Contenido

MODULO 1: Business intelligence y Advanced Analytics

DIA 1 – Ecosistema DATA

- Introducción y contexto
- Entendiendo los datos
- Big Data
- Ejercicios (sin uso de aplicaciones)

DIA 2 – Tecnología I: Transformación digital

- Google Trends es una herramienta para explorar y comparar la tendencia de determinados términos de búsqueda y campos concretos en las consultas que han hecho los usuarios en el motor de búsqueda
- **Insight**: se utiliza en investigación de mercados, marketing, comunicación y en la empresa en general para referirse a un descubrimiento, una idea reveladora que nos da la clave para poder resolver un problema.
- **IoT**: la Internet de las cosas (IoT) describe la red de objetos físicos ("cosas") que llevan incorporados sensores, software y otras tecnologías con el fin de conectarse e intercambiar datos con otros dispositivos y sistemas a través de Internet. Estos dispositivos van desde objetos domésticos comunes hasta herramientas industriales sofisticadas.
- **Marketing Digital:**
  - SEO: todo el tráfico generado de forma orgánica proveniente de buscadores (google, yahoo, bing, duck duck go, etc.)
  - SEM/PPC: todo el tráfico generado por herramientas de publicidad online (google ads, bing ads, facebook ads, linkedin ads, etc.)
  - Social Media: tráfico generado de forma orgánica por las redes sociales en cualquier formato.
  - Marketing de afiliación: tráfico generado por una fuente externa a cambio de una compensación previamente acordada.

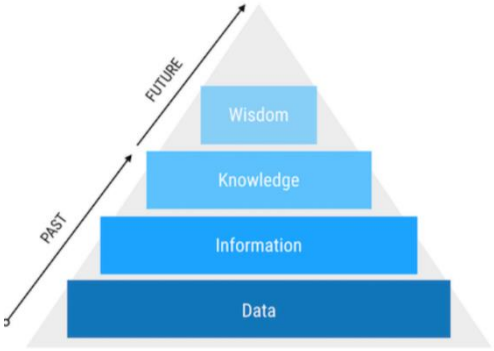
    - Email Marketing: tráfico generado a consecuencia del lanzamiento de campañas de email marketing.
- **Analítica digital:** Google Analytics es la herramienta más utilizada del mundo en el campo de la analítica web. Ofrece información agrupada del tráfico que llega a los sitios web según la audiencia, la adquisición, el comportamiento y las conversiones que se llevan a cabo en el sitio web.

- Búsquedas avanzadas en Google: uso de comandos INFO, SITE, OR, AND, “”, INTEXT, INTITLE

- https://www.xataka.com/basics/25-codigos-funciones-trucos-para-buscar-google-exprimiendo-al-maximo-su-motor-busqueda

DIA 3 – Business Intelligence

- **Estructura**:
    - **Datos**: los cuales por sí solos no aportan ninguna información y están dispersos por toda la organización y en distintos formatos.
    - **Información**: reunión de todos los datos en un formato en el que pueda leerlo. Toda la información que tengamos identificada, categorizada, etiquetada o calculada tras la recogida de datos.
    - BO: Business Operation
        - Estructura y orden básico aplicado a los datos
        - Tablas, documentos, listas, carpetas, etc.
    - BI: Business Intelligence
        - Estructura y orden aplicada a los datos
        - Data Warehouse, DataMart



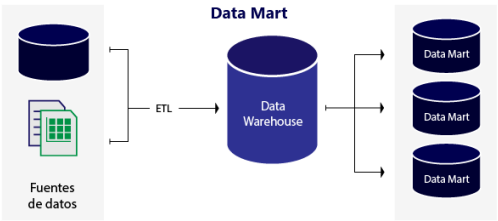
    - **Conocimiento**: se deriva de las personas y es intangible y empírico.
    - **Decisiones**: implica el funcionamiento de un sistema de BI implementado que me permite tomar decisiones.

- Por el momento NO se está utilizando al 100%. No se tienen bien definidos los perfiles necesarios. Las empresas todavía lo están implementando.

