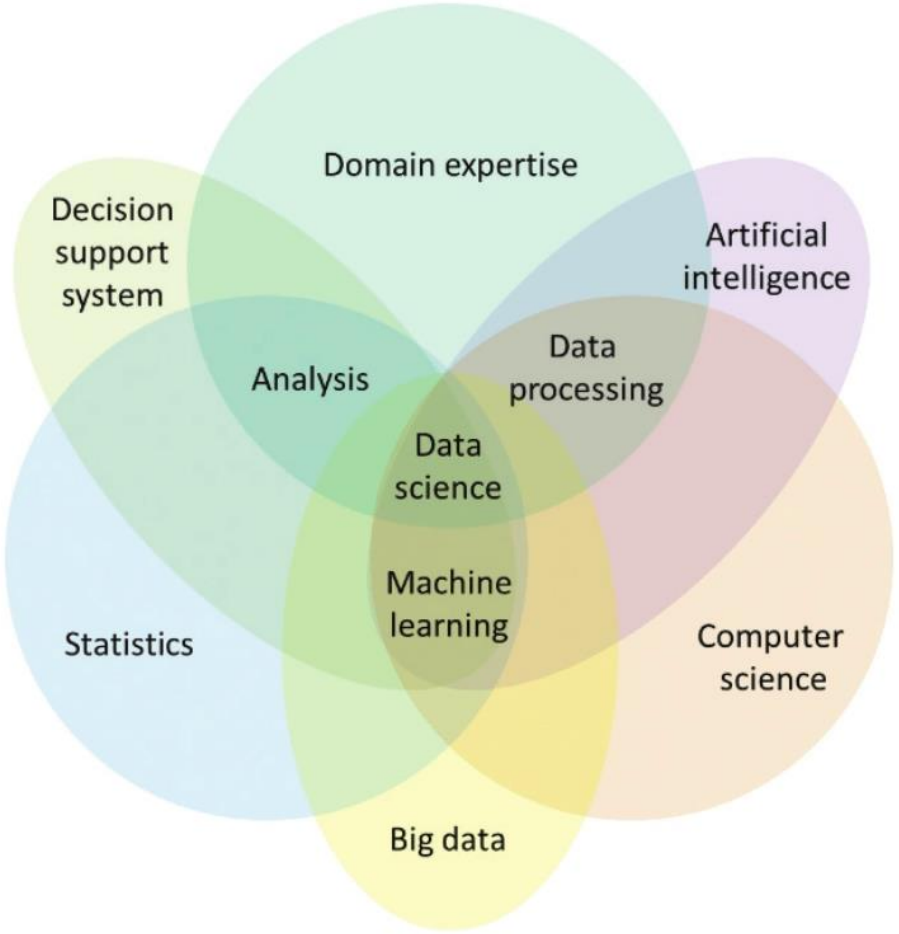
Introduccción al BigData

Ing. Ricardo Velasteguí (M.B.A.)

BigData

- IBM.- Tendencia, enfoque, entendimiento, toma de deciciones, utiliza enorme cantidades de datos (e,s,no e).
- Oracle.- Estrategia holística de gestión de información, tradicional y no tradicional.
- Gartner.- Manejan conjuntos de gran volumen, mejorar comprensión y toma de deciciones.
- Gualtieri.- Solución al crecimiento exponencial de Datos.
- Wikipedia.- Grandes cantidades de datos, patrones repetitivos.
- ISTA . - El análisis y gestión de grandes volúmenes de datos los cuales no pueden ser tratados de la manera convencional.

Entorno BigData



- Correlación con Big Data
  - Experiencia/dominio de campo (Domain expertise)
  - Estadísticas (Statistics)
  - Ciencias de la computación (Computer science)
  - Inteligencia de negocios (Decision support system)
  - Inteligencia Artificial (Artificial intelligence)
- Derivaciones Big Data
  - Análisis de datos (Analysis)
  - Procesamiento de Datos (Data processing)
  - Ciencia de Datos (Data science)
  - Aprendizaje automático (Machine learning)

Características del BigData

- Volumen.- Datos generados por máquinas, redes e interacciones personales, mensajes de redes sociales, clicks de las páginas, aplicaciones móviles, tráfico de red, sensores, etc.
- Velocidad.- Ritmo de adquisición de datos.
- Variedad.- Fuente de datos proveniente de sistemas informáticos, de redes sociales, y imagines, documentos, pdfs.
- Veracidad. – Precisión del dato (que tan fiable o verdadero).
- Valor.- Transformar, analizar y obtener su valor.
- Visualización.- Ver de forma visual los datos (sintetizados).
- Variabilidad.- Obsolescencia casi inmediata, significa que su valor cambia en el tiempo.

