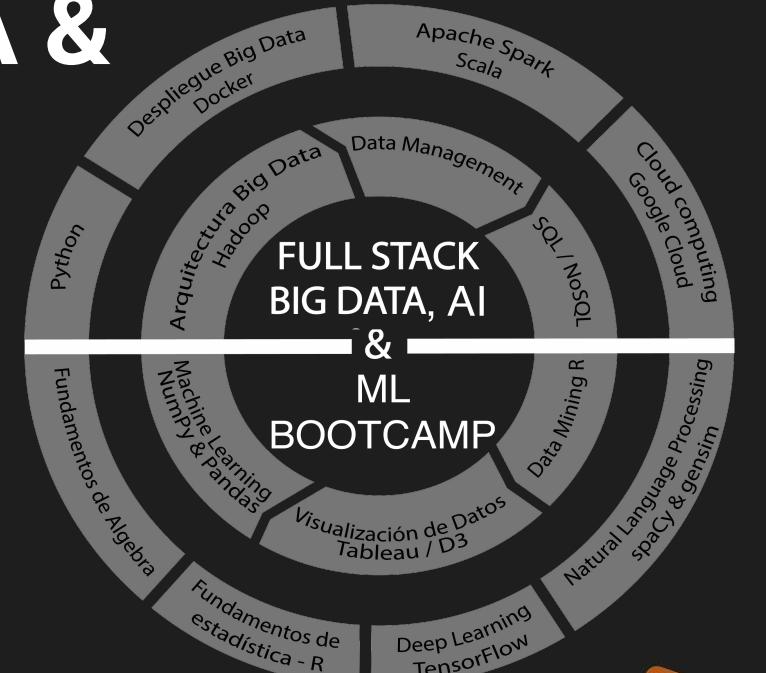


Full Stack Big Data, IA & Machine Learning Bootcamp

Big Data Tour



Motivación



- El crecimiento anual en perfiles Big Data está en torno al 15%. Se estima que en los próximos años superará el 45%
- Hay un problema serio: la falta de personal cualificado
- Todo gira cada día más en torno al Big Data:
 - Lo “conocido”: Inteligencia de Negocio como la conocemos
 - La inteligencia artificial
 - Traducción de documentos
 - Reconocimiento facial
 - Coches autónomos
 - Detección de contenidos inadecuados
 - Asistentes personales
- Hemos pasado de ser unos frikis raros a unos frikis sexys
- La Inteligencia Artificial ha explotado

Nuestra fuerza



"Without data
you're just another person
with an opinion."

W. Edwards Deming

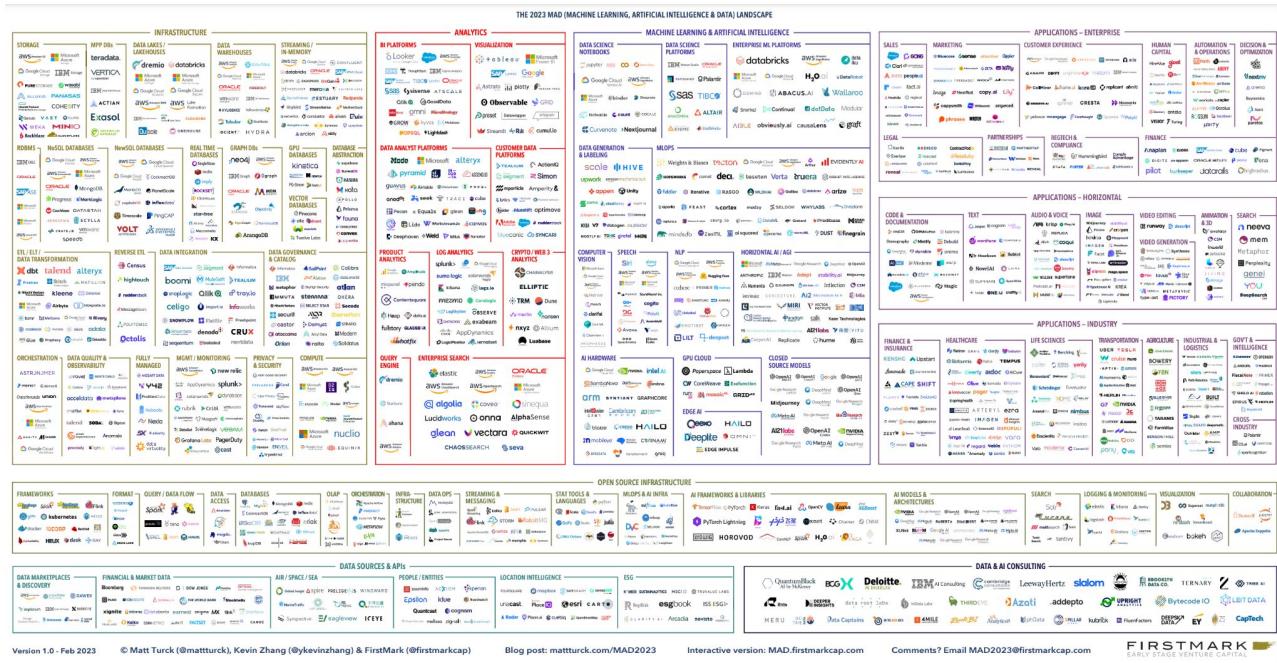
Buzzwords

Confusión

Expresiones de moda

Solape

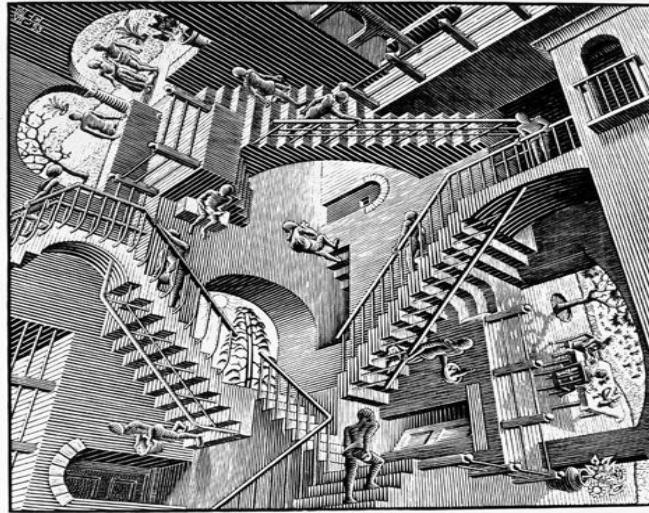
¿Qué ocurre cuando uno se decide a estudiar “Big Data”?



<https://mattturck.com/landscape/mad2023.pdf>

Me perdí...

Ya lo decía Escher...



Estás más perdido que un muelle cayendo por las escaleras de Hogwarts



keepcoding.io



cursos@keepcoding.io



KeepCoding®. Todos los derechos reservados

¿Qué pretendemos conseguir?

**NINGUNA META ES DIFÍCIL
SI LA DIVIDIMOS EN
OBJETIVOS PEQUEÑOS
Y ALCANZABLES.**





Algunas reflexiones antes de comenzar...



¿Utilizáis Amazon, Spotify o Netflix?



