del lenguaje natural o el análisis de imágenes, para extraer información significativa.

Se estima[1] que alrededor del 20 % de los datos que genera y manipula una empresa son estructurados, mientras que el 80 % de los datos restantes son no estructurados.

Figura 3. Ejemplo de relación de datos estructurados y no estructurados generados por una organización empresarial



Otra posible clasificación de los datos puede ser según su naturaleza. Para ello, se definen dos tipos:

- **Datos cualitativos**. Se refieren a características o cualidades observables o deducibles que no pueden medirse numéricamente. Estos datos describen atributos o cualidades como estados emocionales, preferencias u opiniones, entre otros. Los datos cualitativos suelen expresarse en forma de palabras, descripciones o categorías.

  Por ejemplo, el análisis de sentimiento con base en una encuesta o preguntas a clientes sobre un producto o servicio (por ejemplo: «satisfecho», «decepcionado», «neutral», etc.).

[1] Estimación de Gartner. Véase el informe en el siguiente enlace (requiere suscripción): https://www.gartner.com/document/3077117?ref=solrAll&refval=184973918&qid=7ac0f5706d57ff6d65c141149d8da3e4

Los datos no estructurados son en su mayoría cualitativos, ya que proporcionan información que no sigue un patrón predecible y que requiere ciertas técnicas para comprender su contexto y estructura, y obtener conclusiones.

- **Datos cuantitativos**. Estos datos se pueden medir y expresar en forma de números. Representan cantidades numéricas y pueden analizarse utilizando técnicas estadísticas. Los datos cuantitativos pueden ser discretos (pueden tomar valores dentro de un conjunto numerable) o continuos (pueden expresar una cantidad infinita de valores).

  Un ejemplo de dato discreto es la cantidad de existencias disponibles en un almacén o el total de pedidos de un cliente, mientras que un dato continuo puede ser el registro diario de temperatura en una determinada ubicación.

Los datos cualitativos suelen ser, por naturaleza, datos no estructurados, mientras que los datos cuantitativos suelen ser datos estructurados.

Adicionalmente a estas dos clasificaciones, existen otros tipos de clasificación que pueden darse según el contexto y la disciplina en la que se esté trabajando. A continuación, se indican algunos ejemplos:

- **Datos sensibles y no sensibles**. Los datos sensibles son los que pueden requerir protección especial debido a su naturaleza confidencial o su potencial para causar daño a una persona u organización si se divulgan o se utilizan de manera inapropiada. Ejemplos de datos sensibles incluyen información médica, datos financieros u orientación sexual, entre otros.

- **Datos personales y no personales**. Los datos personales se refieren a la información que identifica o puede identificar a una persona específica, como el nombre, la dirección, el número de identificación, etc. Los datos no personales son los que no están directamente relacionados con una persona identificable (como, por ejemplo, la fecha de alta de un cliente).

- **Datos históricos y en tiempo real**. Los datos históricos son los que se han recopilado y almacenado en un momento anterior, y proporcionan información sobre eventos pasados. Los datos en tiempo real son los que se generan y transmiten instantáneamente a medida que ocurren, y permiten la toma de decisiones en tiempo real.

- **Datos transaccionales y no transaccionales**. Los datos transaccionales (también conocidos como datos operativos) son los referentes a transacciones y operaciones comerciales, como compras, ventas, reservas o suscripciones, entre otros. Los datos no transaccionales son los que representan información que no está directamente relacionada con las transacciones a tiempo real, pero

que proporcionan información de referencia o analítica, como datos descriptivos de un cliente o ventas totales del último periodo fiscal.

Estas son solo algunas de las clasificaciones comunes de los tipos de datos. Hay que tener en cuenta también que, en muchos casos, los datos pueden ser mixtos, es decir, pueden combinar diferentes tipos de datos en una misma representación o conjunto.

2. Ciclo de vida del dato: del dato al conocimiento

Un determinado dato puede pasar por diferentes etapas desde su creación o captura hasta su obsolescencia. Aunque las etapas específicas pueden variar según el contexto, generalmente se definen las siguientes fases en el ciclo de vida del dato:

1) **Creación**. En esta etapa, los datos se generan o recopilan por primera vez. Pueden provenir de diversas fuentes, como transacciones, sensores, encuestas, formularios, registros, etc.

2) **Captura y almacenamiento**. Los datos se capturan y se almacenan en sistemas adecuados, como bases de datos relacionales, almacenes de datos, *data lakes* u otras plataformas de almacenamiento. Aquí se realiza el proceso de guardar los datos en un formato y estructura apropiados.

3) **Organización y procesamiento**. En esta fase, los datos se organizan, se limpian, se transforman y se procesan para su uso posterior. Se aplican técnicas de limpieza de datos, normalización, agregación y otras operaciones para asegurar su calidad y coherencia.