- **Modelos de datos**

    - **Data Warehouse**: almacenaje de datos a lo “bruto”

    - **Data Mart**: almacenaje de datos más específicos

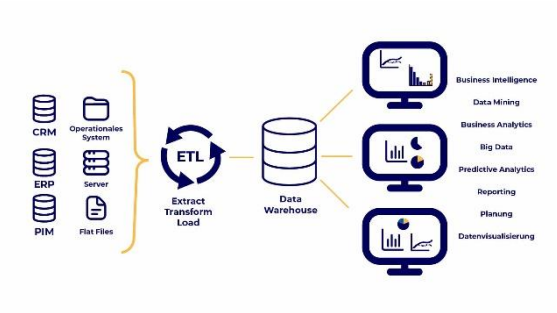


    - Ej.: Data Warehouse = DM de Ventas + DM de Recursos Humanos + DM de Producción

- Estructura **OTLP**: Bases de datos transaccionales (las habituales); rápido de procesar, pero lento para analizar.

- Estructura **OLAP** (también denominada *Cubo OLAP*): análisis multidimensional de datos de forma veloz e interactiva. No está optimizado para transacciones; implican cargas pesadas en procesos. Formatos Estrella y Copo de Nieve

- **ETL**: extracción (datos a lo "bruto"), transformación y carga. Datos que no están optimizados. Existen procesos ELT: "más baratos" a priori.

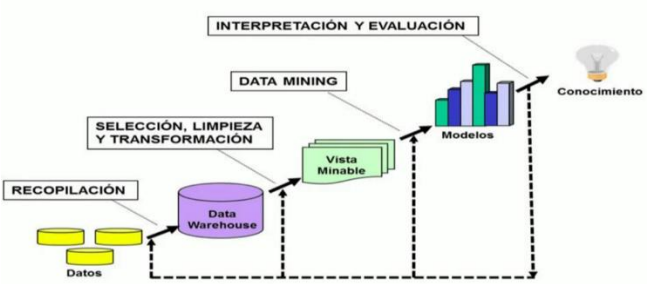


- Las herramientas de BI nacen para dotarnos de mayor flexibilidad y homogeneidad ante la manera antigua de construir esos informes y tomar decisiones.

- **Tipos de salidas** en BI:
    - **DSS**: son los llamados 'Sistemas de soporte a la decisión'. Comprenden informes dinámicos y no requieren conocimientos técnicos. La información está dirigida y adecuada a cada perfil.
    - **EIS**: son los llamados 'Sistemas de información ejecutiva'. Ofrecen indicadores de negocio o KPI y permiten análisis de expectativas y por supuesto, apoyan la toma de decisiones empresariales.
    - **CMI**: también llamados "Cuadro de mando integrales". Orientados a la toma de decisiones por altos puestos directivos y agrupan todos los departamentos de la compañía.

- **Herramientas** BI: *Power BI - Microsoft* (líder), Tableau (2ª), Qlik.

- Google ha comprado *Looker* y lo ha integrado (*Google Cloud Platform*), con vistas a un futuro crecimiento potencial. Quejas por problemas y fallos en *Looker* (lo hemos sufrido en prácticas).
- Minería de datos: es el proceso final para interpretar y evaluar (ver módulo 5). Dos ramas:
    - **Estadística clásica**: se utiliza principalmente con un fin puramente predictivo y para ello podemos hacer uso de: árboles de decisión, *clustering*, análisis de regresión, etc.
    - **Moderna**: está basada en inteligencia artificial y aprendizaje automático (machine learning) y además de predecir, se usa para descubrir conocimiento. *Redes neuronales*, Agrupamiento *k-means*



EJERCICIO

E.2 Descarga el dataset de Kaggle 'Netflix Movies and TV Shows'. Cárgalo en Data Studio e intenta responder a las preguntas que plantea.

- Utilizar archivos CSV:
    - utilizar previamente una hoja de cálculo de Google para unificar formatos
    - No todos los archivos CSV están bien configurados
    - En tipo de archivo, tiene que poner Google Sheet.
- Fuente de datos: https://www.kaggle.com/ Buscar datos con buena valoración. Utilizaremos para la práctica la base de datos de Netflix (https://www.kaggle.com/datasets/shivamb/netflix-shows)
- Conectar con https://lookerstudio.google.com/datasources
- Crear un informe genérico en Looker Studio
- Trastear añadiendo gráficos y controles

Conexión entre tablas (Google Studio)

- Ver video
- Se pueden conectar hasta 5
- Revisar conceptos SQL: LEFT JOIN, RIGHT JOIN, INNER JOIN (intersecciones) https://programacionymas.com/blog/como-funciona-inner-left-right-full-join

Ejercicio: Big Data en Restauración (corrección al día siguiente).