...Entonces os estáis beneficiando del Big Data

- Nuestra **huella digital** es cada vez más importante.
 - Separación entre la huella digital profesional y personal
 - Perfil digital coherente - Borrado de trazas antiguas
- Es importante entender que el big data plantea una **oportunidad** para tu vida digital, y por otro, vale la pena conocer qué **medidas inteligentes** podemos tomar para lograr que el big data trabaje en nuestro beneficio a la par que **protejamos nuestra identidad**.
- Precio de nuestros datos: ¿Deberíamos cobrarlos?



Cuando uno piensa que conoce la necesidad de sus usuarios...



Big Data en el mundo multimedia

- Disney lanza los trailers de Star Wars analizando la expectación en redes sociales
- Se usa en la investigación para averiguar las preferencias de los clientes potenciales sobre argumentos, castings, etc. de series de televisión y películas
- Netflix tiene etiquetadores (taggers) para la clasificación de sus series
- En el caso de “La Gran Muralla” se eligió a Matt Damon como personaje principal ya que su nombre superaba doblemente al de cualquier otro actor en las búsquedas por internet en China.
- Big Data: El “asesino” de los guionistas





¿Conocéis a una persona, dispositivo u organización que no genere datos?





¿Conocéis a una persona u organización que no utilice los datos?





**¿Conocéis a una persona u
organización que no aproveche
los datos que tiene a pesar de
usarlos?**





Contexto en el que surgimos



keepcoding.io



cursos@keepcoding.io

|

KeepCoding®. Todos los derechos reservados

Contexto

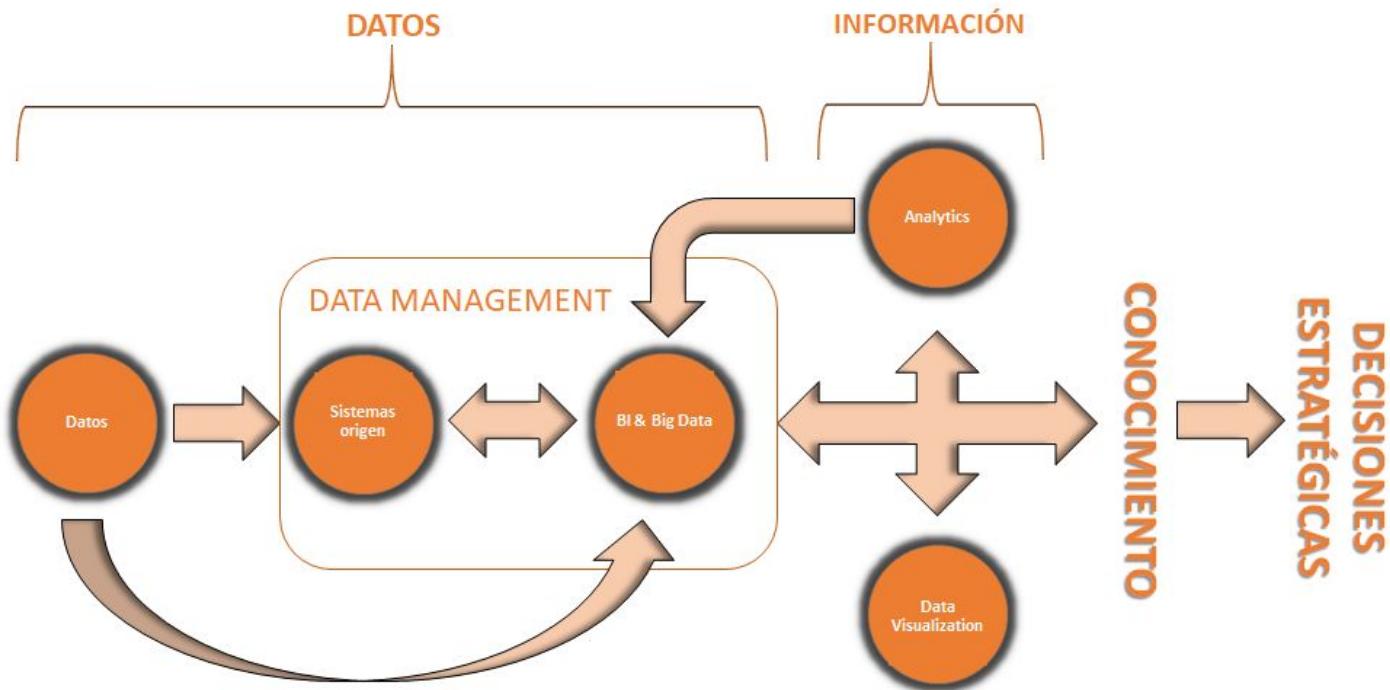
- “Tengo un montón de datos y necesito guardarlos para que después aporte valor en mi empresa”
- Hace años, los datos eran para Business Operation
- Creación del departamento de Inteligencia de Negocio para ayudar a la toma de decisiones
- “El cliente más valioso es aquel que nunca pierdes”



