
BIG
DATA

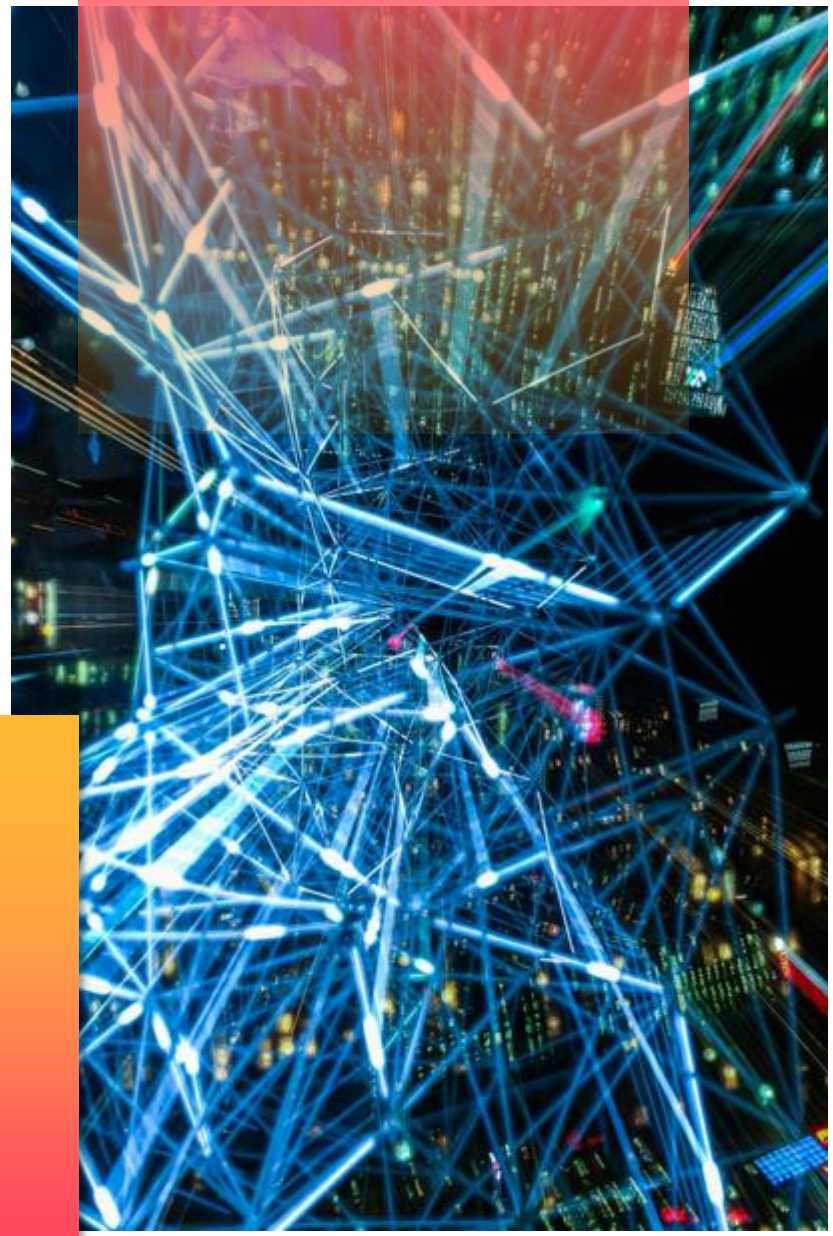
2. La importancia del dato

Introducción

La irrupción de la tecnología digital ha supuesto importantes avances sin precedentes que han transformado nuestro día a día.

Para poder aprovechar al máximo las funcionalidades que nos ofrece la tecnología se ha desarrollado software basado en los datos que recopilan y nos permite explotarlos para hacer crecer nuestras economías.

Los **datos han supuesto una auténtica revolución** a la hora de analizar y transformar la información. Se han convertido en uno de los recursos claves y básicos en nuestro trabajo, es decir, en **materia prima** como ya lo eran la tierra, la mano de obra o el capital.