- Punto de partida
- Objetivos
- Recopilación de datos
- Análisis de datos
- KPIs
- Buyer Persona
- Estrategias de Marketing
- Soluciones

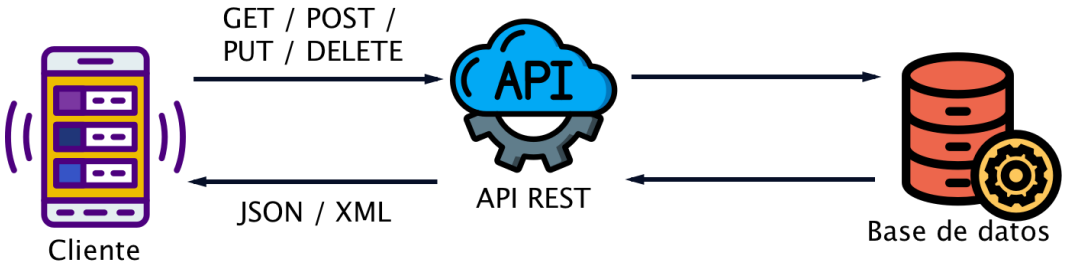
DIA 4 – Tecnología II: Manipulación de datos.

- Documentación: la documentación de un programa puede ser **interna y externa**.
    - La documentación **interna** es la contenida en líneas de comentarios.
    - La documentación **externa** incluye análisis, diagramas de flujo y/o pseudocódigos, manuales de usuario con instrucciones para ejecutar el programa y para interpretar los resultados.
- Fundamentos de programación: tipos de datos, booleanos, variables, funciones, condicionales, bucles.
- Expresiones regulares: https://regexr.com/ https://regex101.com/

- **JavaScript Object Notation (JSON):**
    - es un formato basado en texto estándar para representar datos estructurados en la sintaxis de objetos de JavaScript.
    - https://www.json.org/json-es.html

- Utilizado para transmitir datos en aplicaciones web: enviar algunos datos desde el servidor al cliente, así estos datos pueden ser mostrados en páginas web, o viceversa.
- Es muy común encontrarse en herramientas de analítica de datos de páginas webs y apps objetos JSON que lanzan **eventos** en cada acción que se quiera recoger, los cuales llevan incorporados muchos valores, llaves o parámetros adicionales a los que podemos acceder y extraer fácilmente para nuestro objetivo final.

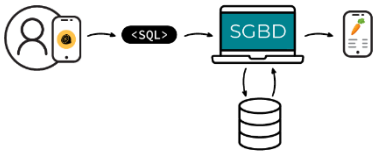
- Una **API**, o interfaz de programación de aplicaciones:
  - Conjunto de reglas que definen cómo pueden las aplicaciones o los dispositivos conectarse y comunicarse -entre sí.
  - **API REST**: cumple los principios de diseño del estilo de arquitectura REST o transferencia de estado representacional. Por este motivo, las API REST a veces se conocen como API RESTful.
  - Capa intermedia entre datos (BD) y las aplicaciones web finales (cliente).



- https://jsfiddle.net/ Test your JavaScript, CSS, HTML or CoffeeScript online with JSFiddle code editor.

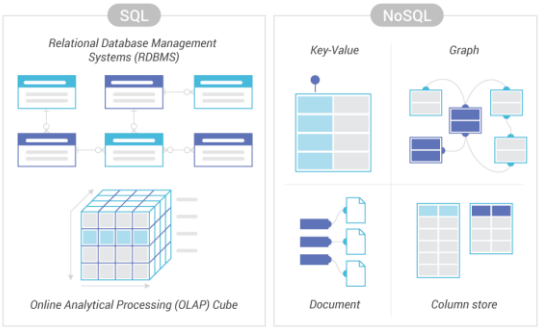
- Ejercicio *json* "juego de tronos"