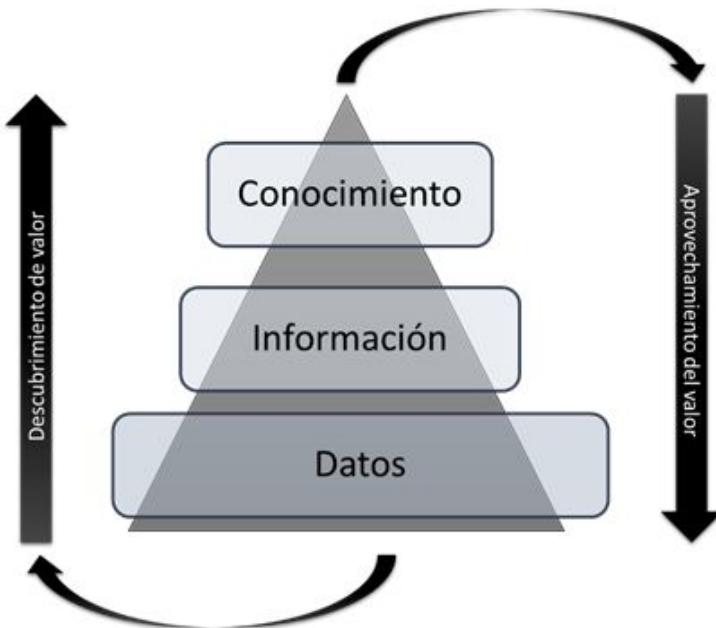


De los datos a las decisiones estratégicas

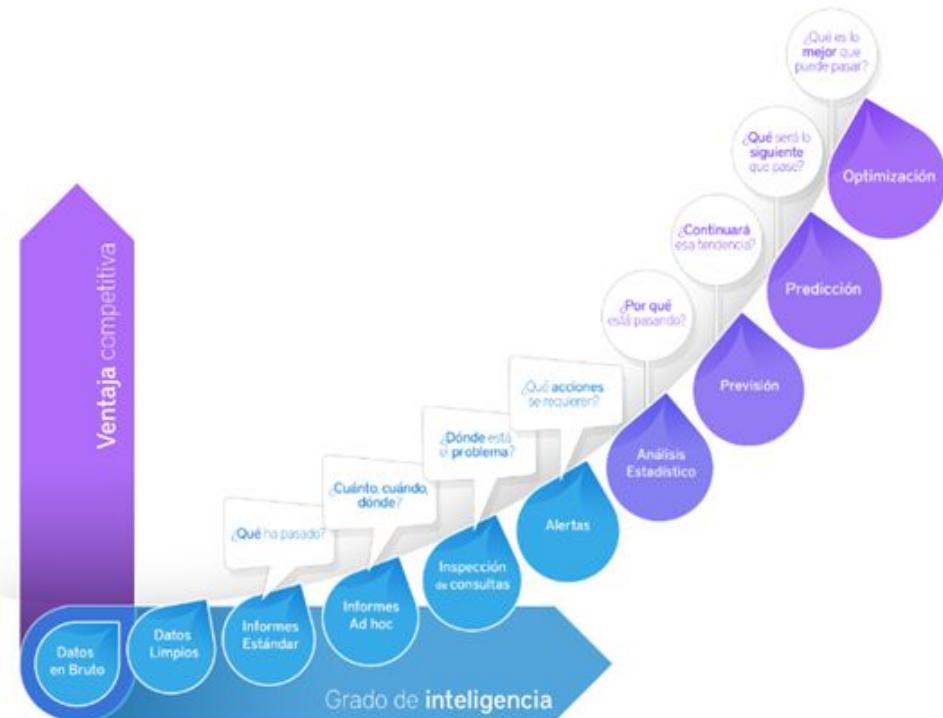
Ciclo de vida del dato



La pirámide del conocimiento



¿Cuál es el DIKW real que necesitamos?



Corporate Performance Management



La pirámide de la información



Organizaciones basadas en la estrategia (1/5)

PRINCIPIOS

- No se puede aplicar una estrategia que no se pueda describir.
- La actividad de la organización es algo más que la suma de las partes, necesita estar alineada con la estrategia.
- Hay que hacer que la estrategia sea el trabajo diario de todos los empleados. No se trata de dirigir desde arriba hacia abajo, sino de comunicar desde arriba hacia abajo.
- Hay que hacer de la estrategia un proceso continuo.
- Movilizar el cambio mediante el liderazgo de los directivos.

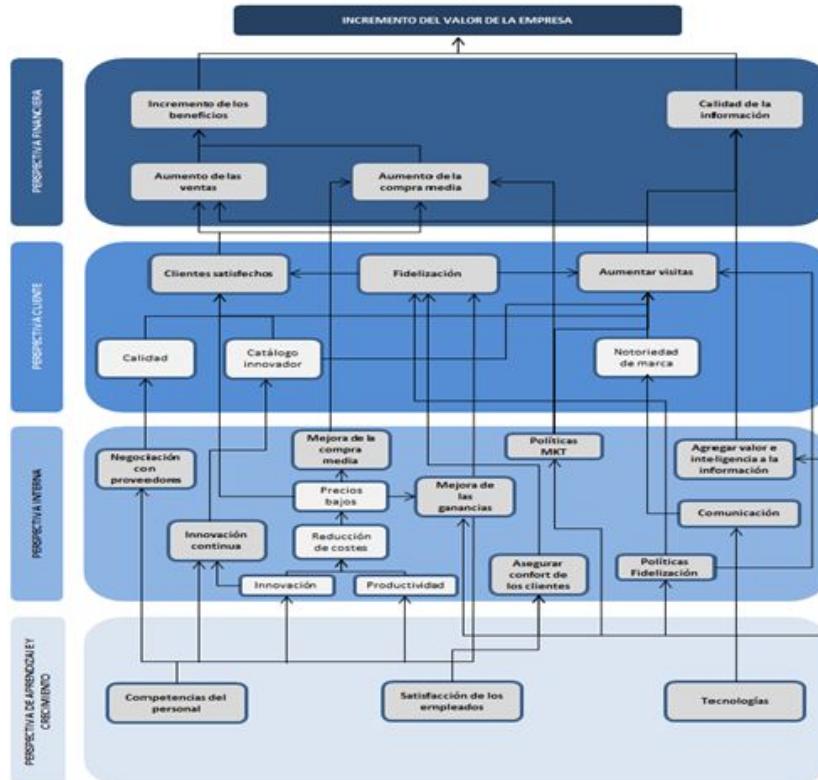
Organizaciones basadas en la estrategia (2/5)

CADENA DE CREACIÓN DE VALOR



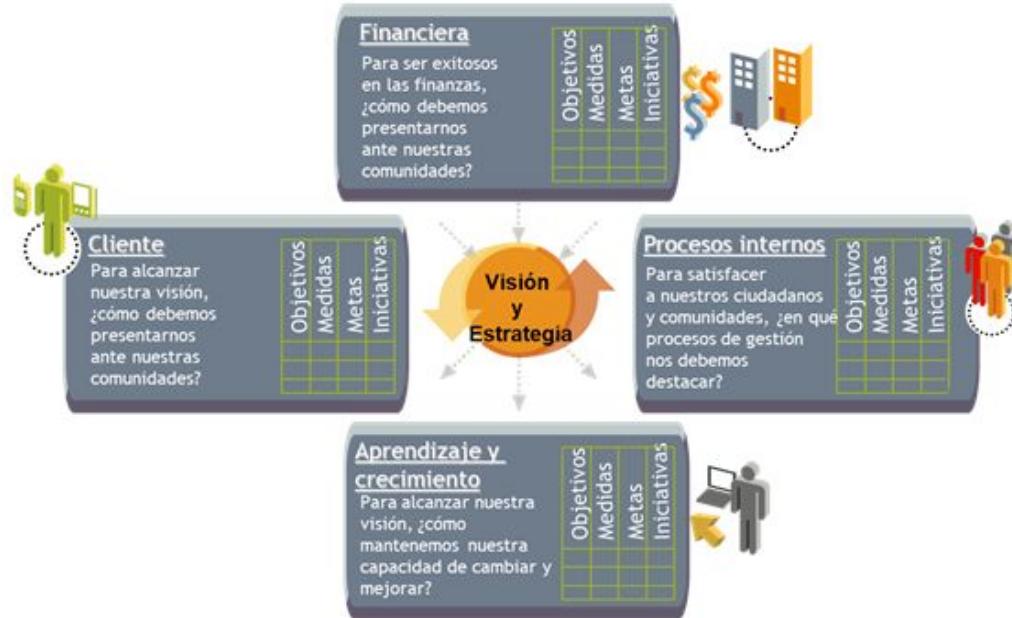
Organizaciones basadas en la estrategia (3/5)

MAPA ESTRATÉGICO



Organizaciones basadas en la estrategia (4/5)

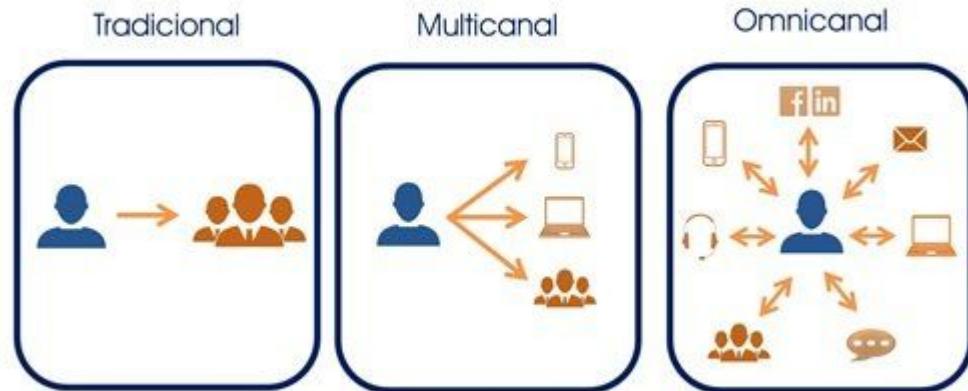
CUADRO DE MANDO INTEGRAL



Organizaciones basadas en la estrategia (5/5)

OMNICANALIDAD

- Conseguir una **experiencia de usuario de calidad** haciendo que la interacción del usuario sea perfecta en cualquiera de los canales, sin que note diferencias que hagan que un canal u otro lo considere mejor.
- **Responder sin demora** a la demanda del usuario en el momento que le surja, porque si no otro se nos adelantará.
- **Personalizar** la experiencia del cliente al máximo en función de sus intereses
- **Simplificar** al máximo los procesos para que al cliente le resulte fácil la interacción.