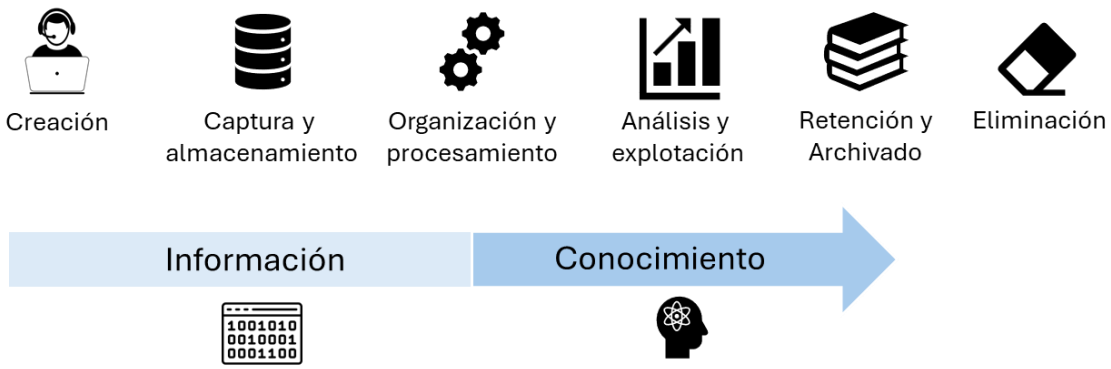
4) **Análisis y explotación**. En esta etapa, los datos se ponen a disposición de los usuarios y sistemas que los necesiten a través de informes, cuadros de mando, interfaces de programación de aplicaciones (API) y exportaciones de datos, entre otros métodos. Una vez disponibles, se utilizan diferentes técnicas para extraer información y conocimientos relevantes o descubrir patrones, tendencias y relaciones, mediante el uso de consultas directas, minería de datos, aprendizaje automático, estadísticas y visualización de datos, entre otros.

5) **Retención y archivado**. En esta fase, los datos se retienen y archivan de acuerdo con las políticas de retención de la organización y los requisitos legales. Algunos datos pueden almacenarse a largo plazo para cumplir con regulaciones o para futuros análisis históricos.

6) **Eliminación**. Los datos que ya no son necesarios o relevantes se eliminan de forma segura según las políticas de retención y privacidad de la organización.

Esto implica la destrucción o el borrado seguro de los datos para proteger la privacidad y cumplir con las regulaciones.

Figura 4. Etapas del ciclo de vida del dato



Es importante destacar que el ciclo de vida del dato no es un proceso lineal y puede implicar iteraciones, retroalimentación y cambios en cada etapa a medida que se obtienen nuevos conocimientos y se ajustan las necesidades y requisitos de la organización.

3. Sistemas de almacenamiento y análisis de datos

Un sistema de almacenamiento de datos es una infraestructura tecnológica diseñada para gestionar y almacenar datos de manera eficiente y organizada. Son sistemas generalmente optimizados para manejar grandes volúmenes de datos. Estos sistemas son esenciales para las organizaciones que necesitan gestionar datos en una variedad de tipos y escalas.

Los sistemas de almacenamiento de datos incorporan, en mayor o menor medida, las siguientes características:

- **Almacenamiento eficiente**. Estos sistemas están diseñados para tratar volúmenes masivos de datos, ya sean datos estructurados, semiestructurados o no estructurados.
- **Organización de datos**. Ayudan a organizar los datos de manera coherente, para facilitar su consulta y acceso.
- **Acceso y manipulación de datos**. Incorporan herramientas o proveen de servicios para realizar consultas sobre los datos y manipularlos mediante operaciones de inserción, modificación y borrado. Asimismo, incorporan lenguajes de programación los cuales permiten ejecutar comandos de consulta, manipulación y análisis de datos.
- **Velocidad y rendimiento**. Los sistemas de almacenamiento de datos deben permitir un acceso rápido y eficiente a los datos, especialmente en aplicaciones y servicios donde la velocidad es crucial.
- **Escalabilidad**. Deben ser capaces de almacenar grandes volúmenes de datos y soportar operaciones masivas sin afectar a su rendimiento.
- **Seguridad**. Deben proporcionar mecanismos de seguridad y encriptación para proteger los datos y permitir que solo los usuarios o servicios con los permisos necesarios puedan acceder a ellos.
- **Disponibilidad**. Deben garantizar la disponibilidad en caso de fallos para no afectar a las herramientas o servicios que dependen de los datos que están almacenados.
- **Recuperación y copias de seguridad**. Deben permitir recuperar de datos en caso de pérdida y crear copias de seguridad regulares para garantizar la integridad de los datos.

- **Integración**. Deben ser capaces de integrarse con otras herramientas y sistemas utilizados por la organización.

Los sistemas de almacenamiento de datos pueden variar en términos de tecnología y diseño. Pueden incluir sistemas de bases de datos tradicionales, sistemas de archivos distribuidos, almacenes de datos (*data warehouses*), sistemas de almacenamiento en la nube y más. Cada tipo de sistema tiene sus propias ventajas y desafíos, y la elección de uno u otro depende de las necesidades específicas de la organización y el tipo de datos que se manejan.