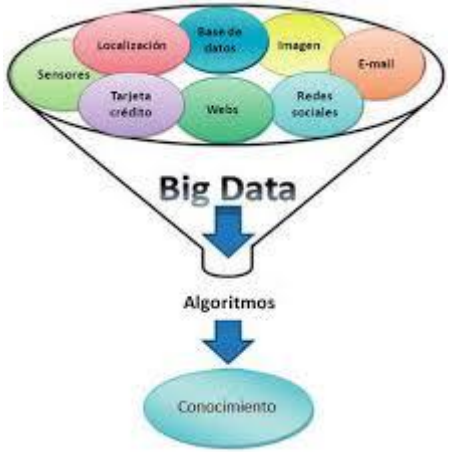
Ventajas e Inconvenientes del BigData

- Ventajas
  - Velocidad en toma de decisions
  - Feedback en tiempo real
  - Profesionales cualificados
  - Mayor eficiencia
  - Conocimiento del Mercado
  - Marketing especializado
  - Fidelización de Clientes

- Inconvenientes
  - Exceso de datos
  - Ataques informáticos
  - Vulneración de la privacidad
  - Seguridad y protección de datos

Fuentes de Datos

- Datos de internet y móviles.
- Datos de Internet de las Cosas.
- Datos sectoriales recopilados por empresas especializadas.
- Datos experimentales.



- Ejemplos:
  - Datos de RFID y NFC

  - Web

  - Telecomunicaciones

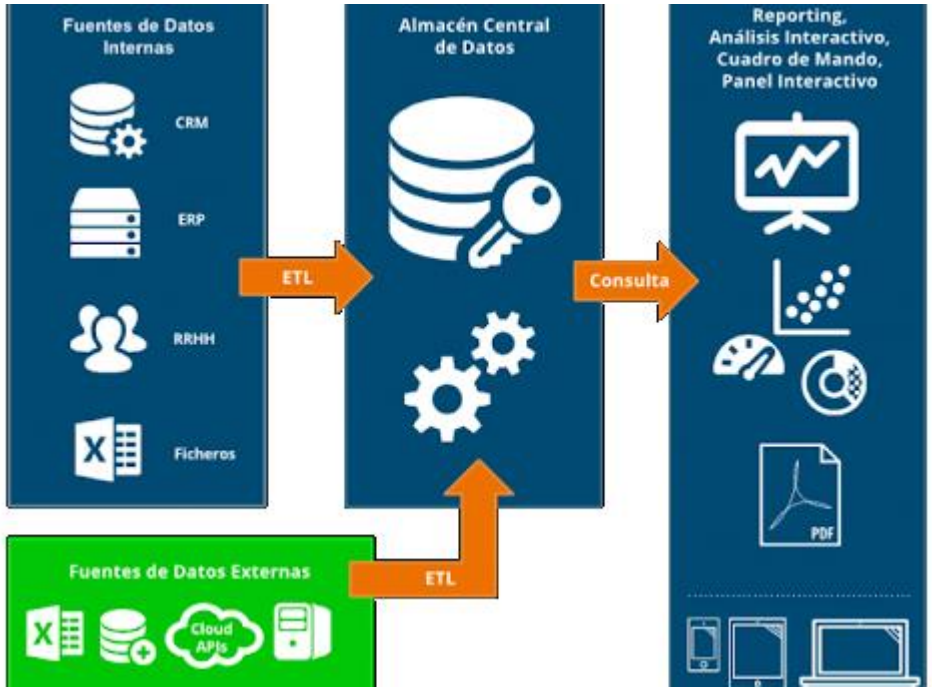
- Smart cities:
  - Seguridad ciudadana
  - Movilidad Urbana
  - Gestión de Energía
  - Gestión del Agua
  - Entre otros (gobiernos, mercados bursátiles, análisis de sentimientos, sensors, transacciones)

Grandes volúmenes de Datos comunes

Dominio	Escenarios de grandes volúmenes de datos comunes	
Servicios financieros	Modelado de riesgo verdadero Análisis de las amenazas y detección de fraude	Vigilancia del Comercio El puntaje crediticio
Medios y Entretenimiento	Los motores de recomendación Focalización	Buscar Calidad La detección y abusos de fraude de clicks
Venta al por menor (retail)	Punto de análisis de las transacciones de venta Análisis de la pérdida de clientes	El análisis de sentimientos (sentiment analysis)
Telecomunicaciones	Prevención del churn de clientes La optimización del rendimiento de la red	Detalles de llamadas (CDR) y su análisis Predicción de fallos de red
Gobierno	Seguridad Cibernética (botnets, fraudes) La congestión del tráfico y re-enrutamiento	Monitoreo Ambiental Monitoreo Antisocial a través de medios sociales
Salud	Investigación del genoma La investigación del cáncer	Detección temprana de pandemias Monitoreo de la calidad del aire

Almacenamiento (Datawarehouse)

- Un **Data Warehouse** sirve para recopilar y administrar datos de diversas fuentes, se utiliza principalmente para conectar y analizar los datos empresariales.