El cliente pasa a convertirse en el centro de todas las estrategias.
Los datos, su almacenamiento y su estudio cobran especial relevancia, puesto que es imprescindible para conocer al cliente y seguir sus interacciones





¿Empresas en la que soléis comprar que utilicen la estrategia de la omnicanalidad?





Datos, datos, datos...¿dónde están?

Lo más tradicional



Nivel I

Contaplus Elite 0001 - Empresa de pruebas: Sage Contabilidad - Ejercicio 2013

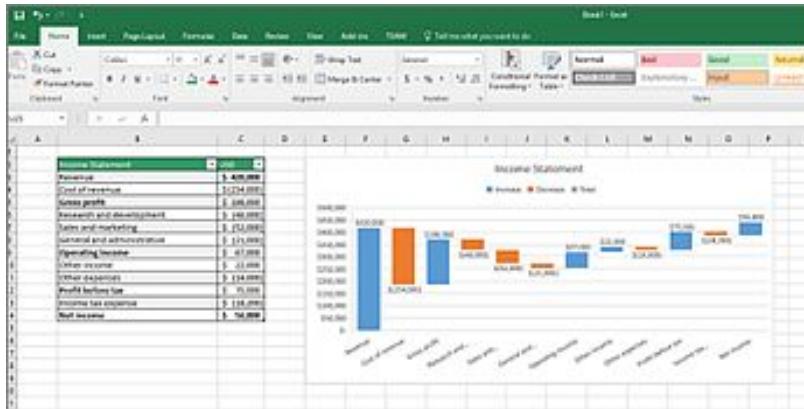
Archivo Nuevo... Imprimir Analíticos Inmovilizado Chequeos Informes Bebé eFacturas Util Ventana Ayuda

Empresa Plan contable Asientos Tesorería Presupuestos Analítica Segmentos Inmovilizado Informes Bebé eFactura Utilidades

Inmovilizado

SALDOS DE SUBCUENTAS

	A	B	C	D	E	F	G	H
1 SALDOS DE SUBCUENTAS	11300000	1290000	1730000	1750000	2010000	2060000	2130000	
2 Enero	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
3 Febrero	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
4 Marzo	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
5 Abril	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
6 Mayo	0,00	0,00	0,00	95.284,60	0,00	0,00	125.621,03	
7 Junio	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
8 Julio	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
9 Agosto	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
10 Septiembre	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
11 Octubre	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
12 Noviembre	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
13 Diciembre	0,00	0,00	0,00	0,00	0,00	0,00	0,00	



Bases de datos tradicionales



Microsoft®
SQL Server

ORACLE

ODBC

SQLite



PostgreSQL

MySQL®



TERADATA



SYBASE®

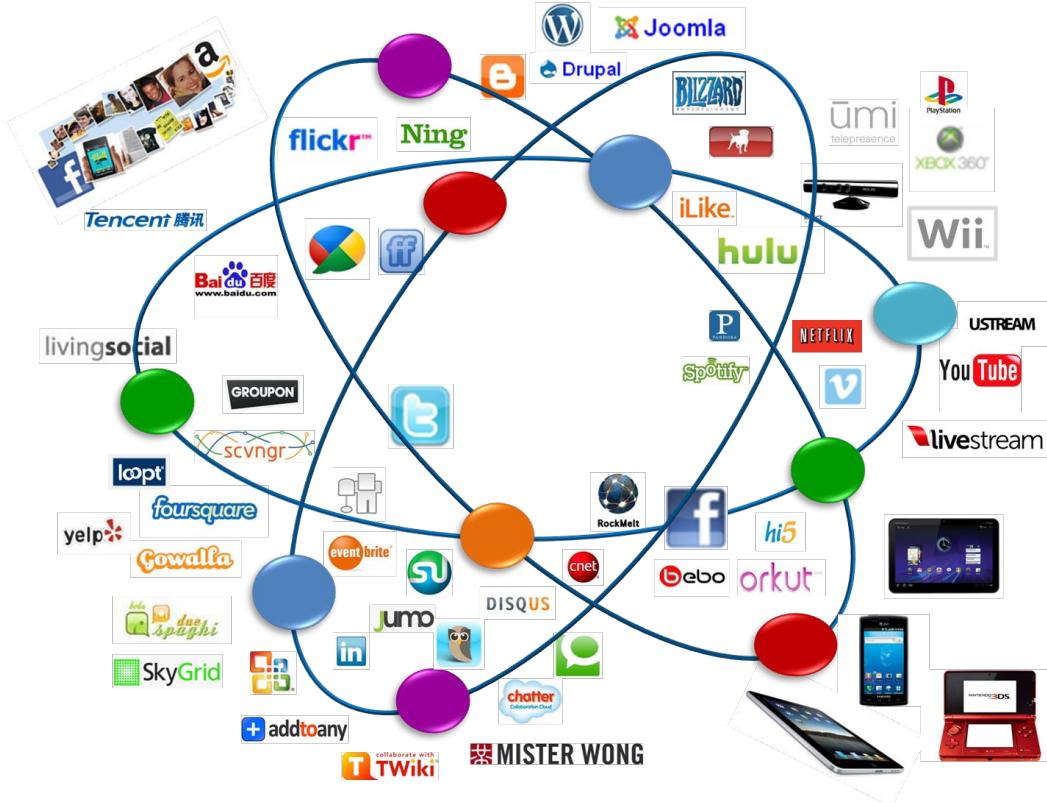
Informix®



Bases de datos NoSQL



Internet





Censo electoral

Sede electrónica

 Buscar

ES E



La vida de las mujeres
y los hombres
en Europa
Edición 2017



másINÉ
Revista digital



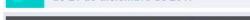
Explica
Estadísticas
territoriales



El IPC en un
clic



Apellidos y
nombres



¿Cuántos
habitantes...?