Bases de datos - SQL



- La gestión de las bases de datos es fundamental para todos los trabajos de estas áreas.
- Un sistema de gestión de bases de datos (**SGBD**, por sus siglas en inglés) o D**ataBase Management System (DBMS)** es una colección de software muy específico, orientado al manejo de base de datos, cuya función es servir de interfaz entre la base de datos, el usuario y las distintas aplicaciones utilizadas.
- Su uso permite realizar un mejor control a los administradores de sistemas y, por otro lado, también obtener mejores resultados a la hora de realizar consultas que ayuden a la gestión empresarial mediante la generación de la tan perseguida ventaja competitiva.
- Permite manejar los accesos diferenciados (identificación, seguridad) y permite interpretar las búsquedas para ingresar, modificar, invertir o suprimir datos.
- Se pueden diferenciar 2 grandes familias de **SGBD**: los SGBD SQL y los SGBD NoSQL.
- **SQL: Relacionales.**

    - Esquema fijo y datos clasificados.
    - Tipo y validez de datos muy importante
    - Necesidad recurrente de escritura y modificaciones de datos sobre elementos específicos (SQL permite modificar fácilmente lineas especificas)
    - Necesidad de búsquedas complejas



- **NoSQL: <u>No relacionales</u>. <u>Modulares</u>.**
    - No necesitan esquema fijo.
    - Necesidad de múltiples búsquedas de lectura.
    - Grandes conjuntos de datos (Big Data)
    - Datos distribuidos (varias fuentes)
- En programación solemos usar el término **CRUD** para referirnos a las operaciones básicas que puedes realizar sobre un conjunto de datos y por sus siglas son:
    - Crearlos: nuevos registros, insertar información.
    - Leerlos (Read): consultar esa información (un registro o una colección de estos registros).
    - Actualizarlos (Update): tomar un registro que ya existe en la base de datos y modificar alguna de las columnas.
    - Eliminarlos (Delete): tomar un registro y quitarlo del almacén.
- **Claves (Keys):**
    - **Primaria o principal - PRIMARY KEY**
        - Identifica de forma única cada registro en una tabla.
        - Deben contener valores únicos y no pueden contener valores NULL.
        - Una tabla solo puede tener una clave principal, que puede consistir en campos simples o múltiples.
        - Una tabla NO tiene porqué tener una clave primaria, si bien es aconsejable.
    - Foránea - **FOREIGN KEY**
        - <u>Clave (campo de una columna) que sirve para relacionar dos tablas</u>.
        - El campo **FOREIGN KEY** se relaciona o vincula con la **PRIMARY KEY** de otra tabla.
        - La tabla secundaria es la que contiene la **FOREIGN KEY** y la tabla principal contiene la **PRIMARY KEY.**
        - La **FOREIGN KEY** es una <u>restricción que no permite que se agreguen o inserten datos que no válidos en la columna de *foreign key*</u>, ya que los valores que se van a insertar deben ser valores que se encuentren o ya estén en la tabla con la que se quiere relacionar.
- **Sentencias SQL:**
    - Create table
    - Operaciones: Insert, Select, Update y Delete
    - Condicionales: WHERE
        - Operadores lógicos: AND, OR, =, ¡=
        - Between

        - Like
    - CASE – WHEN – ELSE – END
    - Tablas cruzadas: LEFT JOIN, RIGHT JOIN, INNER JOIN,
    - Campos calculados
    - Funciones

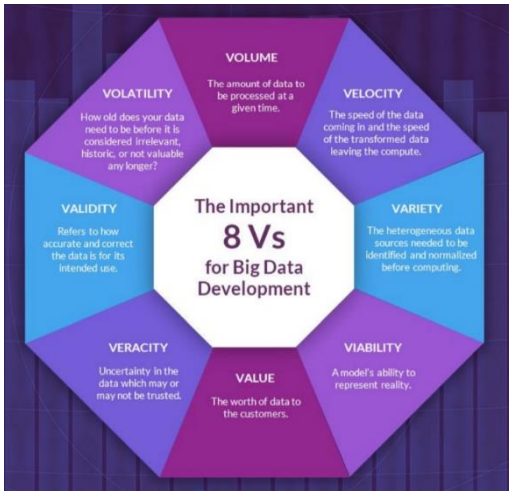
EJERCICIOS

1- Instalación de MySQL y creación de dos tablas: Usuarios y Productos.
2- Ejecutar al menos dos cruzados de tablas para asignar los productos creados por cada uno de los usuarios teniendo en cuenta el ID (id de usuario como foreign key en producto)
    a. JOIN : https://www.tutorialesprogramacionya.com/mysqlya/temarios/descripcion.php?cod=58&punto=64&inicio
3- Mejora el dashboard que hiciste ayer con expresiones regulares y campos calculados en Looker Studio