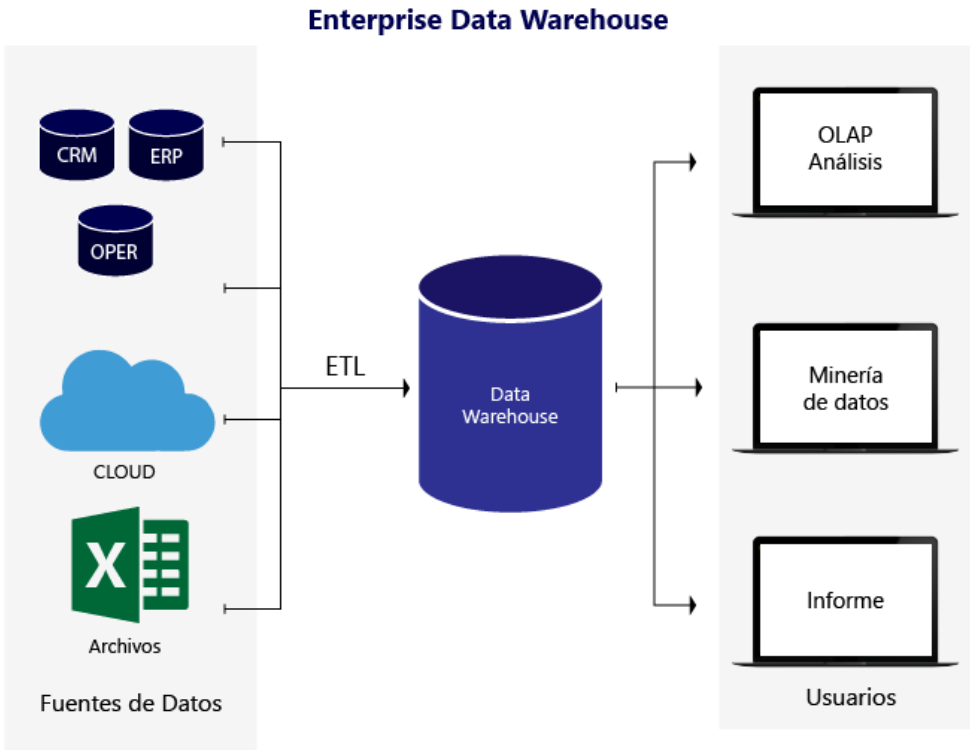
Características:

- Gran almacenamiento
- Consolidación de fuentes de datos
- Servicios: Hosting, Housing, Warehouse, Cloud.
- Tipos:
  - Enterprise Data Warehouse
  - Operational Data Store (ODS)
  - Data Mart
  - Data Lake

- La implementación de un Data Warehouse en las empresas es de suma importancia, ya que permite a los usuarios acceder de una manera fácil a una gran cantidad de datos, analizarlos y proporcionar información importante sobre las diversas actividades de la organización.

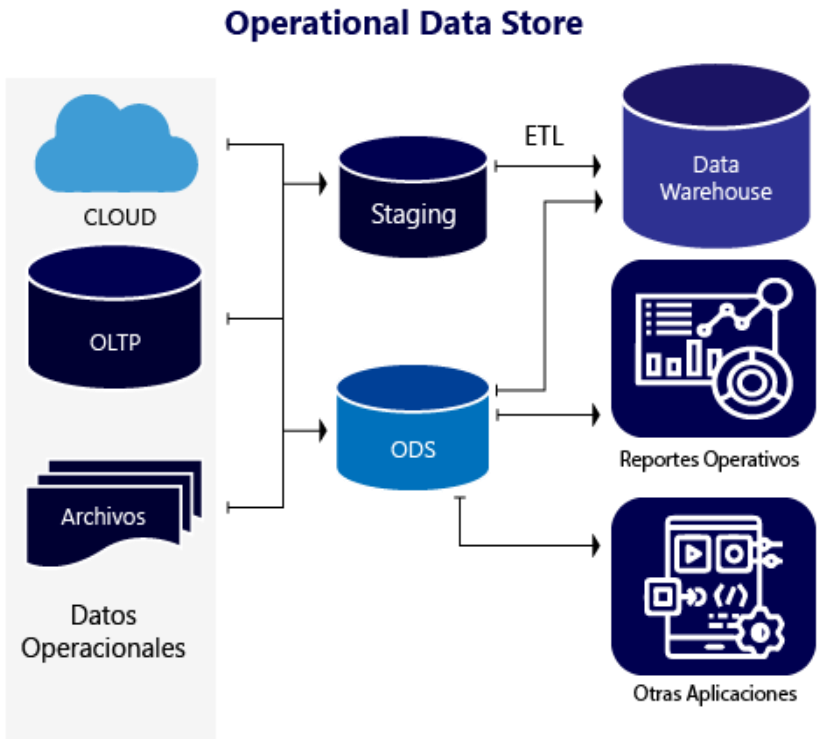
Enterprise Data Warehouse

- Es un almacenamiento de datos centralizado, unifica toda la información de una organización para que toda la empresa pueda tener acceso. Ofrece un servicio en el que apoya la toma de decisiones en la empresa. Se enfoca principalmente en organizar y representar los datos, también se obtiene la capacidad de clasificar los datos según el usuario y dar acceso de acuerdo a las restricciones internas.