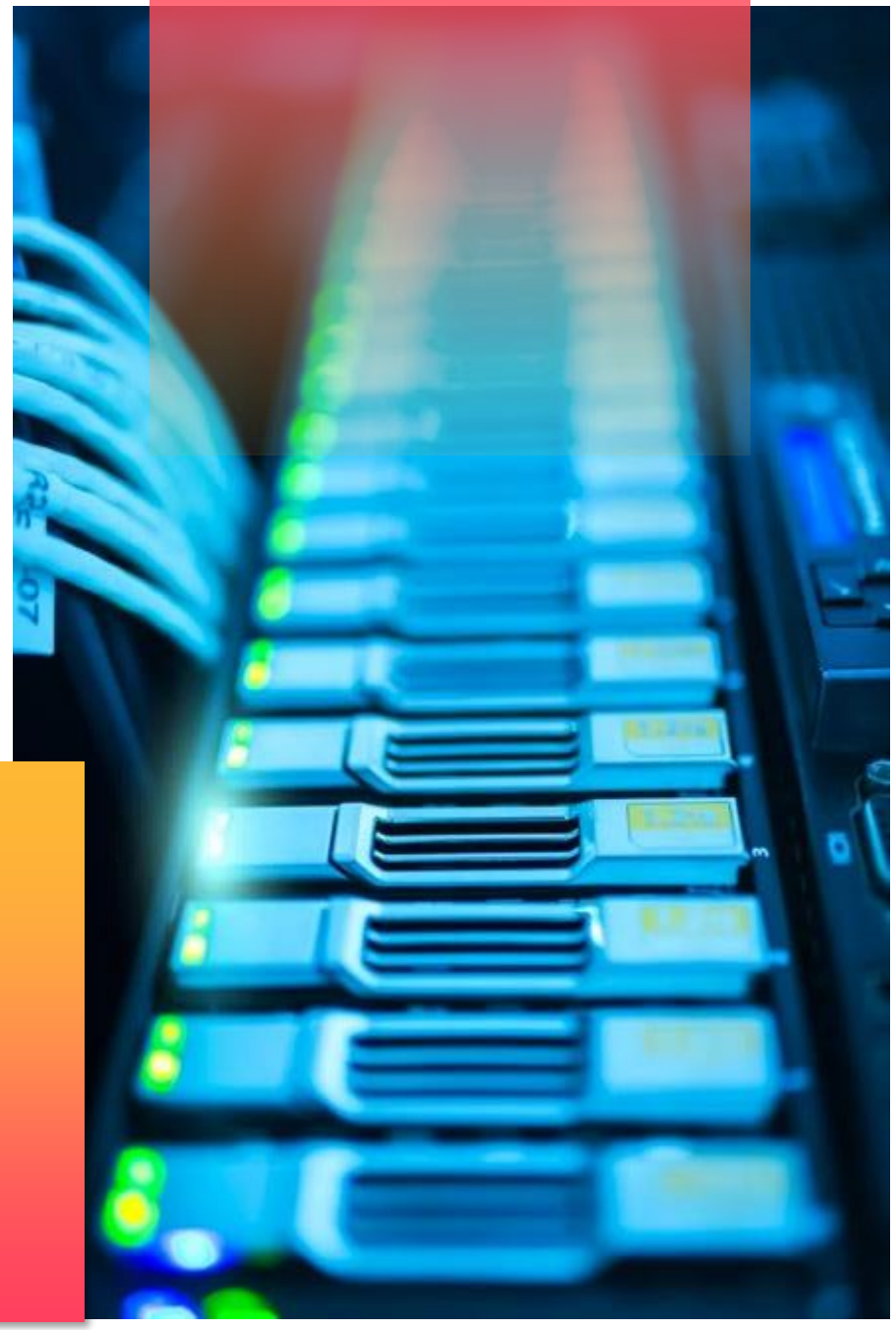


Introducción

Todo este ha sido posible gracias a que la recopilación de datos a pasado de la **observación de patrones** durante cientos de años a un **proceso completamente automatizado** que recopila y analiza los datos a través de una computadora que se comporta como si se tratase de un ser humano.

Los datos son generados por sensores que se encuentran en nuestros móviles, coches, dispositivos, máquinas, farolas… No obstante, en los inicios, ante tal multitud de datos, **era muy difícil y costoso recopilarlos e interpretarlos**. Hoy en día, se ha **democratizado la capacidad de almacenamiento** y se ha **aumentado la capacidad de reutilización** para darles nuevos propósitos con el fin de nuevas formas de conocimiento.

Nuestra economía está centrada en los datos y se calcula que para el 2030 las ganancia del tratamiento eficiente de estos sumen casi 15 billones de dólares al PIB global.



Antecendentes

A lo largo de nuestra historia se han ido recopilando datos, ya sean grabados en una piedra, en un rollo de papiro, en libros…, y, gracias a ellos, ha sido posible colaborar con el avance de la sociedad. No obstante, la cantidad de datos que existía era muy limitada, pero, afortunadamente, los datos han dejado de ser un recurso escaso gracias a nuestra capacidad de conectar varios dispositivos y sensores a Internet que permite generar una cantidad de datos a nivel exponencial. **Se estima que se generan 2,5 cuatrillones de bytes de datos al día.**

Antecedentes

Para verlo de manera más visual piensa que el **90% de datos en todo el mundo se han producido en los últimos dos años.** Un ejemplo de esto lo podemos ver en que los datos comerciales se están **duplicando cada 1,2 años**.

Esta generación de datos exponencial se debe sobretodo a cómo se producen, y es que a medida que crece la cantidad de dispositivos que se conectan a Internet (Internet de la cosas) la cantidad de datos generadas aumenta vertiginosamente.

Disminución del coste de almacenamiento de los datos

La disminución del coste de almacenamiento de los datos ha supuesto la **revolución** de estos. Si echamos la vista atrás, en 1980 el almacenamiento de un gigabyte costaba cientos de miles de dólares, era difícil de conseguir y requería de una persona para su manipulación. Hoy en día apenas cuesta unos céntimos, no se necesita de nadie para poder utilizarlo y podemos comprarlo físico o en la nube.

Esta gran bajada de precio ha hecho posible que los datos supongan **un recurso renovable que se puede combinar con otros** conjuntos de datos y utilizar muchas veces.

Análisis de datos

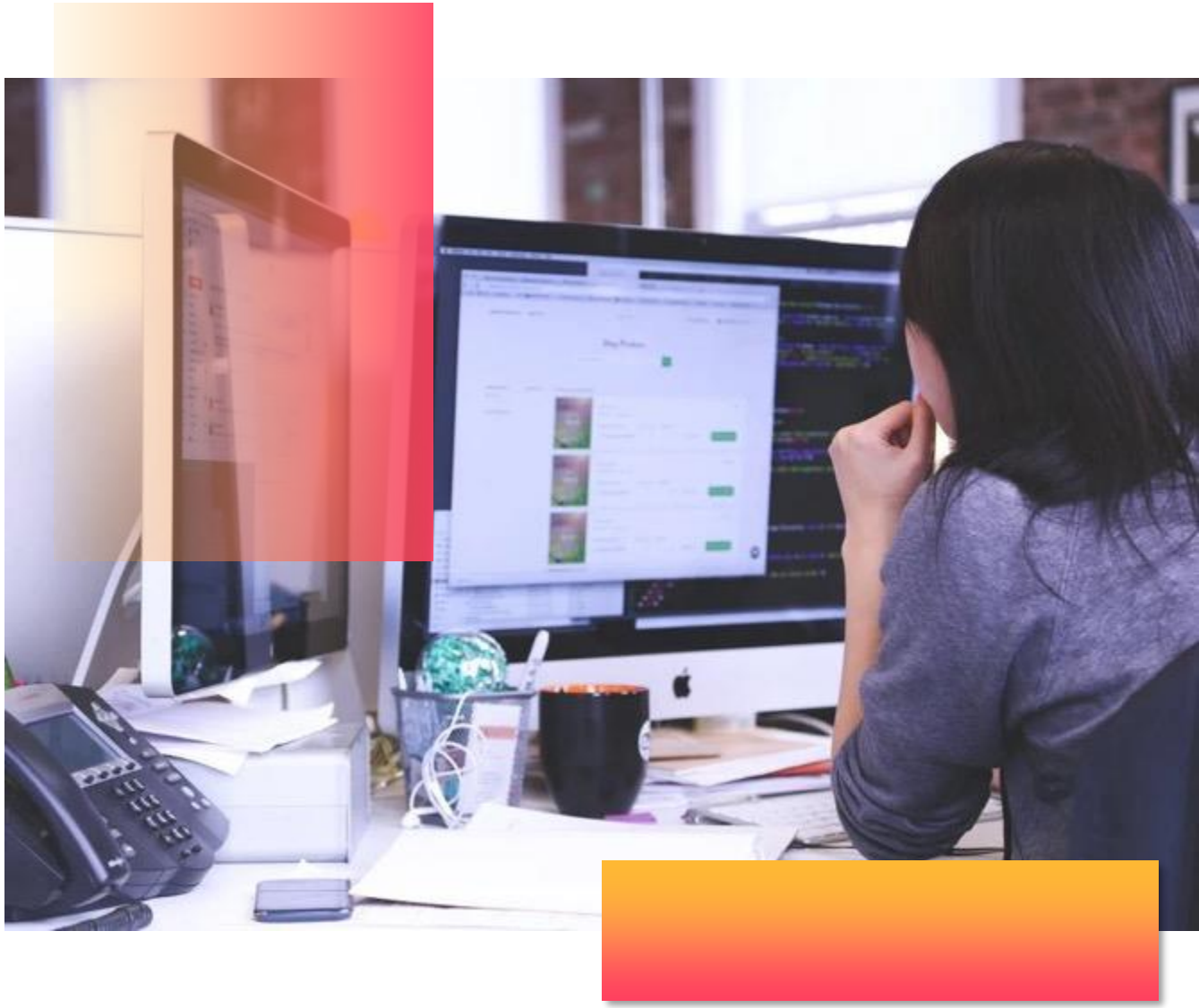
Si los datos no son comprensible no sirven para nada. Para lograr hacerlos comprensibles es necesaria la utilización de un **software** combinado con el ingenio humano.