Elecciones al Parlamento de Cataluña
de 21 de diciembre de 2017



Acceso a todos nuestros vídeos e infografías

- Última hora
 - 23 Nov 17. Estadística del taxi
 - 23 Nov 17. Coyuntura Turística Hotelera. CTH
 - 21 Nov 17. Entrada de Pedidos en la Industria. IEP
 - 21 Nov 17. Índices de Cifras de Negocios en la Industria. ICN
 - 21 Nov 17. Indicadores de actividad del sector servicios. IASS
 - 20 Nov 17. Mujeres y hombres en España

Más noticias

Indicador	Periodo	Valor	Variación (%)
IPC	2017M10	102.668	1,6
EPA. Ocupados (miles)	2017T3	19.049,2	2,82
EPA. Tasa de paro	1 2017T3	16,38	-2,53
PIB	24 2017T3	—	3,1
Población total (miles)	3 2017	46.528,9	0,19

1. Valor en %. Variación: diferencia respecto a la tasa del mismo periodo del año anterior.

2. Ofrece el número encuestado, ref. 2010. Datos corregidos de efectos estacionales y de calendario.

3. Cifras de población a 1 de enero de 2017. Datos provisionales.

4. Datos avance

Aquí se muestra un gráfico en formato Flash

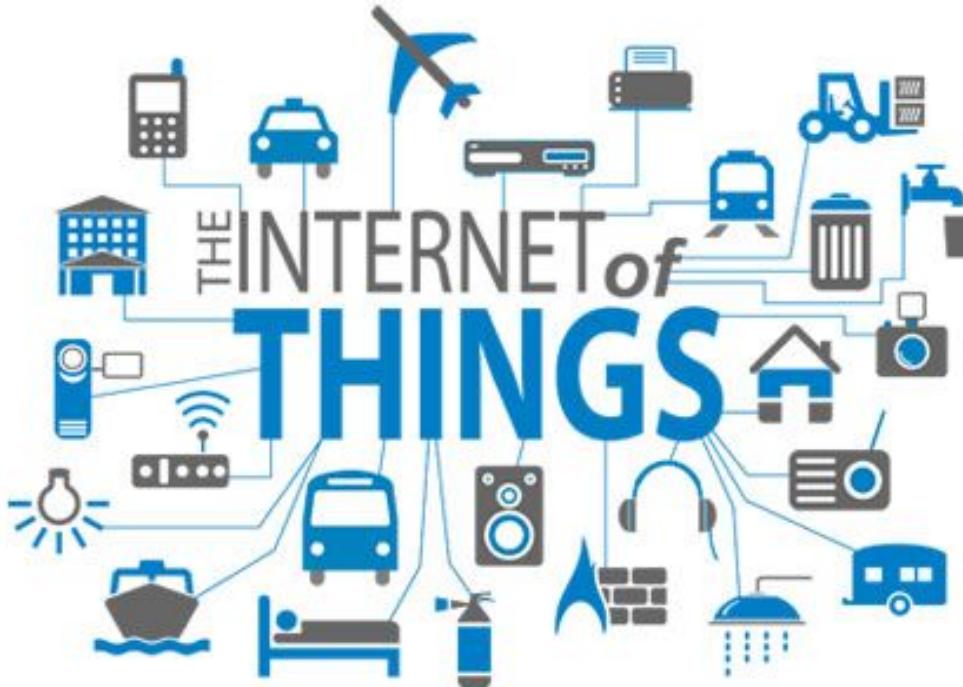
1. Para poder visualizarlo, se debe utilizar [este Visor Flash](#)
2. A continuación, en la ventana emergente que aparecerá, pulsar sobre el botón "Permitir".
3. La página se volverá a cargar mostrando el contenido en ese formato.

Otros datos interesantes



Protección de la identidad

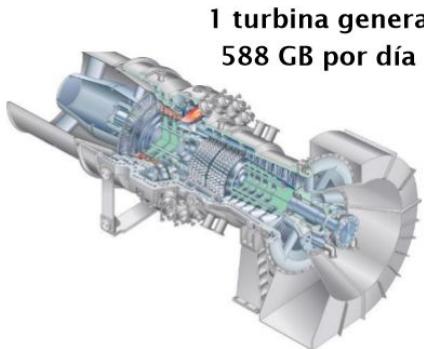




Los datos que no están en Internet

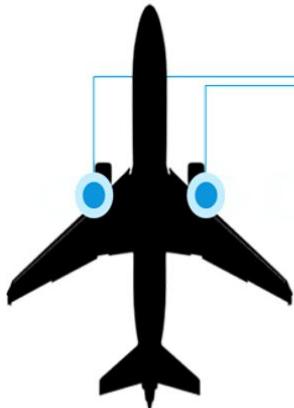


270 millones de usuarios generan 100 GB por día



1 turbina genera 588 GB por día

Sensor data from a cross-country flight



20 TB × 2 × 6 × 28,537 × 365

20 terabytes of information per engine every hour

twin-engine Boeing 737

six-hour, cross-country flight from New York to Los Angeles

of commercial flights in the sky in the United States on any given day.

days in a year

= **2,499,841,200 TB
(2,5 ZB al año)**

Fuente: HP



keepcoding.io



cursos@keepcoding.io



KeepCoding®. Todos los derechos reservados

Resumiendo

- Bases de datos propias
 - Internet
 - Máquinas y sensores
 - Dispositivos móviles
 - Administraciones públicas
 - Asociaciones y organizaciones
 - Open Data
 - ...
- ¿Alguno más?



¿Orígenes de datos "singulares"?



La saliva!



¿Por qué me vigilan si yo no soy nadie?



<https://youtu.be/NPE7i8wuupk>



Data Management



keepcoding.io



cursos@keepcoding.io

|

KeepCoding®. Todos los derechos reservados

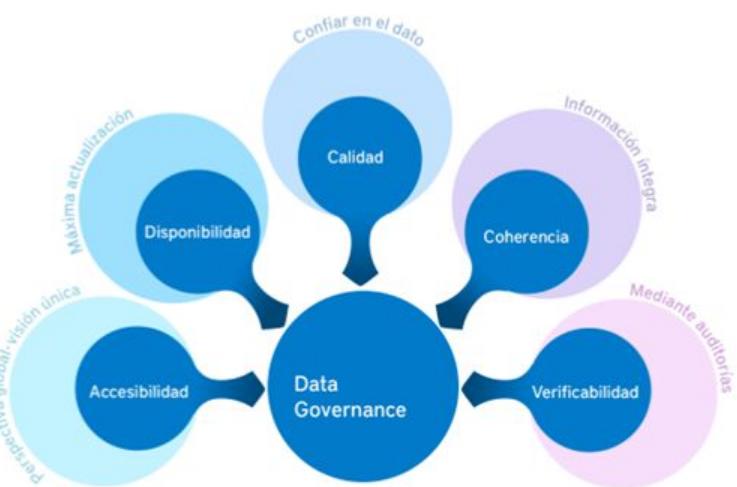
Data Management



La gestión de los datos, que engloba el conjunto de todas las disciplinas relacionadas con gestionar los datos como un activo valioso, es lo que se conoce como **Data Management**.

El fin del Data Management es tener una visión unificada de todos los datos de la empresa así como de su entorno de negocio.

Data Governance



Cada vez disponemos de más información en las organizaciones, pero la manera de gestionarla no ha evolucionado al mismo ritmo en la mayoría de las ocasiones, lo que provoca que las organizaciones muchas veces **no saben ni qué datos tienen ni dónde encontrarlos**.

Por tanto, se convierte en algo fundamental a día de hoy el ser capaz de obtener la **trazabilidad** de cualquier dato para saber qué caminos sigue la información en su flujo de negocio. De esto se hace cargo la disciplina del **Data Governance**.



Data Architecture



Los datos hay que **almacenarlos**, pero no de cualquier manera, sino siguiendo unas **reglas específicas**. De ello se encarga el **Data Architecture**, que describe los procesos, sistemas y organización humana necesarios para almacenar, acceder, mover y organizar los datos.



Data Modeling & Design



Una vez establecidas las reglas de cómo almacenar los datos, el siguiente paso es llevarlas a cabo.

De esto se encarga la especialidad de **Data Modeling & Design**, que modela y diseña las bases de datos, se encarga de su **implementación y soporte**, de manera que los datos puedan ser usados como recursos.