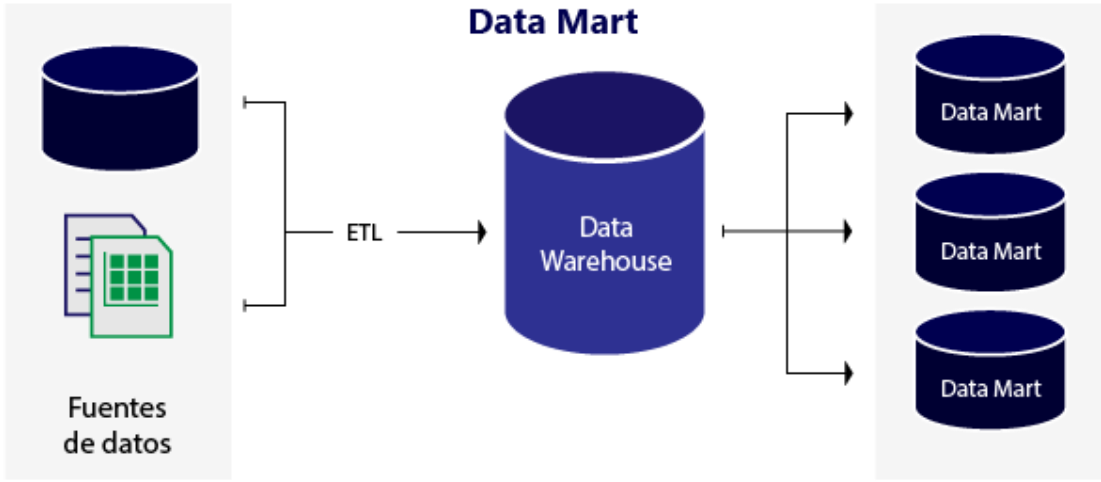
Operational Data Store (ODS)

- También conocido como ODS, es un almacén de datos, que cuando el almacenamiento de datos y los sistemas OLTP no admiten las necesidades de los informes de las organizaciones. En ODS, todo el almacenamiento de datos se actualiza en tiempo real y/o con baja latencia de actualización, y por eso mismo se utiliza habitualmente para actividades rutinarias, como es el almacenamiento de registros de la operación de la empresa y transacciones de venta. En pocas palabras, es un tipo de base de datos que se utiliza habitualmente como un área lógica provisional para un almacén de datos.