Las **herramienta analíticas** ordenan el conjunto de datos con el fin de ayudar a descubrir pautas y nuevas ideas y tendencias desde el punto de vista estadístico. Mediante una base de datos y algoritmos estadísticos estas herramientas de software analítico permiten extraer información valiosa en tiempo real entre una gran cantidad de datos que nos ayudarán a tomar decisiones de manera más rápida y con mayor precisión.

Economía del dato

Debido a la aparición del dato se están produciendo **un aumento en la productividad** de las empresas. Pero no solo afecta a las empresas que los implantan, los efectos económicos de los datos se están **expandiendo a muchos sectores**, por ejemplo, en EE.UU. cada puesto de trabajo relacionado con los datos genera otros tres puestos de trabajo de manera indirecta.

Además, no solo implica aumentar el crecimiento económico de una empresa, también se trata de crear un **nuevo motor de creación de empleo**, ya que se están creando miles de nuevos trabajos nuevos y muy bien remunerados.



Economía del dato. ¿Cómo afecta a los diferentes sectores?

Producción

Es el sector que más almacena datos y esto se transforma en una mayor eficiencia, producto de mayor calidad y distribución más efectiva.

Finanzas

Mejora de la eficiencia operativa y disminución de fraudes.

Agricultura

Producción de mejores alimentos utilizando menos recursos.

Salud

Aumento de la esperanza de vida.

Transporte

Ahorra tiempo y dinero a los viajeros. Un transporte más eficiente supone ahorrar combustible y por ende reducir las emisiones de CO2.

Energía

Reducción del consumo energético.

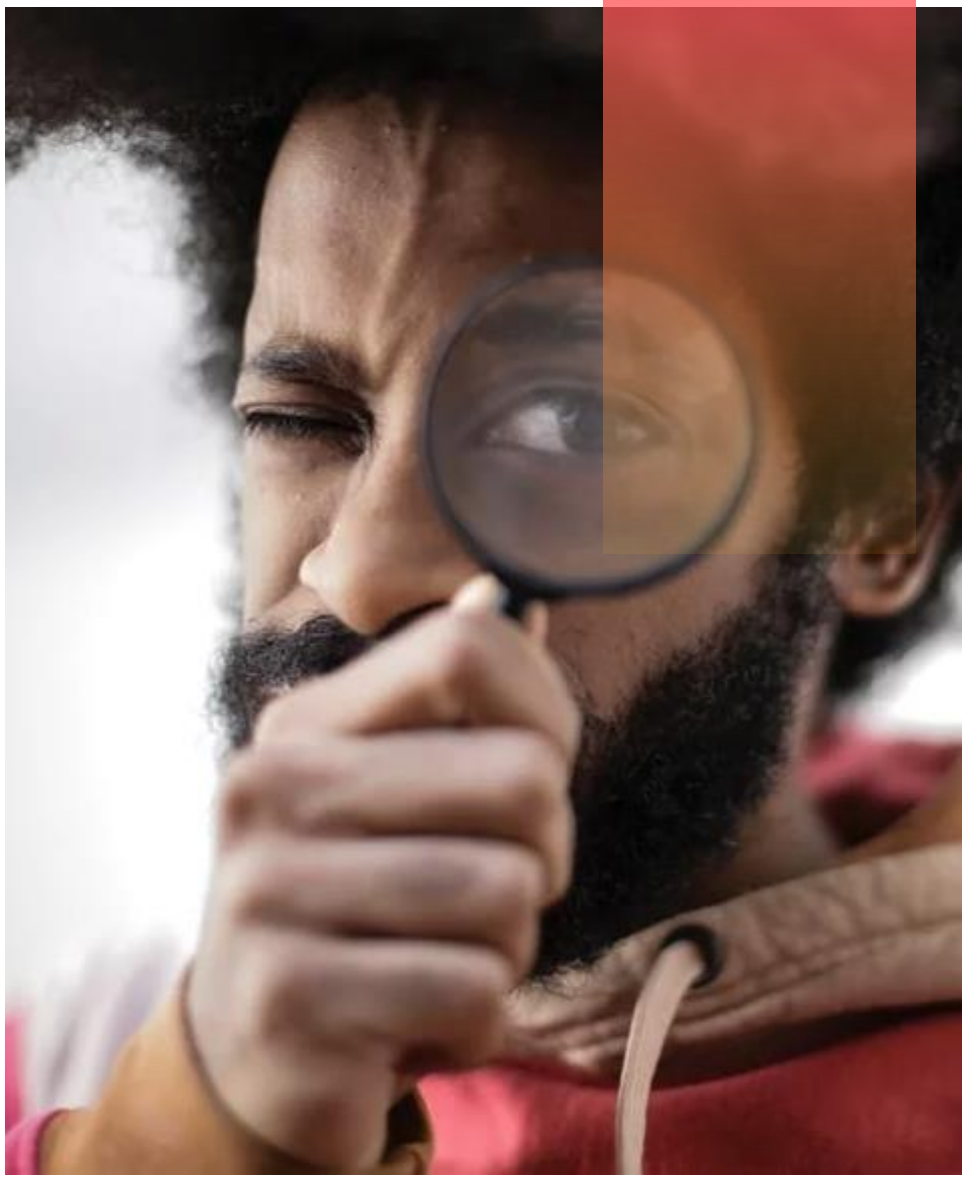
Mejor experiencia para el consumidor

La innovación de los datos dota a los consumidores de mayor información con la que tomar sus decisiones. Además, las empresas pasan de una producción en masa a una personalización en masa.