Data Storage



Un punto importante que hay que tener en cuenta en todo lo relacionado al almacenamiento es que hay que **evitar** que sea cada aplicación la que **decida** cómo guardar los datos y para ello el **Data Storage** tiene que ser el que vele por controlar cómo, cuándo y qué se almacena en cada uno de los sistemas.

Data Security



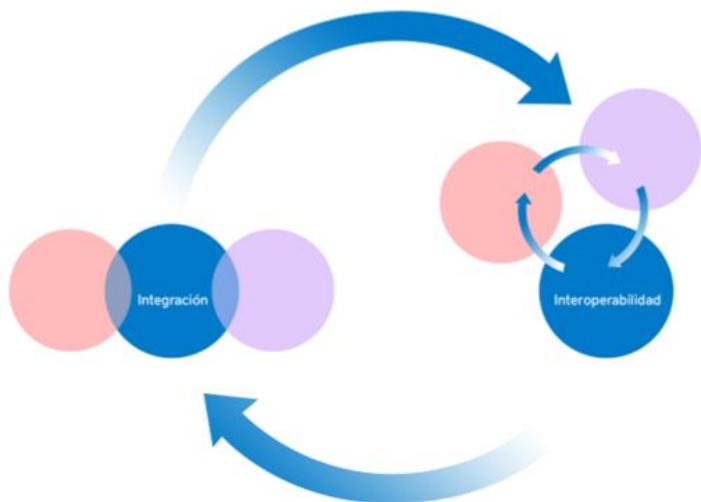
En los tiempos que corren no podemos olvidarnos tampoco del tema de la seguridad.

El módulo de **Data Security** se encarga de la **protección** de los **datos** contra el acceso, visualización, modificación o eliminación no autorizada, ya sea accidental, intencionada o maliciosa.

Se llevan a cabo todos los mecanismos, acciones y políticas necesarias para **garantizar** la **seguridad** de los datos en el entorno de la empresa.



Data Integration & Interoperability



En una organización **coexisten muchas aplicaciones y sistemas**, pero si no existe una comunicación entre todas ellas se está perdiendo eficacia y visión, lo que conlleva que la capacidad de decisión se vea disminuida.

La **integración** implica entender cómo la información se guarda en los distintos sistemas para que, al **interactuar** entre ellos, no sólo se conecten, sino que se entienda.

Master Data



Los **datos maestros** son utilizados en varios procesos de la organización, por lo que es de vital importancia **estandarizarlos** en los **diferentes sistemas** donde se almacenan.

El negocio necesita tener una visión de 360º sobre estos datos para tomar decisiones alineadas a la visión de la empresa.

Meta-Data



El módulo de **Meta-Data** es el que se encarga de integrar, controlar y proporcionar metadatos.

Los **metadatos** describen, etiquetan y caracterizan los datos a los que se refieren, haciendo más **fácil** su **interpretación** y **utilización**.

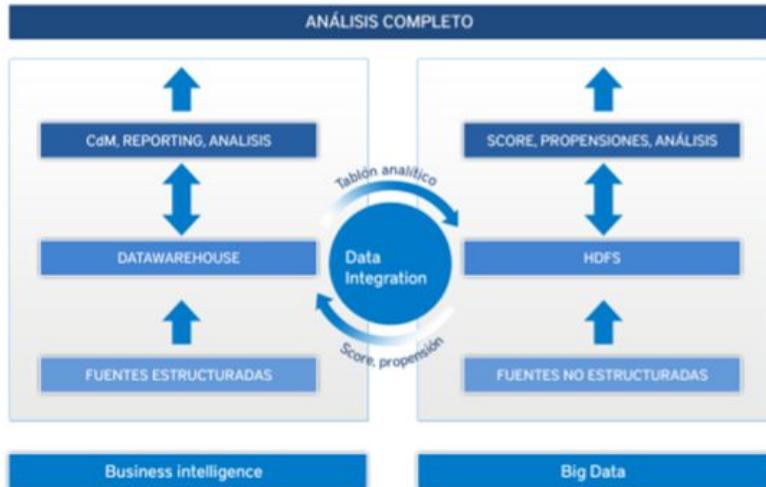
Data Quality



De gran importancia es la disciplina de **Data Quality**, que se encarga de definir, controlar y mejorar la calidad de los datos.

Es necesario que los **datos** sean de calidad, esto es, que sean **fiables, precisos, consistentes** y que proporcionen una **visión única**.

Business Intelligence & Big Data



La disciplina del **BI & Big Data** tiene como objetivo proporcionar una **visión integrada** de la información de la empresa para dar **apoyo** a la **toma de decisiones**.

Se ocupa de los datos tanto históricos como actuales y de las analíticas oportunas.





BBDD relacionales

Definición de base de datos

- Una BBDD es una colección electrónica de información diseñada para satisfacer unas determinadas necesidades:
 - > Pueden contener información de diversas fuentes
 - > Proporcionan mecanismos para extraer rápidamente los datos
 - > Permiten compartir información entre los distintos departamentos de una empresa.
- Las BBDD son uno de los pilares de la informática.
- Las bases de datos están por todos lados a nivel informático.



Base de datos relacionales (1/2)

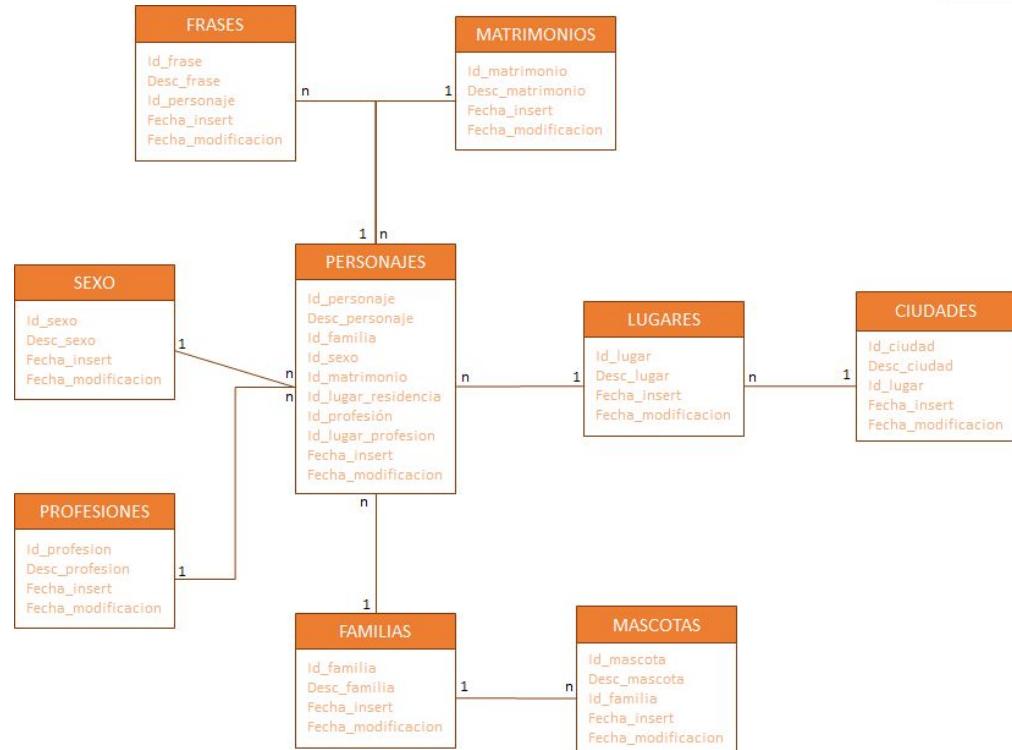
- En un principio existían dos modelos de bases de datos: el jerárquico y el de redes, de los cuales el de redes era el que más se adaptaba a las necesidades reales de almacenamiento y clasificación.
- Posteriormente, aparecieron las bases de datos relacionales
- Una BBDD relacional se compone de tablas, que a su vez están compuestas de campos, los cuales están formados por filas y columnas.