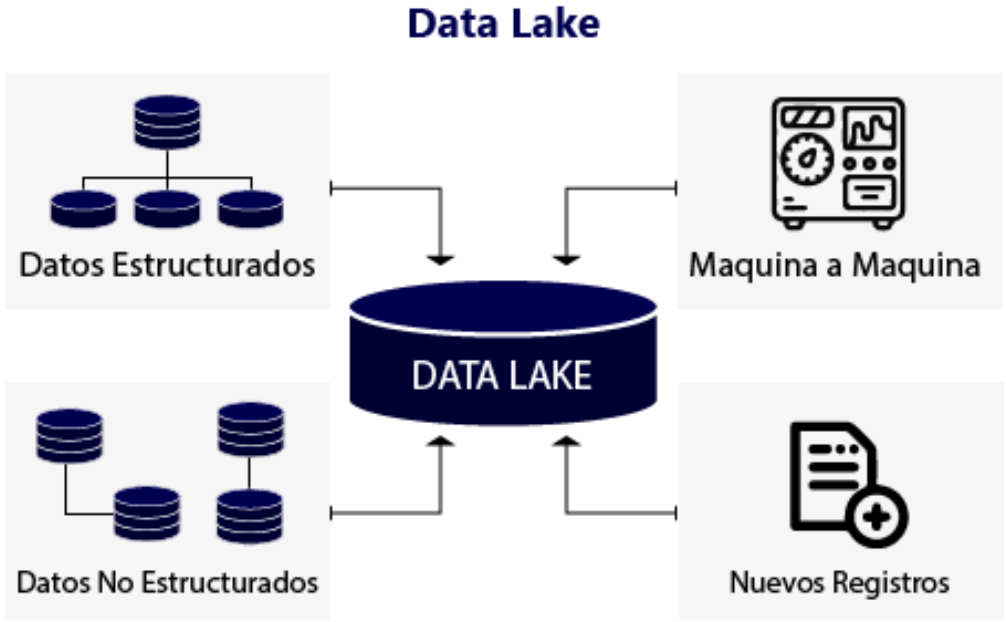
Data Mart

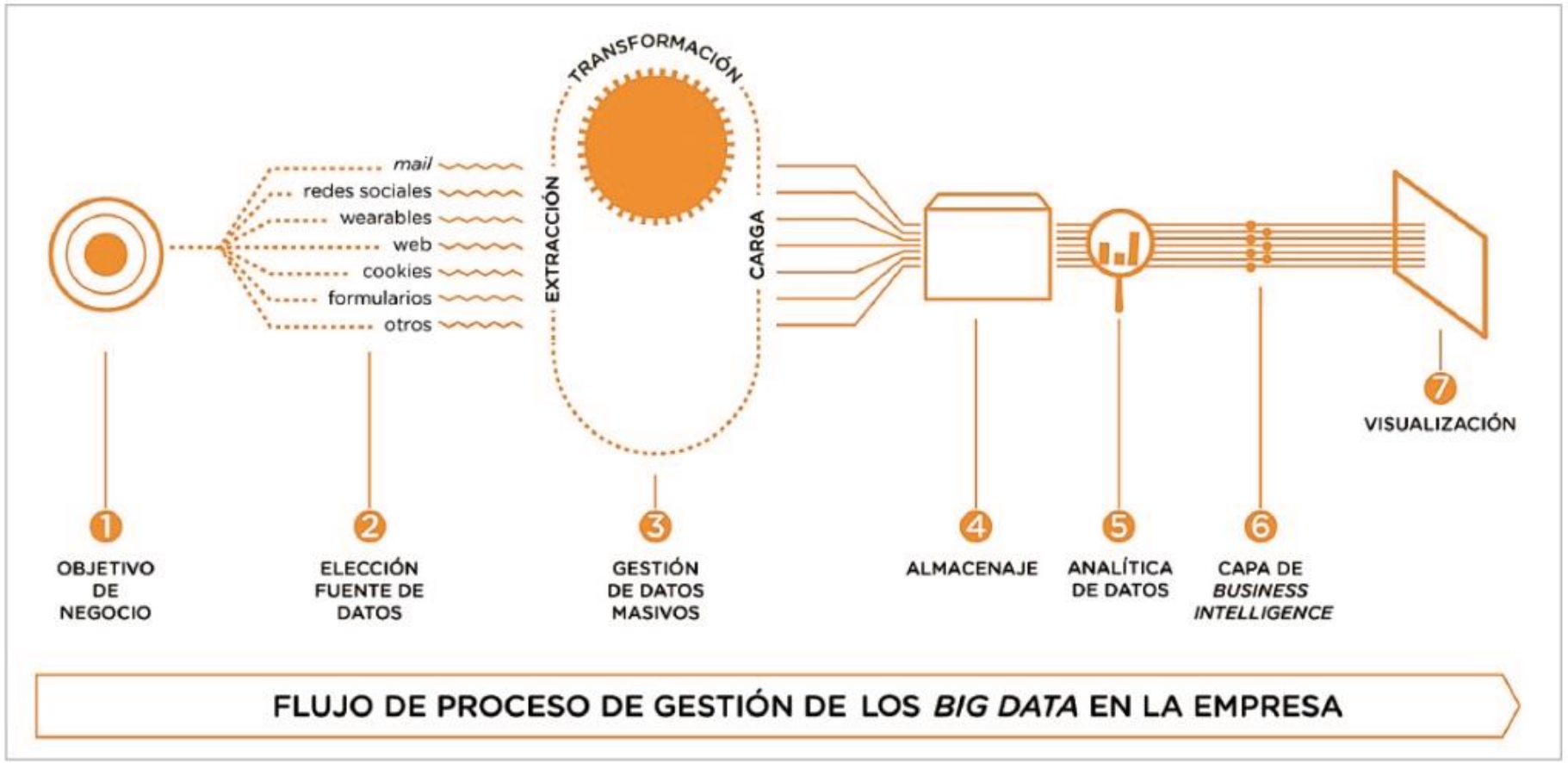
- Un Data Mart es un subconjunto del almacenamiento de datos orientado a un área específica, también conocido como base de información departamental. Está especialmente diseñado para una línea de negocio particular, como ventas o finanzas.
- En un Data Mart independiente, los datos pueden recopilarse directamente de las fuentes. Tiene diversas funciones como organizar la información para posteriormente analizarla, realizar indicadores (KPI), evaluar los objetivos del sector que se está analizando, etc.
- El objetivo es realizar un análisis detallado sobre lo que ocurre en un negocio.



Data Lake

- Es un repositorio centrado en almacenar gran cantidad de datos estructurados y sin estructurar sin importar su escala. Permite almacenar los datos tal cual vienen sin necesidad de ser estructurados. Ejecuta diferentes tipos de análisis, dashboards y visualizaciones, además de generar procesos de big data, análisis en tiempo real y de Machine Learning que facilitan la toma de decisiones.





FLUJO DE PROCESO DE GESTIÓN DE LOS *BIG DATA* EN LA EMPRESA

Definición de Big Data

Conjunto de tecnologías que permiten
...
de grandes conjuntos de datos distribuidos.

- Recopilación
- Almacenamiento
- Gestión
- Análisis
- Visualización

Tecnologías diseñadas para el tratamiento de datos

- BATCH – por lotes
- STREAMING – Tiempo real

Datos

- Estructurados
- Semiestructurados
- No Estructurados

Tipos de procesamiento del Big Data

Procesamiento por lotes BATCH

Permite procesar volúmenes de datos en tiempos espaciados, por ejemplo cada 10 minutos, 1 hora o diario.

El sistema dispone de lotes o batch en el que almacena toda la información que va obteniendo hasta completar un periodo.