Estos sistemas también suelen incorporar funcionalidades o herramientas (creadas por el propio fabricante o por terceros) para proporcionar capacidades de **análisis de datos**. A través de estas capacidades, es posible:

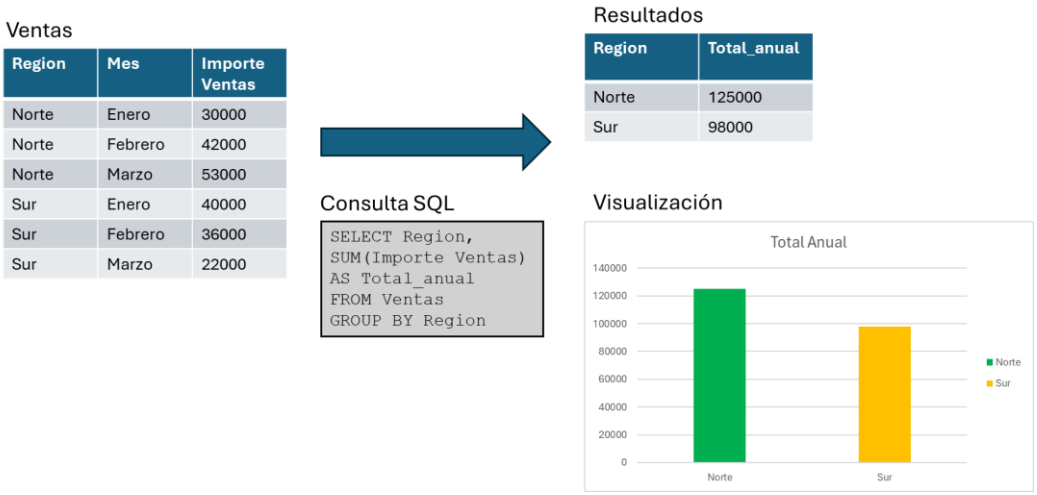
- **Crear y procesar consultas analíticas**. Por ejemplo, la obtención de medidas como el total o la suma a partir de una agrupación o filtrado de datos. Los sistemas también permiten realizar consultas más complejas, como análisis de series temporales, regresiones, clusterización (*clustering*) y otros. Para realizar estas consultas se suele disponer de herramientas y lenguajes de consulta como SQL (Structured Query Language), que veremos más adelante.

- **Ejecutar consultas complejas con eficiencia en el rendimiento**. Estos sistemas incorporan técnicas de almacenamiento y procesamiento que manejan grandes volúmenes de datos y ofrecen un rendimiento rápido en la ejecución de consultas.

- **Integrar datos de múltiples tablas o bases de datos**. Las consultas analíticas incorporan funciones que permiten agrupar y obtener datos relacionados provenientes de diferentes tablas o bases de datos.

- **Visualizar datos y resultados de consultas**. Utilizando diversas herramientas de visualización, los resultados de las consultas pueden representarse gráficamente para proporcionar una comprensión más intuitiva. Estas herramientas ofrecen una amplia variedad de visualizaciones, desde las simples, como una tabla o un gráfico de barras, hasta las más complejas, como cuadros de mando integrales.

Figura 5. Ejemplo sencillo de consulta analítica y visualización de datos



3.1. Base de datos SQL y NoSQL

A la hora de almacenar datos, los sistemas de almacenamiento digital pueden clasificarse en dos categorías principales: **SQL** (relacionales) y **NoSQL** (no relacionales).

El término *SQL* es un acrónimo de ***Structured Query Language***, el cual es un lenguaje de programación que permite consultar, manipular y cambiar datos en una base de datos relacional. Este término se popularizó en la década de 1970 con el desarrollo y la adopción de sistemas de gestión de bases de datos relacionales (RDBMS, por sus siglas en inglés). Aunque el lenguaje y los conceptos relacionados con SQL comenzaron a desarrollarse en la década de 1970, no fue hasta la década de 1980 que se convirtió en un estándar ampliamente reconocido y utilizado en la industria de las bases de datos. Su uso perdura hasta el día de hoy y está estrechamente ligado al concepto de bases de datos relacionales.

El término ***NoSQL***, que significa ***Not Only SQL*** o ***No Solo SQL***, hace referencia a bases de datos que proporcionan un mecanismo de almacenamiento y consulta diferente a las tablas y relaciones utilizadas en las bases de datos relacionales. Aunque estas bases de datos existen desde la década de 1960, el término *NoSQL* no se acuñó hasta principios de la década de 2000, a medida que las organizaciones se enfrentaban a desafíos relacionados con el manejo de grandes volúmenes de datos no estructurados, como los generados por redes sociales, aplicaciones web y dispositivos conectados. Los sistemas NoSQL surgieron para abordar problemas de escalabilidad, flexibilidad y rendimiento que las bases de datos relacionales tradicionales no podían resolver de manera eficiente.

Las diferencias entre las bases de datos SQL y NoSQL radican en su estructura, modelo de datos, escalabilidad y casos de uso. A continuación, se realiza una comparación de las principales diferencias entre ambos tipos de bases de datos:

- **Modelo de datos**. Las bases de datos SQL utilizan un modelo de datos tabular basado en tablas con filas y columnas, estando sus tablas relacionadas entre sí. En comparación con las SQL, las bases de datos NoSQL utilizan diversos modelos de datos, como documentos, gráficos, clave-valor y columnas. No se adhieren a un esquema fijo y pueden adaptarse a estructuras más flexibles.

- **Estructura de datos**. Las bases de datos SQL requieren de un esquema predefinido y rígido que debe respetarse a la hora de introducir y manipular datos. Dependiendo de la complejidad de la base de datos, los cambios en el esquema pueden ser complicados de abordar. Sin embargo, las bases de datos NoSQL no requieren de un esquema fijo, lo que les permite agregar, modificar o eliminar campos fácilmente, sin interrumpir la funcionalidad.