Base de datos relacionales (2/2)

- Las bases de datos relacionales tienen la ventaja de "relacionarse" entre sí sin la necesidad de duplicar una gran cantidad de información, basadas en un lenguaje estándar llamado SQL (Structured Query Language), el cual es, podríamos decir, la razón para que las bases de datos relacionales tengan un éxito tan arrollador.

Ejemplo

- **Id** → identificador del registro en la tabla
- **fecha_insert** → fecha en la que se insertó el registro
- **fecha_modificacion** → fecha en la que se ha modificado el registro
- **Desc** → descripción de un lugar, una ciudad, el nombre de un personaje, de una mascota, etc.



Claves primarias y foráneas

➤ Clave primaria (Primary Key)

- > Consiste en una o más columnas cuyos datos contenidos son utilizados para identificar de manera única cada fila en la tabla.
- > No puede haber duplicados.
- > No puede tener registros vacíos.
- > Sólo puede haber una por tabla.
- > Se almacena en un índice.

➤ Clave foránea (Foreign Key)

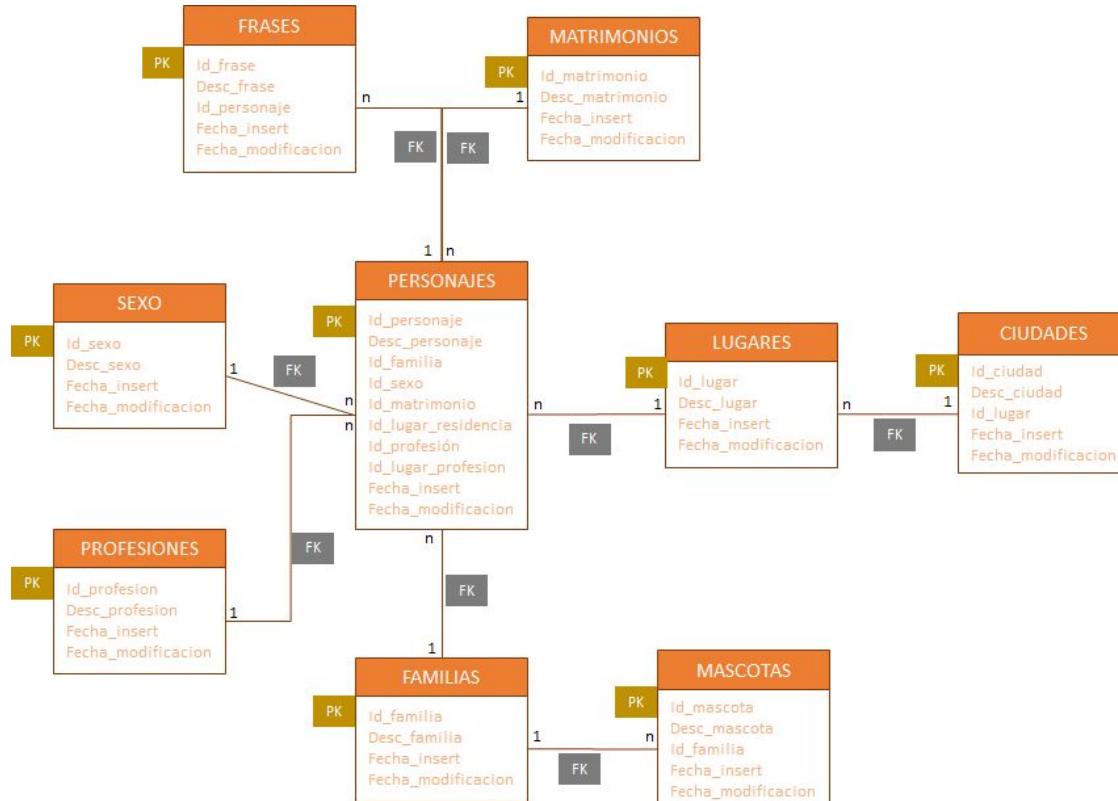
- > Es un grupo de una o más columnas en una tabla que referencian a la clave primaria de otra tabla.
- > Puede formar parte de la clave primaria
- > Puede contener duplicados
- > Puede tener registros vacíos.
- > En la misma tabla puede haber varias.



Formas normales

- NF1
 - > La tabla contiene una clave única
 - > No tiene datos repetidos
- NF2
 - > NF1
 - > Todo registro debe depender únicamente de la clave principal
 - > Las columnas pueden depender de otras tablas, pero de campos que sean clave en sus respectivas tablas
- NF3
 - > NF2
 - > No puede haber columnas que dependen de otras columnas que no sean clave principal

No era tan difícil :)



The Hello World Collection

<https://helloworldcollection.github.io/>

The Hello World Collection

"Hello World" is the first program one usually writes when learning a new programming language. Having first been mentioned in Brian Kernighan's [tutorial to the B programming language](#), it became widely known through Kernighan & Ritchie's 1978 book that introduced ["The C Programming Language"](#), where it read like this:

```
main() {
    printf("hello, world\n");
}
```



Since then, Hello World has been implemented in just about every programming language on the planet. This collection includes **567 Hello World programs** in as many more-or-less well known programming languages, plus **76 human languages**.

The programs in this collection are intended to be as minimal as possible in the respective language. They are meant to demonstrate how to output Hello World as simply as possible, not to show off language features. For a collection of programs that tell more about what programming in the languages actually is like, have a look at the [99 Bottles of Beer](#) collection.

The Hello World Collection, started in 1994, was compiled with help from [many people around the world](#). It is the biggest collection of Hello World programs on the Internet, and the only one collecting human languages as well. To contribute, send your program to info@helloworldcollection.de. Begin your contribution with a comment in the respective language. [Real](#) programming languages only please.

Click [here](#) for a list of all contributors and other sources.

Click [here](#) for related links.

Click [here](#) for brief history of the Hello World Collection.

Support the [KDE Education Project](#) with our exclusive [Hello World merchandise](#) — T-shirts, mugs and more!

Last update: Oct 21, 2017.





Business Intelligence



keepcoding.io



cursos@keepcoding.io

|

KeepCoding®. Todos los derechos reservados

Contexto en el que surge

- Entorno que cambia rápidamente
- Cada vez hay más datos y menos tiempo
- Hay que hacer las preguntas adecuadas
- Se encargan los sistemas de información
- Soporte a las transacciones
- Distribuyen la información
- Soporte a la toma de decisiones



Definición de Business Intelligence

El **Business Intelligence** es un término paraguas que abarca los procesos, las herramientas y las tecnologías para convertir datos en información, información en conocimiento y planes para conducir de forma eficaz las actividades de los negocios.

The DataWarehouse Institute



El Business Intelligence se encarga de...