Procesamiento en tiempo real - STREAMING

Permite procesar volúmenes de datos en tiempos lo más parecido a tiempo real que se pueda, hablamos de ordenes de 100 mili segundos a segundos.

Tipos de datos en Big Data

Datos Estructurados

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

No estructurados

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, consumían 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 vellorí 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.

Semiestructurados


```
{
    "marcadores": [
      {
        "latitude": 40.416875,
        "longitude": -3.703308,
        "city": "Madrid",
        "description": "Puerta del Sol"
      },
      {
        "latitude": 40.417438,
        "lonaitude": -3.693363,
      },
    ]
}
```

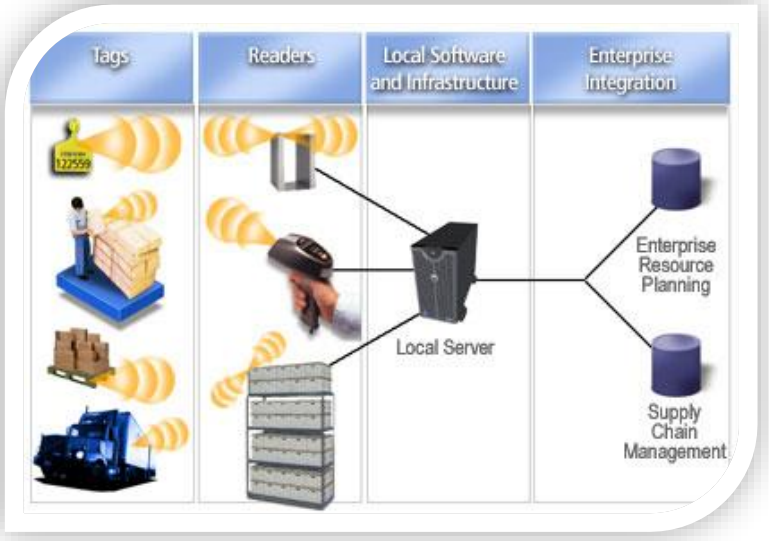
Gracias

Ing. Ricardo Velasteguí (M.B.A.)

Fuente de Datos

- RFID - Radio Frecuency Identification Tag, que traducido es una etiqueta de identificación por Radiofrecuencia, y consiste en un pequeño tag que contiene un número de serie único. Se coloca sobre objetos como paletas de envío o paquetes de productos. La etiqueta se puede adherir a todo tipo de cosas como mercancías, contenedores de envío, vehículos, etc. Un escáner electrónico puede usar señales de radio para leer o rastrear la etiqueta de identificación.

- Varias aplicaciones de los datos de identificación por radiofrecuencia en el comercio minorista son la gestión de activos, el seguimiento de la producción y el envío y la recepción.



- Las siglas NFC hacen referencia a su nombre en inglés, Near Field Comunication, que podemos traducir como comunicación de campo cercano.

- Se trata de una tecnología inalámbrica de alta frecuencia con un radio de acción es muy bajo, por lo que para utilizarla deber estar a un mínimo 15 cm del dispositivo con el que interactúas.

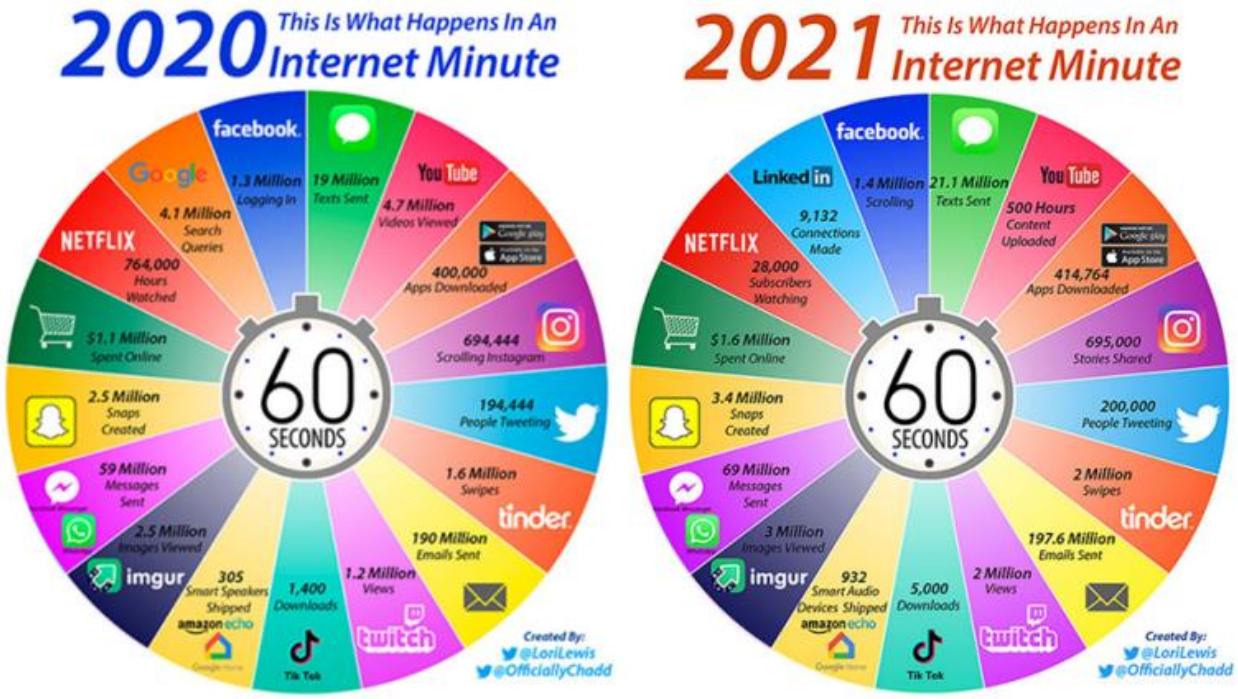
- Su funcionamiento se basa en la creación de un campo electromagnético en el que, mediante inducción, se genera un intercambio de información entre ambos dispositivos.