Tratamiento del dato

El volumen enorme de los datos hace imposible su tratamiento y análisis a partir de las herramientas de bases de datos y analíticas convencionales. Debido a este hecho ha sido necesario desarrollar **herramientas de Big data** que sean capaces de manejar dichos datos. Entre los **beneficios** que aportan encontramos:

- Toma de **decisiones más inteligentes**.
- Obtención de un **mayor conocimiento**.
- Elaboración de **soluciones óptimas**.
- Desarrollo **de mejores productos centrados en el cliente**.
- Aumentar la **lealtad del cliente**.
- **Proceso automatizados** con un análisis predictivo y descriptivo más preciso.



Marco arquitectónico Big data

Las herramientas de big data deben tener un **marco arquitectónico especial** para el tratamiento de los datos. Esta estructura se base en **capas donde cada una tendrá una función particular** que permite que los datos se vayan canalizando en función de los requisitos del sistema de procesamiento por lotes o del sistema de procesamiento de flujo.

Marco arquitectónico Big data

¿Cómo debe ser la estructura de capas?

- **Capa de ingestión de datos:**
  En esta primera capa los datos son clasificados en función de su prioridad.

- **Capa de recopilación de datos:**
  Se canalizan los datos de la capa de ingestión haciendo hincapié en el transporte de datos. Los componentes están desacoplados con el fín de apoyar el desarrollo de las capacidades analíticas.

- **Capa de procesamiento de datos:**
  Esta capa será la primera donde se realiza la analítica a partir de los datos obtenidos de la anterior.

- **Capa de almacenamiento de datos:**
  En esta capa aumenta el tamaño de los datos que se trataran, por lo tanto comienza uno de los grandes retos del big data, el almacenamiento eficiente. Deberemos de encontrar una solución de almacenamiento eficiente.

- **Capa de consulta de datos:**
  La función de esta capa será reunir el valor de so datos más útiles para al siguiente capa, para ello se llevará a cabo un procesamiento analítico sólido.

- **Capa de visualización de datos:**
  El dato se transforma en conocimiento y los usuarios de los canales de datos podrán extraer valor de los datos. Es en esta capa donde se podrán tomar las decisiones en función de los datos obtenidos.

Marco arquitectónico Big data

Las dos arquitectura más comunes que existen son **Arquitectura Lambda** y **Kappa.** La principal **diferencia** entre ambas será el flujo del tratamiento de datos ya que mientras Lambda utiliza procesamiento batch y streaming, Kappa utiliza solo procesamiento streaming.

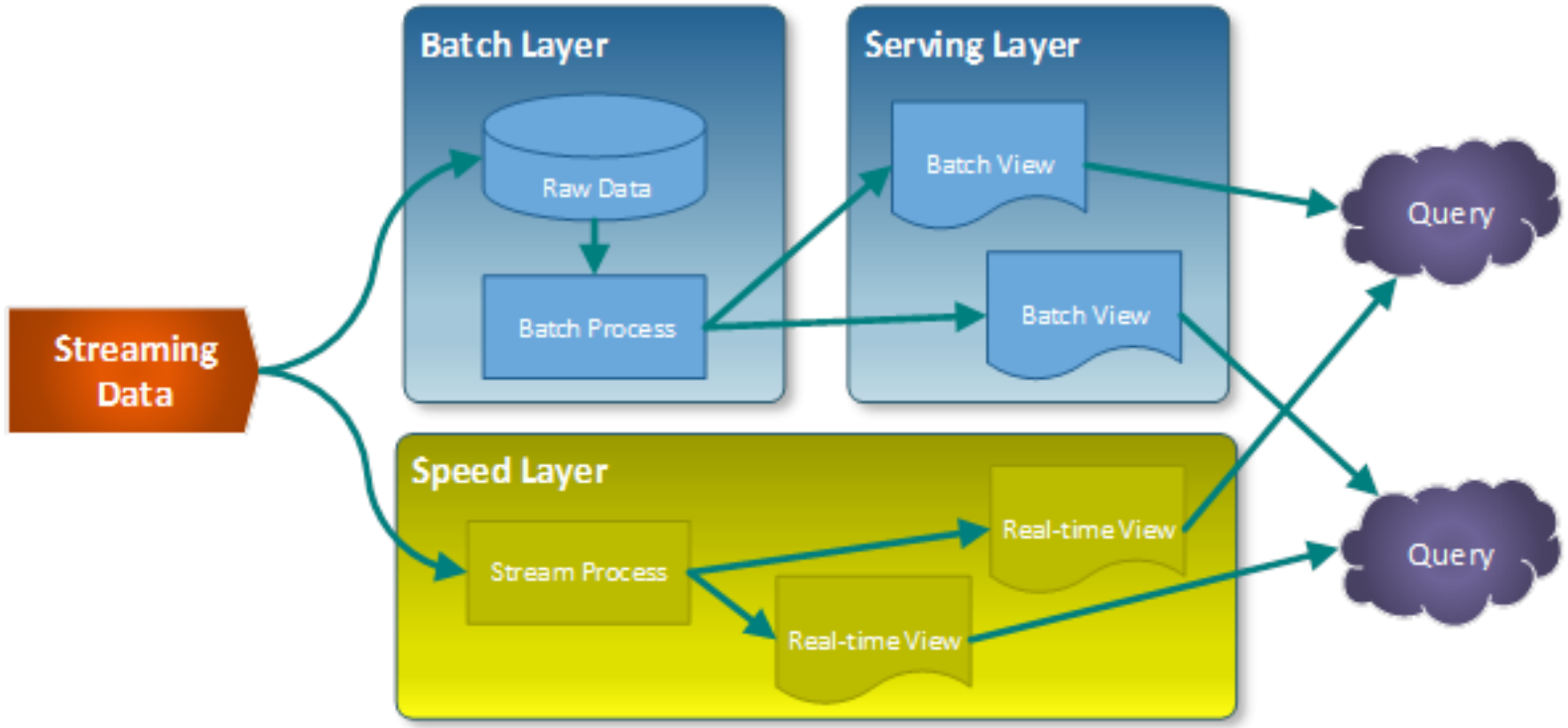
- **Batch**: el proceso de los datos que tiene un inicio y un fin en el tiempo.
- **Streaming**: el proceso de datos no tiene un fin temporal, está continuamente recibiendo y tratando nueva información.