- **Escalabilidad**. Las bases de datos SQL escalan verticalmente, normalmente en un mismo servidor, y pueden requerir un aumento de hardware para añadir más capacidad de almacenamiento o memoria. Esto puede llegar a ser bastante costoso cuando se manejan grandes volúmenes de datos. En contraposición, las bases de datos NoSQL pueden escalar horizontalmente, de modo que se pueden añadir más servidores para distribuir los datos de forma más eficiente, lo que las hace más apropiadas para infraestructuras basadas en la nube.

- **Consultas**. Las bases de datos SQL están optimizadas para consultas estructuradas y complejas, utilizando el lenguaje SQL. En cambio, las bases de datos NoSQL utilizan una variedad de lenguajes y formatos para interactuar con los datos, por ejemplo lenguajes de consulta basados en JSON, JavaScript o líneas de comandos, y pueden incluir también el propio SQL o variaciones de este. Esta flexibilidad facilita el manejo de datos no estructurados.

- **Casos de uso**. Las bases de datos SQL son adecuadas para aplicaciones que necesitan integridad de datos y relaciones complejas, como sistemas de gestión de relaciones con el cliente (CRM) o sistemas de gestión de contenidos (CMS). Las bases de datos NoSQL son ideales para aplicaciones que gestionan grandes volúmenes de datos no estructurados, como redes sociales, aplicaciones móviles y aplicaciones de internet de las cosas (IoT).

A grandes rasgos, las principales diferencias entre bases de datos SQL y NoSQL pueden sintetizarse tal como muestra en la tabla 2.

Tabla 2. Tabla resumen de las diferencias entre sistemas SQL y NoSQL

	Bases de datos SQL	Bases de datos NoSQL
Modelo de datos	Utiliza un único modelo de datos, el tabular.	Utiliza diversidad de modelos de datos.
Estructura de datos	Con un esquema predefinido.	Sin un esquema predefinido.
Escalabilidad	Vertical.	Horizontal.
Consultas	Lenguaje SQL.	Variedad de lenguajes.
Casos de uso	Aplicaciones que necesitan integridad de datos y relaciones complejas.	Aplicaciones que gestionan grandes volúmenes de datos no estructurados.

En el momento de preparar este módulo, algunos ejemplos de bases de datos SQL y NOSQL del mercado eran los siguientes:

Figura 6. Ejemplos de bases de datos SQL y No SQL existentes en el mercado

3.2. Almacenes de datos (*data warehouses*)

A medida que las organizaciones comenzaron a almacenar grandes volúmenes de datos provenientes de diversas aplicaciones y fuentes, surgió la necesidad de contar con un repositorio central donde pudieran almacenarse en una estructura consolidada, de modo que diferentes áreas de negocio pudieran consultarlos y analizarlos.

Para abordar estas necesidades se introdujo el concepto de **almacén de datos** o *data warehouse*. Este término se introdujo en la década de 1980 mediante un artículo publicado en 1988 por los investigadores de IBM, Barry Devlin y Paul Murphy. Posteriormente, autores como Bill Inmon y Ralph Kimball ayudaron a popularizar este concepto desde enfoques ligeramente diferentes.

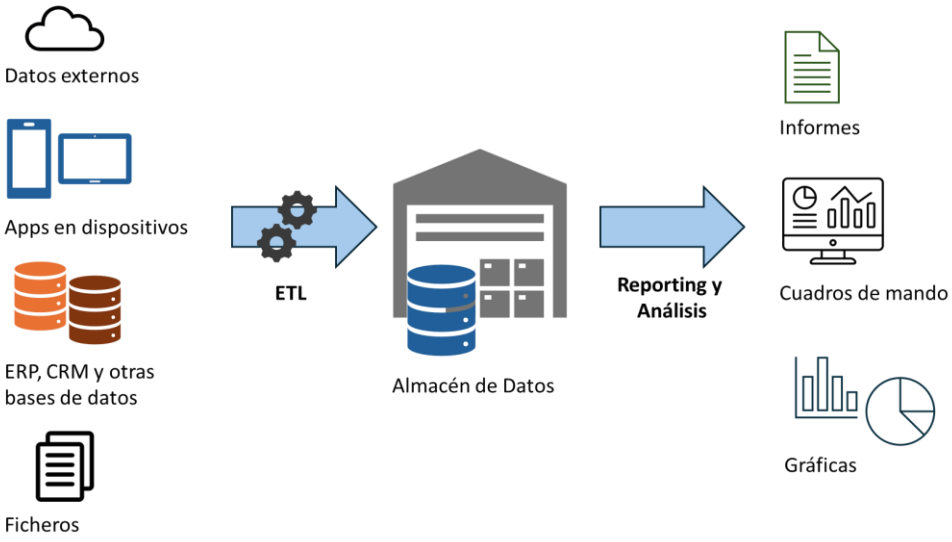
En un almacén de datos, los datos se almacenan en un formato estructurado, generalmente en una base de datos relacional. Estos datos se obtienen de diversas fuentes, como sistemas transaccionales, ficheros de datos maestros, archivos y bases de datos externas. Mediante una serie de procesos, estos datos se transforman y se cargan en el almacén de datos para su análisis. Estos procesos son conocidos como **ETL** (*extract, transform and load*).