- Para que la tecnología NFC funcione son necesarios dos dispositivos, un emisor y un receptor.

- Todos los dispositivos móviles de última generación disponen de tecnología NFC.

Fuente de Datos

- El gráfico que se muestra, ilustra lo que sucede dentro de Internet en un minuto típico en 2021. Como siempre, hay un par de categorías diferentes rastreadas en este gráfico que el año pasado, pero la mayoría son las mismas y las que se trasladan son, una vez nuevamente, (casi) todo en comparación con el año pasado, algunos más que otros. Por ejemplo:



- No hace falta decir que la pandemia puede haber influido en algunos de estos aumentos en el uso de Internet, ya que muchas personas tenían más tiempo libre (y menos opciones de entretenimiento con espectáculos y cines cerrados).

- Los mensajes de texto de iPhone por minuto aumentaron de 19 millones por minuto a 21,1 millones;
- Las compras en línea por minuto han aumentado de $ 1,1 millones a $ 1,6 millones;
- Los mensajes a través de Facebook Messenger y WhatsApp por minuto han pasado de 59 millones a 69 millones ;
- Los golpes de yesca por minuto han aumentado de 1,6 millones a 2 millones;
- Los correos electrónicos enviados por minuto han aumentado de 190 millones a 197,6 millones;
- Los snaps creados en Snapchat por minuto han aumentado de 2.5 millones a 3.4 millones (¡oh, snap!);
- Las visualizaciones en Twitch por minuto han aumentado de 1,2 millones a 2 millones ;
- Y el mayor salto fue el de Tiktok: ¡de 1400 descargas por minuto a 5000 (más de 3,5 veces más)!

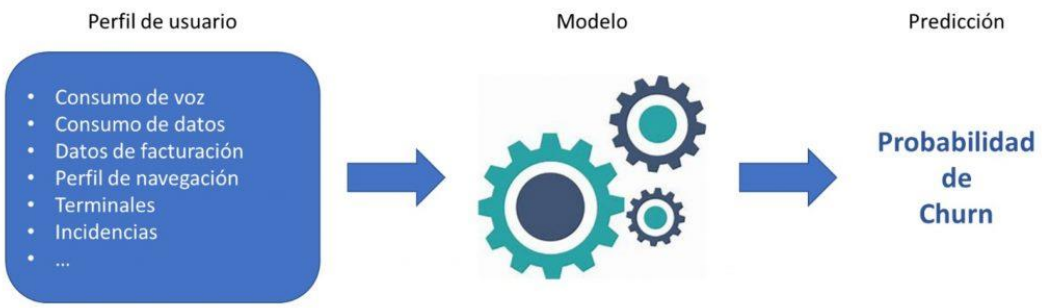
Fuente de Datos

- Las operadoras de telecomunicaciones pueden obtener información muy importante sobre sus clientes, pero por diversos motivos, aún no han podido extraer y explotar todo el valor estratégico y rentabilidad económica de dichos datos.
- Los datos son tanto estructurados (perfil del cliente, peticiones de servicios, tarificación, incidencias técnicas generadas, etc.), como no estructurados (documentos, vídeos, imágenes, contenido Web, localización, presencia, DPI señalización, "logs", grabaciones del "contact center", etc.) y parcialmente estructurados (perfil del cliente enriquecido con CDRs o "call data records" e información externa como blogs, foros, redes sociales, etc.).
- Los operadores tienen información demográfica del cliente, saben dónde vive, saben dónde está, cuándo se conecta, cuánto ancho de banda consume, qué sitios Web visita, qué aplicaciones utiliza, etc.

- Mediante DPI (Deep Packet Inspection) incluso podrían obtener información adicional en tiempo real sobre las necesidades y los gustos e intereses del cliente, si bien actualmente se encuentran limitados por cuestiones legislativas sobre privacidad y confidencialidad.
- Toda esta información adquirida de diversas fuentes debe ser organizada y luego analizada para dar soporte a la toma de decisiones.
- En efecto, "big data" es una poderosa tecnología que pueden explotar para predecir y reducir el ratio de abandono de clientes, impulsar la fidelidad de éstos mediante ofertas especiales o que combinen diversos productos, proporcionar productos y servicios personalizados, mejorar la eficiencia interna, optimizar la infraestructura de red, etc. La información obtenida, puede ser además utilizada para ofrecer nuevos productos y servicios a empresas.