Resumiendo, con el procesamiento batch seremos capaces de procesar volúmenes de datos en tiempos espaciados (ej. Cada 10 minutos). Mientras que con el procesamiento streaming podremos procesar datos casi al instante en que son producidos.

Marco arquitectónico Big data

Arquitectura Lambda

Este tipo de arquitectura fue desarrollada en 2012 por Nathan Marz con el objetivo de obtener un sistema robusto tolerante a fallos, tanto humanos como de hardware. Además, este sistema debería ser linealmente escalable y permitir realizar escrituras y lecturas con baja latencia.



Marco arquitectónico Big data

Arquitectura Lambda

1. El sistema recoge la información y la envía a la capa Batch ya la capa de streaming (speed layer).

2. La capa Batch será la encargada de procesar la información sin modificar. Una vez procesada realizará un tratamiento mediante un proceso batch mediante el cual se obtendrá las Batch Views. Las batch views se utilizarán en la capa que ofrece los datos para ofrecer la información ya transformada al exterior.

3. Una vez obtenidos los batch views, la capa Serving Layer los indexará para que puedan ser consultados con baja latencia.

4. Por otro lado, la capa Speed Layer solo tendrá en cuenta los datos nuevos y compensará la alta latencia de las escrituras que se producen en la capa Servin Layer.

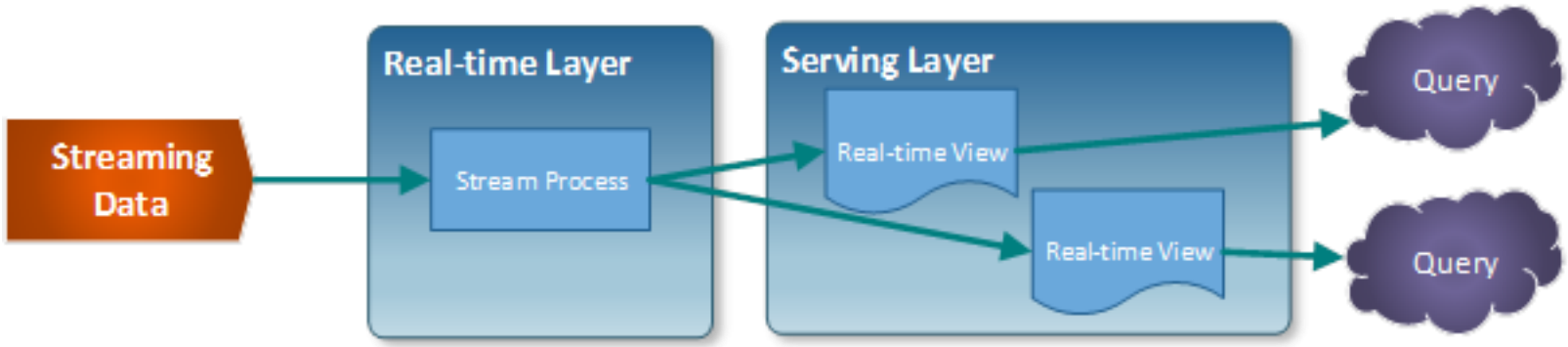
5. La combinación de los resultados de las barth views y de las vistas en tiempo real darán lugar a las respuesta a las consultas realizadas.

Para ver ejemplos reales de este tipo de arquitectura puedes acceder a la web http://lambda-architecture.net/

Marco arquitectónico Big data

Arquitectura Kappa

Este tipo de arquitectura fue desarrollada en 2014 por Jay Kreps a raíz de detectar posibles puntos débiles en la arquitectura Lambda. Esta arquitectura consiste en eliminar la capa batch del proceso y solo dejar la capa de streaming, ya que al no tener un fin temporal, está continuamente procesando datos nuevos. Por lo tanto, la evolución de la arquitectura Lambda que propone Jay Kreps se basa en una simplificación de esta arquitectura que permite procesar los datos en tiempo real y el reprocesamiento continua en un motor de flujo único.



Marco arquitectónico Big data

Arquitectura Kappa

Esta arquitectura trata todas las operaciones como stream, es decir, las operaciones batch son un subconjunto de las operaciones de streaming.

1. En la primera capa, la capa de almacenamiento, los datos continuamente son guardados sin modificarse.
2. En la segunda capa se maneja la información de manera asíncrona de tal forma que diferencia la información en:

   Event time: tiempo en el que se genera la información

   Ingestion time: momento en que se los sistemas reciben la información.

   Processing time: tiempo en el que se está procesando la información.

3. En la última capa, la capa de servicio, se muestran los resultados de la información procesada y de la información original.

No es un reemplazo para la arquitectura Lambda, excepto donde se ajusta su caso de uso.

Marco arquitectónico Big data

Lambda vs Kappa

Lambda

Encaminada al análisis tradicional de los datos.

Almacena datos de manera periódica y procesa grandes volúmenes de datos estáticos (batch) al mismo tiempo que se procesas datos dinámicos (streaming). Combina volumen y velocidad.

Kappa:

Encaminada al análisis en tiempo real.

Minimiza el tiempo al no almacenar los datos, ya que los procesa cuando los recibe.

Recomputa los datos en la capa streaming a no ser que se produzca un cambio en la lógica de negocio, que en este caso, lo hará en al capa batch.