Un almacén de datos ofrece los siguientes beneficios respecto a las bases de datos:

- **Facilita el acceso a los datos para su análisis**, ya que proporciona un único punto de acceso a los usuarios de negocio que necesitan consultar y analizar la información.
- **Mejora la toma de decisiones** al consolidar los datos de la organización en un único lugar, de modo que facilita la obtención de información y su análisis.
- **Optimiza el rendimiento**, ya que los almacenes de datos están diseñados para el análisis y la generación de informes, e incorpora relaciones y medidas útiles para la organización.
- **Mejora la calidad de los datos** al ser sometidos a limpieza y transformación antes de cargarlos en el almacén de datos, lo cual es esencial para evitar decisiones incorrectas basadas en datos erróneos.
- **Fortalece la gobernanza de datos** al centralizar la información, de modo que simplifica la gestión del acceso de usuarios y asegura la corrección y actualización de los datos.
- **Reduce costes** al eliminar la necesidad de múltiples fuentes de datos para análisis y generación de informes, cada una con sus propias complejidades para acceder a la información.

- **Proporciona escalabilidad** al estar diseñado para satisfacer las necesidades de crecimiento de las organizaciones.

Figura 7. Ejemplo de componentes de un almacén de datos



El almacén de datos se describe en detalle en el apartado 1, «*Data warehouse*», del módulo «*Sistemas de bases de datos analíticas*» del curso.

3.3. Persistencia de los datos

En el ámbito de las tecnologías de la información, el concepto de **persistencia** hace referencia al estado de un sistema que perdura más allá del proceso que lo creó o actualizó. Esto se logra guardando dicho estado en un sistema de almacenamiento. La **persistencia de datos** se refiere a la capacidad de los datos de mantenerse en la base de datos que los almacena en un estado estable y accesible.

Por lo general, los sistemas transaccionales (como el sistema de facturación de una empresa o el sistema de gestión de incidencias de un centro de atención telefónica) almacenan los datos generados durante un periodo determinado. A medida que estos datos crecen, surgen desafíos relacionados con el coste de almacenamiento, la administración y el rendimiento, entre otros. Un enfoque para abordar estos problemas es almacenar datos con una determinada antigüedad en un almacenamiento denominado ***histórico*** y eliminar estos datos del almacenamiento original. A este proceso se le conoce como ***historificación***.

Es posible utilizar el almacén de datos como soporte para el histórico de datos; no obstante, es importante destacar que la estructura de los datos en un almacén de datos

no necesariamente coincide con la de los sistemas de origen. Esto se debe a que la estructura del almacén de datos está optimizada para el análisis y la generación de informes. Esta diferencia en la estructura podría dificultar el proceso de restauración de datos históricos en el sistema de origen, en caso de ser necesario.

Asimismo, la información almacenada en un almacén de datos también puede ser historificada con el objetivo de mejorar el rendimiento y proporcionar solamente los datos necesarios para el análisis y la toma de decisiones.

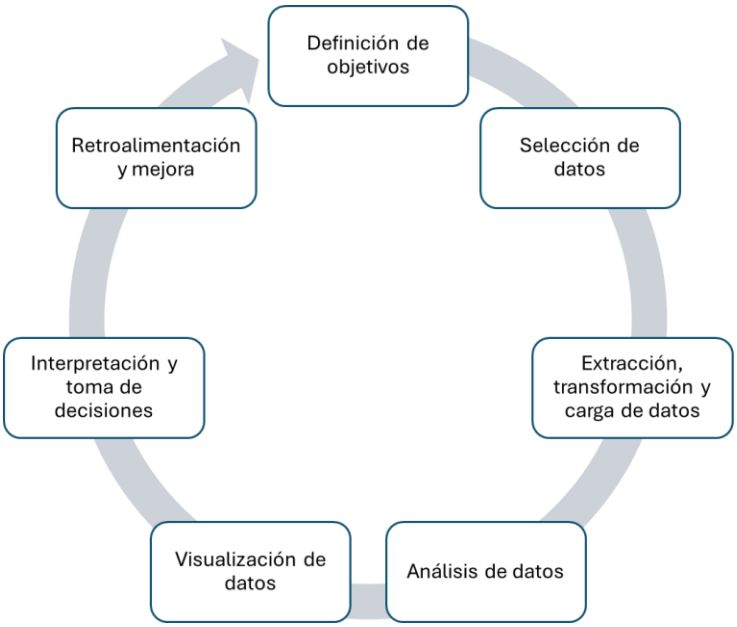
En conclusión, la persistencia de datos en un sistema transaccional tiende a ser más breve que en un almacén de datos. Esto se debe a que el primero se utiliza para operaciones que requieren un rendimiento óptimo, mientras que el segundo se emplea para el análisis y la generación de informes, donde la calidad de la información almacenada es más relevante que el rendimiento en las operaciones de inserción y actualización de datos.

3.4. El proceso de análisis de datos

Los analistas y usuarios de negocios que trabajan con datos suelen atravesar varias etapas o fases, desde que definen sus objetivos de análisis hasta que finalmente toman decisiones basadas en la visualización y el análisis de los datos.

La figura 8 ilustra este proceso. Es importante destacar que en dicho proceso participan diversos actores, desde analistas y expertos en gestión y manipulación de datos hasta procesos automatizados que ejecutan diferentes pasos.

Figura 8. Fases de un proceso de análisis de datos



A continuación, se explican las diferentes fases:

1) **Definición de objetivos**. El primer paso consiste en establecer el objetivo del análisis que se necesita realizar. El analista deberá responder a preguntas como las siguientes:

   - ¿Qué datos y cálculos necesito visualizar?
   - ¿Cuál es el mejor formato de visualización que me ayude a analizar estos datos?
   - ¿Qué patrón necesito identificar en los datos visualizados?