Looker Studio. Ejercicios

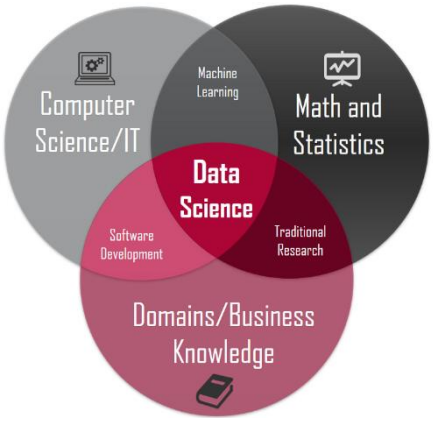
- Funciones https://support.google.com/looker-studio/table/6379764?hl=es
- Expresiones regulares https://support.google.com/looker-studio/answer/10496674?hl=es&ref_topic=7570421
- Ejercicio: regiones del mundo
    - Añadir campo/dimensión en editor de fórmulas y fórmula CASE
    - Añadir métrica en editor de fórmulas y fórmula SUM

DIA 5 – Big Data, IA y Machine Learning

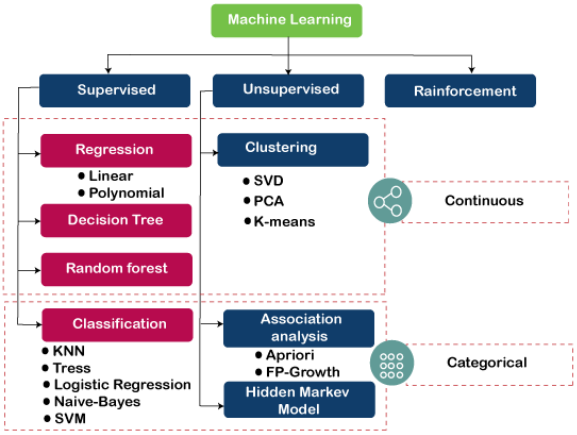


- El **Big Data** no es más que un campo dentro de todo lo que conocemos como ciencias de la computación o computer science.
- **3 elementos básicos** que conforman el Big Data:
    - Velocidad a la que se consume la información.
    - Variedad de información.
    - Volumen de información.
    - Pero todo esto, no nos sirve de nada si no incorporamos un componente de: VALOR. Si los datos no sirven para aprender, descubrir o analizar, todo lo anterior no sirve de nada.
    - Viabilidad, Veracidad, Validez, Volatilidad.

- Es importante no caer en el *hype* del marketing y la innovación asociado al término de Big Data.

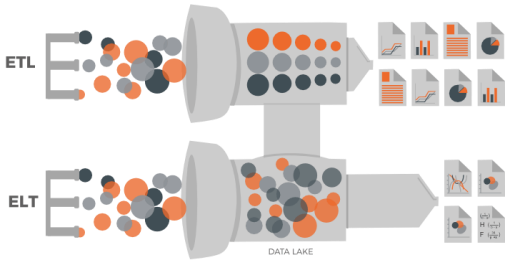


- **Data Science: AI y Machine Learning**
- Revisar las diferencias entre Big Data, IA, Machine Learning y Deep Learning (repaso de unidades 1 y 2)
- Cuando hablamos de **ML**, lo que aporta valor e importancia a su existencia y desarrollo es su capacidad para poder aplicar modelos predictivos → revolución de esta disciplina, enmarcada dentro de la amplia **IA.**
- Ejemplos de aplicaciones y de su crecimiento: Tik Tok, LinkedIn, etc.
- Ejemplos de publicidad programática: la compra programática nos permite el uso de data de los usuarios (histórico y en tiempo real) y nos capacita para abordar la personalización del mensaje publicitario. Esto se traduce en dos grandes ventajas: Posibilidad de segmentar de manera muy precisa el público objetivo al que queremos impactar.
- El **ML** basa toda su potencia en la aplicación de algoritmos para aplicar el aprendizaje automático y desarrollar así modelos predictivos. Algoritmos supervisados y no supervisados. Ejemplos.