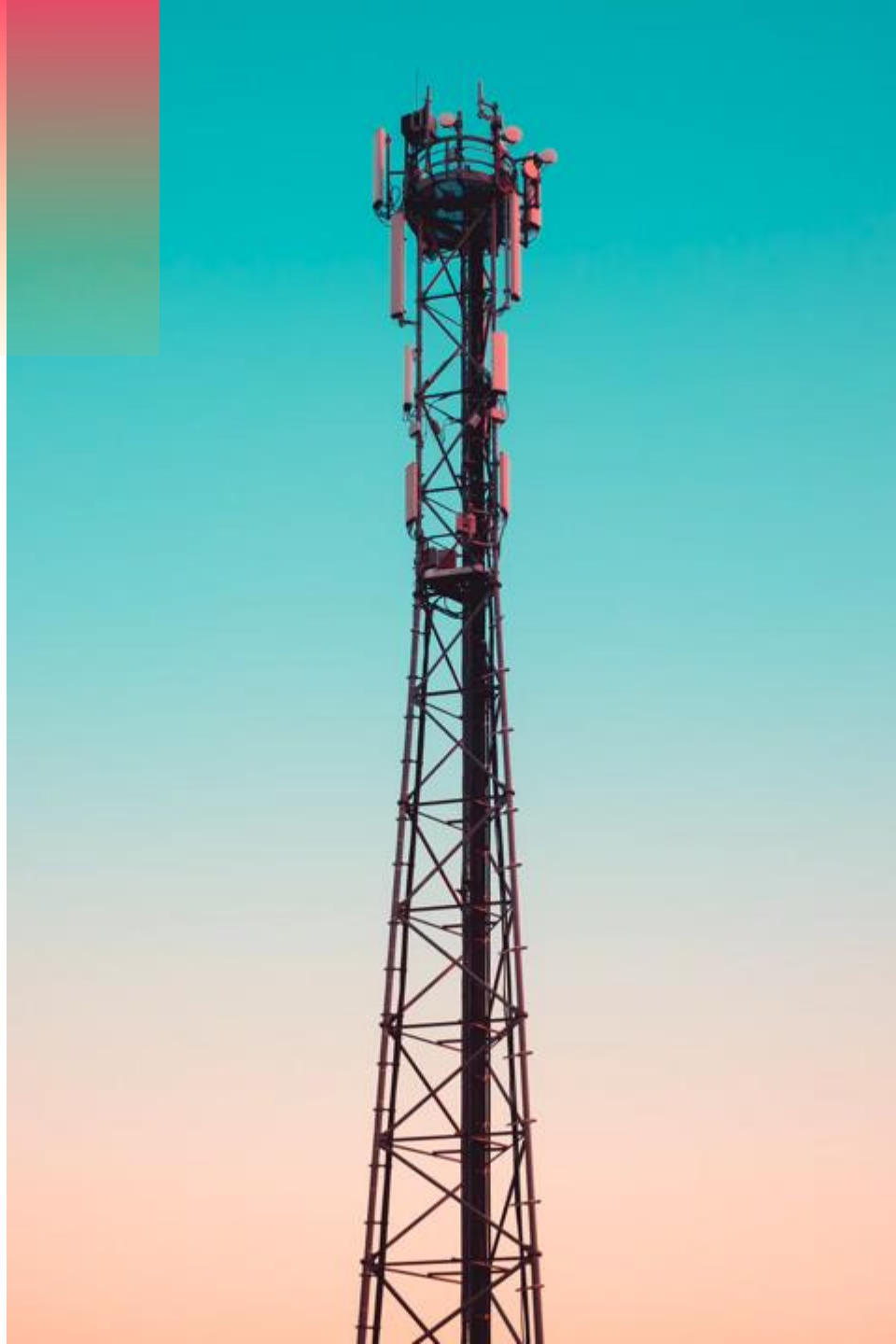
Marco arquitectónico Big data

Ejemplos arquitecturas

Vamos a ver las dos tipos de arquitecturas a través de un ejemplo:

Supongamos que deseamos conocer la cercanía de los usuarios a una determinada antena de telefonía móvil en tiempo real. Para ello utilizaríamos los datos que nos proporcionan los dispositivos móviles, ya que cada vez que el usuario recibiese cobertura de esa antena se generaría un evento. Este evento se procesaría en la capa de streaming y podríamos dibujar sobre un mapa el desplazamiento del usuario con respecto a su posición anterior. Por lo tanto utilizaríamos la arquitectura Kappa.

Por otro lado, si necesitásemos acceder a todo el conjunto de datos sin penalizar el rendimiento utilizaríamos la arquitectura Lambda.