2) **Selección de datos**. En esta fase se identifica los datos que son relevantes para el análisis. Por ejemplo, en una base de datos relacional, esto implica elegir las tablas y columnas de la base de datos que se necesitan para lograr el objetivo del análisis.

3) **Extracción, transformación y carga de datos (ETL)**. Este proceso implica extraer los datos necesarios de diversas fuentes, aplicar transformaciones para prepararlos para el análisis y luego cargarlos en una base de datos destinada a dicho análisis, como un almacén de datos. Las transformaciones pueden abarcar desde la limpieza de datos hasta la combinación de diversas fuentes, el cálculo de nuevas métricas y el ajuste de la estructura para que sea adecuada para el análisis. Normalmente, este proceso está automatizado y se ejecuta de manera recurrente, lo que permite que los analistas cuenten con la información necesaria al realizar el análisis. Sin embargo, en casos de datos recién adquiridos, es esencial llevar a cabo este proceso antes de iniciar el análisis.

4) **Análisis de datos**. En esta fase se definen las técnicas de análisis adecuadas pudiendo crear modelos si es necesario. Estos modelos pueden ser predictivos, clasificatorios, de agrupamiento y de regresión, entre otros, según el problema en cuestión.

5) **Visualización de datos**. Esta fase implica la representación gráfica de los datos analizados tanto para comprender de forma intuitiva los datos obtenidos como para detectar información más compleja, como relaciones, tendencias y patrones, que no son fácilmente deducibles a partir de la visualización de los datos en formato tabular. Estas visualizaciones, como gráficos, mapas, diagramas y tablas, permiten una comprensión más rápida y profunda de la información. Esta fase es esencial cuando deben presentarse los resultados de manera clara y convincente a las partes interesadas.

6) **Interpretación y toma de decisiones**. Una vez visualizados, los resultados obtenidos del análisis se interpretan para responder a las preguntas planteadas y alcanzar los objetivos del análisis. Esto implica extraer conclusiones

significativas y tomar decisiones informadas basadas en la información derivada. Estas decisiones pueden ser estratégicas, operativas o tácticas, según el contexto.

7) **Retroalimentación y mejora**. Después de tomar decisiones basadas en el análisis, es importante evaluar la efectividad de esas decisiones y retroalimentar los resultados al proceso de análisis para mejoras futuras.

Como se puede observar, el análisis de datos es un proceso iterativo, y las fases mencionadas no siempre ocurren de manera lineal ni en un solo ciclo. A menudo, se requiere volver atrás, refinar pasos anteriores y ajustar en función de los resultados y las nuevas preguntas o retos que puedan surgir.

Dentro del gran ciclo de vida del dato, intervienen distintos perfiles de usuario. Pueden destacarse dos grandes grupos, que se corresponden con ambos extremos del ciclo. Por un lado, están los **usuarios de las aplicaciones operacionales**, las aplicaciones ubicadas al inicio del ciclo de vida, que dan soporte a los procesos de negocio, y son el origen del dato. Estos usuarios utilizan las aplicaciones para la operativa del negocio u organización y, por tanto, el análisis de datos o de toma de decisiones es para ellos una actividad secundaria o irrelevante. Por otro lado, están los **usuarios de las aplicaciones destinadas al análisis de datos o a la toma de decisiones**, acostumbrados a trabajar con los datos al final del ciclo de vida del dato, una vez cargados y publicados para su consumo.

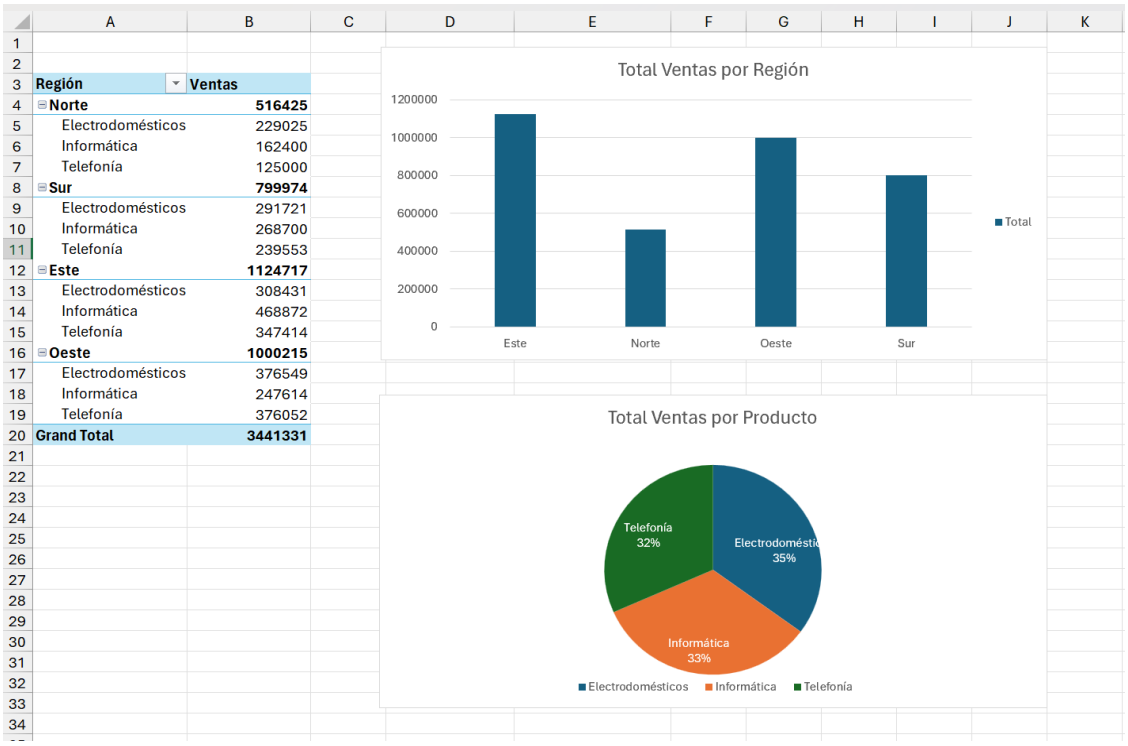
En el módulo «Análisis de datos» del curso ampliaremos detalles sobre este último colectivo.