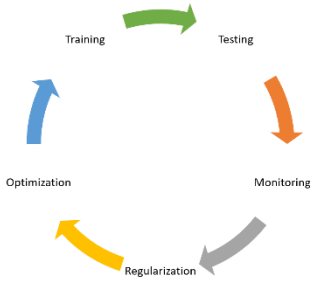
- **Aprendizaje supervisado:** modelos más utilizados dentro del ML. Para ello, necesitamos contar con unas entradas y salidas de la información sobre la cual vamos a aplicar el aprendizaje automático.
  - **Clasificación**: elige entre una lista de opciones previamente definidas y limitada.
  - **Regresión**: predecir números reales o números con infinitas posibilidades.
- **Aprendizaje NO supervisado**: los algoritmos se aplican para aprender de datos con elementos no etiquetados buscando patrones o relaciones entre ellos. En este caso no necesitarían delimitar el número de entradas y salidas.
  - **Clustering**: clasifica en grupos los datos de salida. Concepto de *centroide*
  - **Asociación**: descubre reglas dentro del conjunto de datos.
- **ELT (Extraer, cargar, transformar)**: método diferente de acercarse al flujo de datos, en el que los datos extraídos se cargan primero en el sistema de destino.

- Las transformaciones se realizan después de que carguemos los datos en e l almacén de datos.
- Los datos primero se copian en el *data lake* y luego se transforman in situ.
- Funciona bien cuando el sistema objetivo es lo suficientemente potente como para manejar transformaciones a gran escala.



- ELT generalmente se usa con bases de datos NOSQL, un dispositivo de datos o una instalación en la nube.

- **Data Lake**: para almacenar datos de forma no estructurada (ej: texto, imágenes, artículos, correos, etc.). Repositorio central.
    - Ventajas: más flexibles, no diseñados de antemano.
    - Ejemplo: Amazon S3
- ¿Ventajas e inconvenientes Data Lake VS Data Warehouse? ¿ETL VS ELT?
    - Depende de coste, pero cada vez menos (cruzar con tamaño de la empresa)
    - Depende del tipo y cantidad de datos.
    - Depende del esfuerzo que tenga que hacer la empresa para transformar esos datos y darles valor.
    - En un ELT se depende mucho de la tecnología que se utilice
- ML: Entendiendo el proceso:
    - Hacerse las preguntas adecuadas.
    - Identificar los datos y prepararlos.
    - Aplicación del algoritmo correcto.
    - Evaluación y ajuste de modelo.
    - Uso y presentación de modelo.



- Video aprendizaje supervisado VS no supervisado (VIDEO).
    - Concepto de *espacios latentes*.
    - El NO supervisado señala un camino muy prometedor
    - Cajas negras: algoritmos con matemáticas, estadísticas, etc.
- Videos de Redes Neuronales. Explicación de funcionamiento con el ejemplo de gafas VR + nachos

EJERCICIOS

Ejercicio 2: con la ayuda de la extensión 'Data Miner', haz un scraping y descarga los datos de las casas en venta en tu ciudad existentes en la web pisos.com. Ahora, haz una regresión múltiple en una hoja de Excel para predecir un precio en base a m2 y/o número de habitaciones.

- Uso de Data Miner en web www.pisos.com

- Exportar/grabar a Excel
- Realizar gráficos de dispersión

Recursos de aprendizaje
https://www.codecademy.com/learn/paths/bi-data-analyst Cursos
https://8weeksqlchallenge.com/ Ejercicios SQL
(respuestas en:
https://github.com/bcamandone/Data_Analysis_SQL/tree/main/8%20Week%20SQL%20Challenge)

VIDEO ejemplo Red Neuronal:
https://colab.research.google.com/drive/1QH7yhAmklHxBRi1d-dZcI3Y8uN5WNbnF?usp=sharing

EXAMEN DE VIERNES