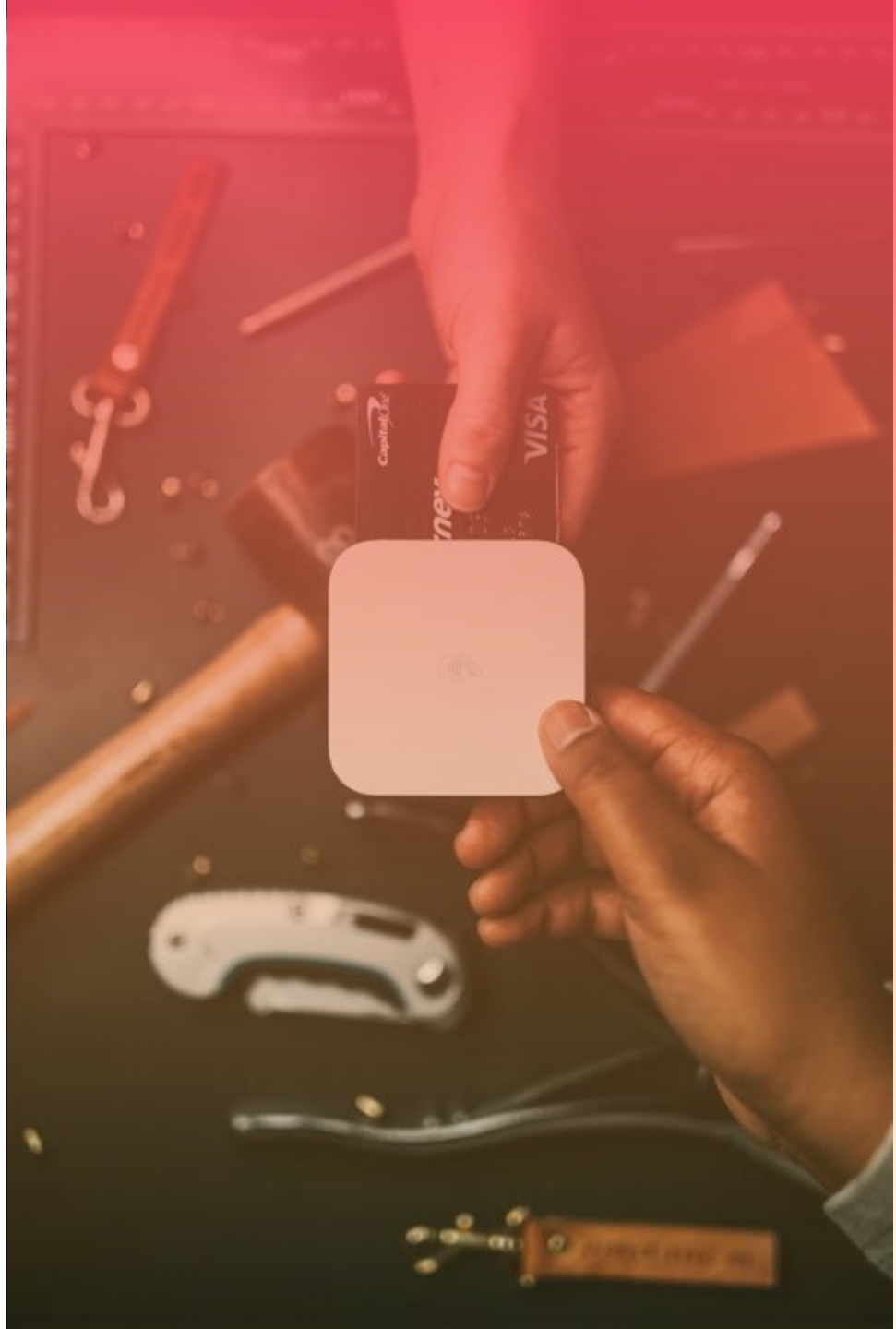


Marco arquitectónico Big data

Ejemplos arquitecturas

Un ejemplo de arquitectura Lambda podría ser un sistema que recomiende post a los usuarios que navegan por un blog en función de los gustos de estos. En este caso, la capa batch será la encargada de entrenar un modelo que mejorará las predicciones. Por otro lado, la capa streaming se encargará de las valoraciones en tiempo real de los usuarios.

Por último, vamos a ver un ejemplo muy claro de la arquitectura Kappa. Imagina que deseamos bloquear una tarjeta de crédito que aparenta ser fraudulenta. En este caso no podremos utilizar la arquitectura Lambda, ya que nos llevaría varios minutos realizar la consulta. En este caso es mejor tomar la decisión casi inmediatamente, por lo que utilizaremos la arquitectura Kappa.



Tipología del dato

Siguiendo la clasificación de datos de Joyanes (2013), se distinguen tres categorías de datos según su procedencia:

- **Datos estructurados**. Se presentan en un formato o esquema bien definido y que poseen campos fijos. Son hojas de cálculo, archivos, bases de datos tradicionales provenientes de CRM, ERP, etc., que han sido recolectados por profesionales del marketing en algún momento.

- **Datos semiestructurados**. No tienen formato definido, pero sí contienen etiquetas u otros marcadores con el fin de clasificar los elementos de estos. En esta categoría encontramos textos con etiquetas XML y HTML.

- **Datos no estructurados.** Los más numerosos. Son datos de tipo indefinido, almacenados principalmente como documentos u objetos sin estructura fija ni bajo ningún patrón concreto. Pueden ser generados por máquinas y personas. Son archivos de audio, vídeo, fotografía y formatos de texto libre como emails, SMS, artículos, WhatsApp, etc.

Tipología del dato

Al contrario de lo que se piensa, mayoría de los datos generados son de tipo personal. La mayoría de las veces los datos son creados por sensores conectados a una máquina. No obstante cuando tratemos datos de tipo personal tendremos que tener en cuenta la LODP.

LOPD.

Actualmente, 250 millones de personas utilizan internet cada día en Europa. Los consumidores valoran enormemente su privacidad en línea

NORMATIVA PROTECCIÓN DE DATOS

El Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos entró en vigor en 2016 y deroga la Directiva 95/46/CE, siendo aplicable directamente, sin necesidad de transposición al ordenamiento jurídico español, desde el 25 de mayo de 2018.

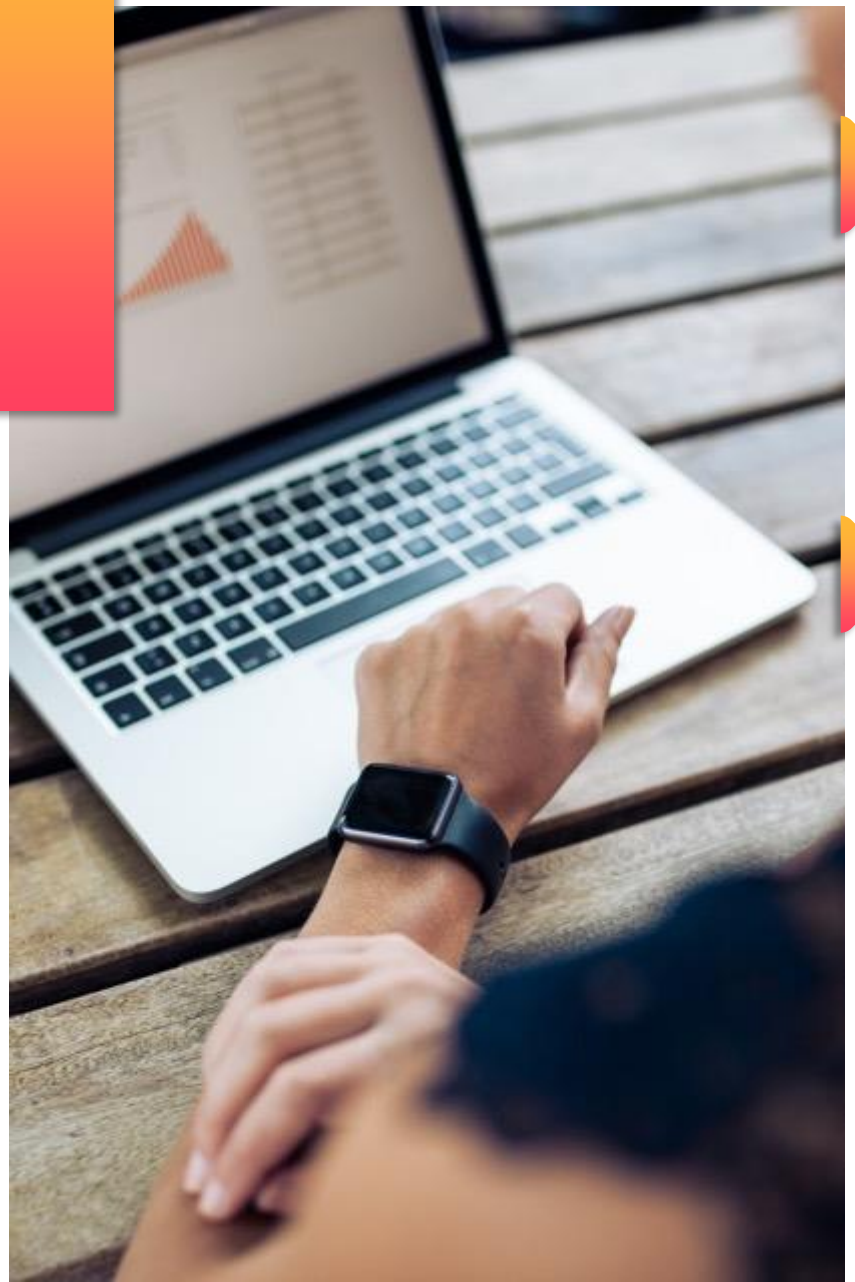
Comienza también a aplicarse la nueva LOPD (Ley Orgánica de Protección de Datos de Carácter Personal): La Ley Orgánica de Protección de Datos Personales y Garantía de los derechos digitales (LOPDGDD) que especifica y complementa ciertos apartados del RGPD, sin modificarlo.