3.5. Herramientas para el análisis de los datos

Dentro de la fase de visualización de datos, es importante contar con herramientas de visualización que sean capaces de interpretar y mostrar de forma gráfica los resultados. Existe un amplio abanico de herramientas, desde *open source* (código abierto) hasta comerciales. Las organizaciones pueden usar más de una herramienta de visualización en función del tipo de datos y resultados a analizar. A continuación, se muestra una clasificación de los tipos de herramienta más comúnmente utilizados:

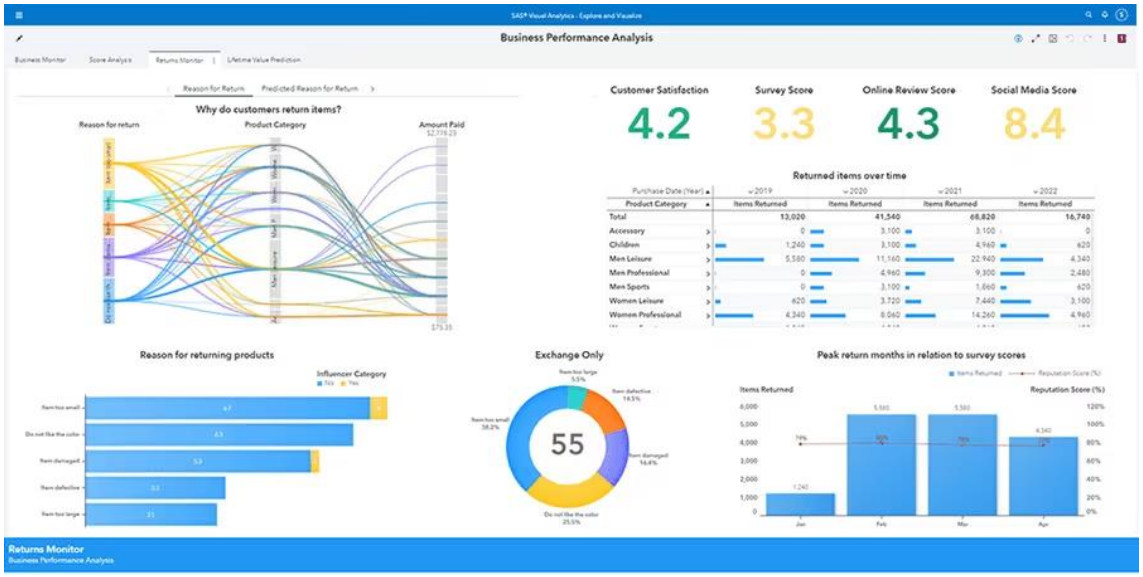
- **Hojas de cálculo**. Las hojas de cálculo son aplicaciones que permiten a los usuarios organizar, almacenar y analizar datos en tablas formadas por filas y columnas. Ofrecen diversas funciones para realizar cálculos matemáticos, estadísticos y financieros, así como para crear gráficos y visualizaciones. En esta categoría se encuentran aplicaciones como Microsoft Excel y Google Sheets.

Figura 9. Ejemplo de informe en Microsoft Excel



- **Software estadístico**. Estas herramientas contienen funcionalidades que permiten realizar análisis estadísticos complejos, como el análisis de varianza, la regresión lineal o las pruebas de hipótesis. En esta categoría existen aplicaciones como SAS (Statistical Analysis System), SPSS (Statistical Package for the Social Sciences) o PSPP (alternativa de código abierto de la herramienta SPSS).