Big Data para la prevención de churn

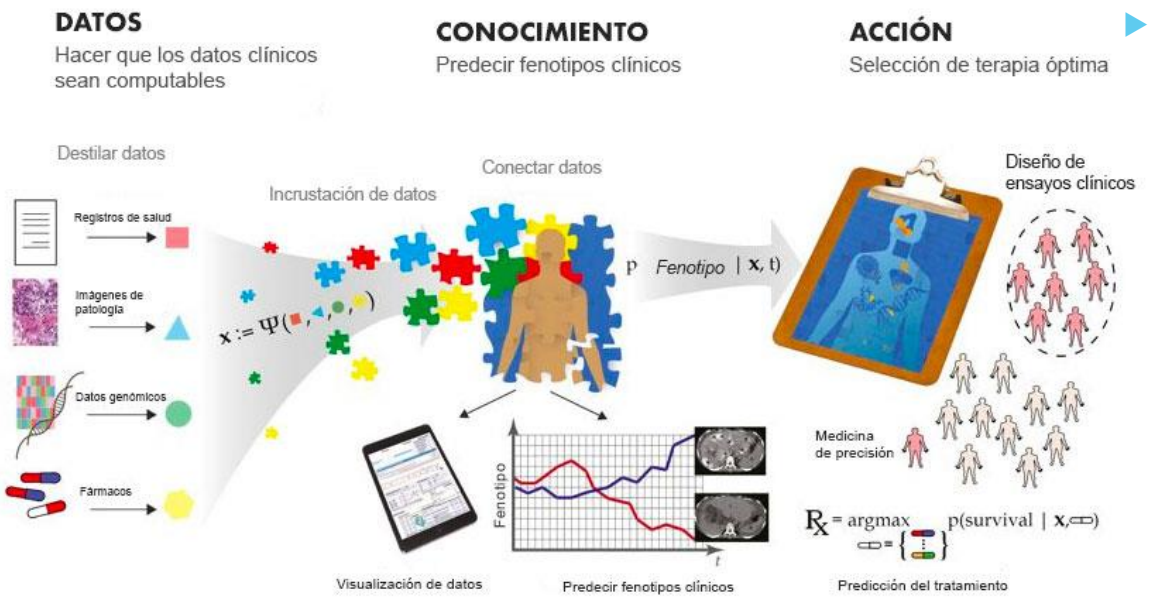
- El Churn es la métrica que mide la pérdida de clientes. Independiente del motivo, él calcula cuántas personas cancelaron cierto servicio o dejaron de comprar cierto producto. La razón de la discontinuidad puede tener su explicación en diferentes factores. Para las empresas queda, entonces, la prevención.
- La forma más acertada de prevenir el Churn es analizar los datos. Con un mapeo concreto del comportamiento de los clientes en toda su jornada de compra, es posible identificar cuáles son los puntos de ruptura que los hacen desistir de consumir algo. Haciendo un análisis constante de los datos, la empresa es capaz de identificar qué clientes están en riesgo de discontinuar la compra y actuar para evitarlo.



- Paso a paso del uso de datos para evitar el Churn

1. Mapear el cliente
2. Mantener una base de datos (CRM – Customer Relationship Management, CDP - Customer Data Platform)
3. Reconquista del cliente

BigData en la prevención del cáncer



- Una enfermedad no afecta igual a una persona sedentaria que una activa; tampoco es el mismo si una persona tiene predisposición genética a sufrir una dolencia y otra, no; o si vive en el campo o en la ciudad. Las variables que intervienen en el desarrollo del cáncer son diversas, pero su conocimiento y análisis pueden ayudar a entender mejor la enfermedad, personalizar los tratamientos y predecir factores de riesgo con más antelación. El big data es un aliado en la lucha contra el cáncer.

- La investigación contra el cáncer cada vez está más relacionada con las técnicas de datos masivos. Las aplicaciones son muchas: desde tomar decisiones en los tratamientos oncológicos hasta predecir cómo evolucionará el cáncer en un paciente concreto. El sector salud es consciente y aumenta cada año la inversión en este tipo de informaciones.
- Como los datos disponibles y recopilados son muchos, son necesarias herramientas y procesos muy complejos. Lo más importante es discernir cuáles son los datos relevantes y "tener un equipo multidisciplinario muy coordinado para analizarlos bien y generar información de valor“.
- El cáncer es la segunda causa de muerte en el mundo, según la Organización Mundial de la Salud (OMS). Cada año mueren unos nueve millones de personas por esta enfermedad. Según la comunidad científica, la clave no sólo se encuentra en el tratamiento, sino, especialmente, en la prevención y la reducción de riesgos de sufrirla.