¿Qué problemas previos trata de resolver el RGPD?

Hasta la aparición del Reglamento existía normativa al respecto, como la Directiva 95/46/CE, pero se detectó que en materia **de protección de los datos en la Unión no se aplicaba de manera uniforme**, incrementando la inseguridad jurídica, con una percepción generalizada de riesgo sobre los datos y en particular en relación con las actividades online.

¿Qué problemas previos trata de resolver el RGPD?

Desde el 25 de mayo de 2018, con la aplicación efectiva del Reglamento general de protección de datos, hay un único conjunto de normas de protección de datos para todas las empresas que operan en la Unión Europea (UE), con independencia de dónde tengan su sede. Además, desde el punto de vista del interesado, que es la persona física cuyos datos personales son objeto de tratamiento, “El RGPD de la Unión Europea le ayuda a tomar el control de esta información mediante varios derechos clave que le conceden mayor poder para protegerse”.



¿Quién supervisa el cumplimiento RGPD?

AEPD: Agencia Española de Protección de datos

¿Cuál es el grado de cumplimiento de cada empresa?

La **AEPD** ofrece herramientas que permiten valorarlo

Veamos la LOPDGDD


LEY ORGÁNICA DE PROTECCIÓN DE DATOS PERSONALES Y GARANTÍA DE LOS DERECHOS DIGITALES
https://vimeo.com/338819761
02:27

¿Qué factores de riesgo debemos tener en cuenta?

Debemos establecer hasta qué punto una actividad de tratamiento de datos personales puede causar algún tipo de **daño** a los titulares de datos

La herramienta gratuita **Facilita RGPD** establece 3 factores de riesgo que vamos a ver a continuación.

FACTORES DE RIESGO 1

Si su empresa pertenece a alguno de los siguientes **sectores**:

- Sanidad
- Entidades bancarias y financieras
- Actividades de servicios sociales
- Publicidad
- Videovigilancia masiva

FACTORES DE RIESGO 2

Si su empresa pertenece a alguno de los siguientes **sectores**:

- Datos que revelen origen étnico o racial
- Datos de opiniones políticas o religión
- Datos de afiliación sindical (excepto cuotas sindicales)
- Datos genéticos
- Datos biométricos dirigidos a identificar de manera unívoca a una persona

Otras categorías especiales:

- Datos de salud física o mental
- Datos relativos a la vida sexual o la orientación sexual
- Datos relativos a condenas o a infracciones penales
- Geolocalización

FACTORES DE RIESGO 3

Si su empresa realiza alguno de los siguientes tratamientos:

- Hacer o analizar perfiles.
- Hacer publicidad y prospección comercial masiva a potenciales clientes.
- Prestar servicios de explotación de redes públicas o de comunicaciones electrónicas.

Otros tratamientos de riesgo:

- Gestionar a miembros o asociados de partidos políticos, sindicatos, iglesias, confesiones o comunidades religiosas, fundaciones u otras entidades sin ánimo de lucro con finalidades políticas, filosóficas, religiosas o sindicales.
- Gestión y control sanitario o venta de medicamentos.
- Gestión de historiales clínicos o sanitarios.

¿A qué tratamientos se aplica el Reglamento?

Totalmente automatizados: cuando se tratan datos personales en una base de datos o se guardan en la nube. La elaboración de perfiles se basa en un tratamiento automatizado.

Tratamiento no automatizado: aquel que, por ejemplo, consiste en guardar contratos en un archivador y se ordenan conforme a un criterio lógico de búsqueda, tal como el número del contrato u otro identificador (apellido, número de cliente, etc.) que se asigne. También podría ser, por ejemplo, el tarjetero con tarjetas de visita.

Tratamiento parcialmente automatizado: que conlleva operaciones, procesos o procedimientos tanto automatizados no automatizados. Por ejemplo, una empresa que se dedica a elaborar perfiles pero cuenta con un archivador en el que almacena y ordena otros datos.

¿Cuál es el objeto del RGPD?

La protección de las personas físicas en lo que respecta al tratamiento de los datos personales.

La libre circulación de los datos personales.

¿QUIÉN QUEDA EXCEPTUADO DEL RGPD?

No se aplica el tratamiento:

- En el ejercicio de una actividad fuera del derecho de la Unión.
- Por los Estados miembros cuando sean tratamientos relativos a política exterior y seguridad común.
- En actividades personales o domésticas efectuadas por personas físicas.
- Por parte de las Autoridades competentes con fines penales.

OBLIGACIONES

Las principales obligaciones del responsable del tratamiento en virtud del RGPD son:

- Registro de actividades del tratamiento.
- Protección de datos desde el diseño y por defecto.
- Análisis de riesgos.
- Facilitar el ejercicio de los derechos.
- Notificar violaciones de seguridad a la autoridad de protección de datos y, cuando impliquen un alto riesgo, al interesado.
- Evaluación de impacto relativa a la protección de datos, cuando el tratamiento de los datos implique un alto riesgo para el interesado.

OBLIGACIONES

Otras obligaciones del responsable:

- Designar a un delegado de protección de datos en los casos previstos en el RGPD o la LOPDGDD.
- Contratar a encargados del tratamiento que reúnan garantías suficientes para cumplir con la normativa sobre protección de datos, firmando el correspondiente contrato que les vincule y establezca sus obligaciones.
- Transparencia e información más detallada cuando los datos personales se obtienen del interesado o de otra fuente.
- Medidas de seguridad en atención a cuál sea el riesgo que implique el tratamiento, lo que podría incluir una auditoría en cualquier momento, si así lo determina el responsable del tratamiento.

¡Lo conseguiste!

Has llegado al final de la unidad