Figura 10. Ejemplo de visualización de datos con SAS

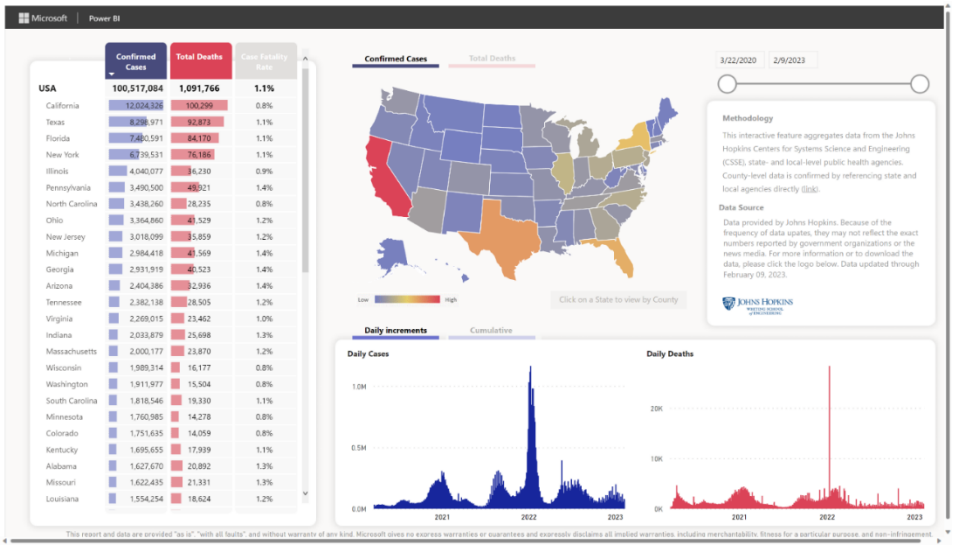


Fuente: página web oficial de SAS

- **Herramientas de inteligencia empresarial (*business intelligence*)**. Las herramientas de inteligencia empresarial permiten conectarse a diferentes orígenes de datos (por ejemplo, hojas de cálculo, bases de datos transaccionales o cubos OLAP) y dar forma al modelo de datos que se utilizará para el análisis. Estas herramientas ofrecen capacidades como el análisis in-memory[2] y permiten crear desde informes sencillos hasta complejos cuadros de mando. En este grupo se encuentran herramientas como Tableau, Microsoft Power BI o QlikView.

[2] Método mediante el cual los datos se almacenan en la memoria RAM para permitir un acceso y análisis más rápidos.

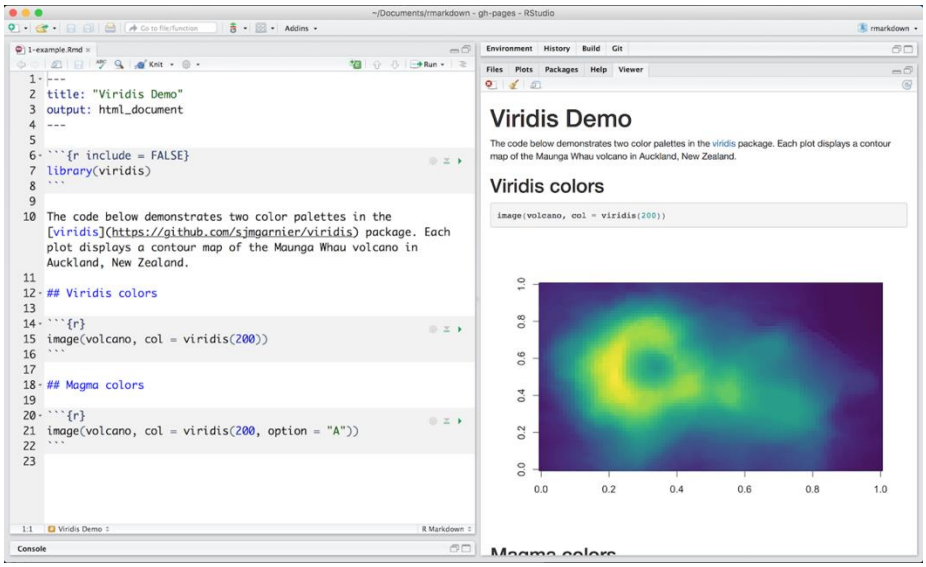
Figura 11. Ejemplo de informe en Power BI



Fuente: learn.microsoft.com

- **Lenguajes de programación**. Además de los propios lenguajes que pueden implementar las herramientas comentadas en los puntos anteriores, existen lenguajes de programación y librerías de lenguajes más genéricos especializados en el análisis estadístico y gráficos. Dentro de esta clasificación se encuentran lenguajes como R o librerías del lenguaje Python como Pandas, NumPy y Matplotlib.

Figura 12. Ejemplo de visualización utilizando el paquete R Markdown de R



Fuente: https://rmarkdown.rstudio.com/

Resumen

En este módulo se aborda de manera exhaustiva el concepto de dato, tanto en su definición general como en su relevancia para el análisis de información.

Inicialmente, se presenta una taxonomía detallada de los datos y se categorizan en estructurados, semiestructurados y no estructurados, cada uno con aplicaciones específicas. Además, se exploran otras dimensiones de clasificación, como datos cualitativos frente a cuantitativos y datos sensibles frente a no sensibles.

Posteriormente, el módulo desglosa las etapas del ciclo de vida de un dato, desde su generación hasta su análisis, archivado y eventual eliminación. Este segmento sienta las bases para entender cómo se manejan los datos a lo largo del tiempo.

La siguiente sección se adentra en los sistemas de almacenamiento y análisis de datos, destacando sus características clave. Se ofrece una comparación entre las bases de datos SQL y NoSQL, una distinción que ha ganado relevancia con el auge de los *big data*. Además, se introduce el concepto de almacén de datos o *data warehouse*, que se examinará con más detalle a lo largo del curso.

Para concluir, el módulo esboza el proceso de análisis de datos y presenta una estructura de las fases que lo componen. Se cierra con una descripción de las herramientas analíticas más comúnmente empleadas en este campo.