- Analizar la información de forma continuada en el tiempo
- Explorar para comprender qué sucede
- Descubrir relaciones entre variables, tendencias, patrones.
- Almacenar información en tablas interrelacionadas en un DataWarehouse.
- Objeto de análisis y objetivo concreto.
- Comunicar los resultados y efectuar los cambios.



Principios en los que se basan los sistemas de información

- **Percibir** lo que está ocurriendo en la compañía
- **Recordar** lo que ya ocurrió
- **Aprender** de lo recordado
- **Actuar** en base a lo aprendido



Sistemas operacionales vs informacionales

	SISTEMA OPERACIONAL	SISTEMA INFORMATICO
Fuente de datos	Datos operacionales, constituyen las fuentes originales de datos.	Datos consolidados y ficheros externos
Propósito del dato	Ejecución y control de las principales tareas de gestión.	Ayuda en el proceso de análisis, planificación y toma de decisiones.
Inserciones y actualizaciones	Inserciones y actualizaciones cortas y rápidas realizadas por los usuarios finales.	Actualizaciones periódicas y extensas para el refresco de datos.
Consultas	Estandarizadas y simples, con respuestas típicas de unos pocos registros.	Normalmente consultas complejas que requieren de agregaciones, cruces y filtrado de datos.
Velocidad de procesamiento	Normalmente muy rápidos.	Dependiente de la cantidad de datos almacenados y la complejidad de las consultas. El refresco del cálculo de algunas de ellas puede suponer horas.
Necesidades de espacio	Puede ser relativamente pequeño si los datos históricos se archivan.	Normalmente mayores necesidades de espacio, debido al almacenamiento de datos históricos.
Diseño de la base de datos	Altamente normalizada, con gran cantidad de tablas.	Normalmente desnormalizada, con menor número de tablas: modelos de datos en estrella, copo de nieve, etc.
Backup y recuperación	Backup imprescindible, ya que la pérdida de datos puede conllevar pérdida de dinero y consecuencias legales.	Backups realizados con cierta regularidad. En ocasiones la recuperación consiste en la nueva recarga de datos desde los sistemas OLTP.

Oportunidad de aprendizaje

- Cada interacción con el cliente, cada transacción, cada venta, cada acción comercial es una **oportunidad de aprendizaje**.
- Para aprender es necesario almacenar de manera conjunta, organizada, consistente y útil las fuentes de datos. Es lo que se conoce como el **Datawarehousing**.
- El Datawarehousing permite a las compañías recordar lo que se ha percibido.

Datawarehousing (1/2)

Un **DataWarehouse** es una colección de datos:

- Orientados
- Integrados
- No volátiles
- Variables en el tiempo

que ayudan a la toma de decisiones en la entidad en la que se utiliza

Bill Inmon

Un **DataWarehouse** es una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis

Ralph Kimball



Datawarehousing (2/2)

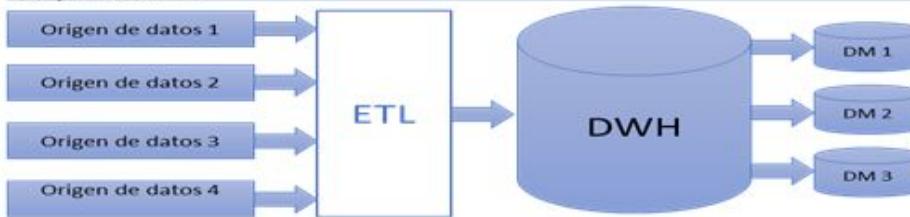
- Un Datawarehouse no se debe comprar, se debe construir
- Implantación por fases
- Implica a varios departamentos de la empresa
- El diseño no es único ni estándar



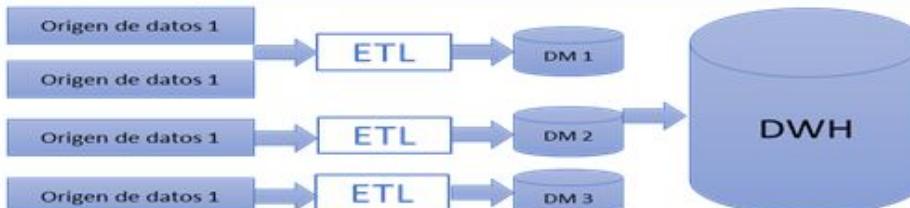
Datamarts

Un **DataMart** es un subconjunto del DataWarehouse que proporciona una visualización personalizada.

Top-Down

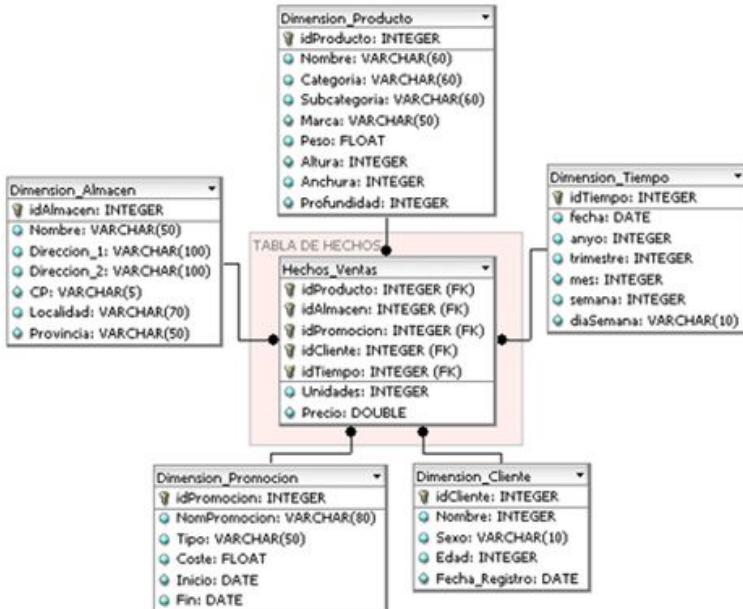


Bottom-Up

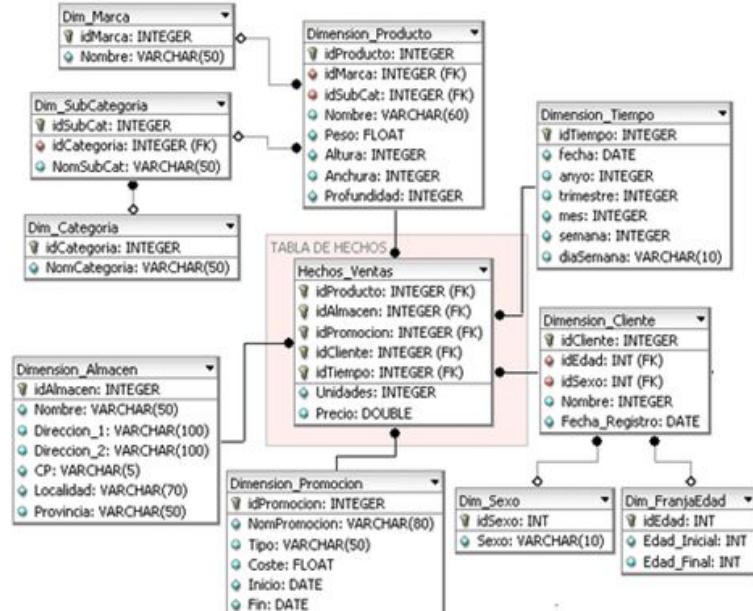


Modelización de la estructura

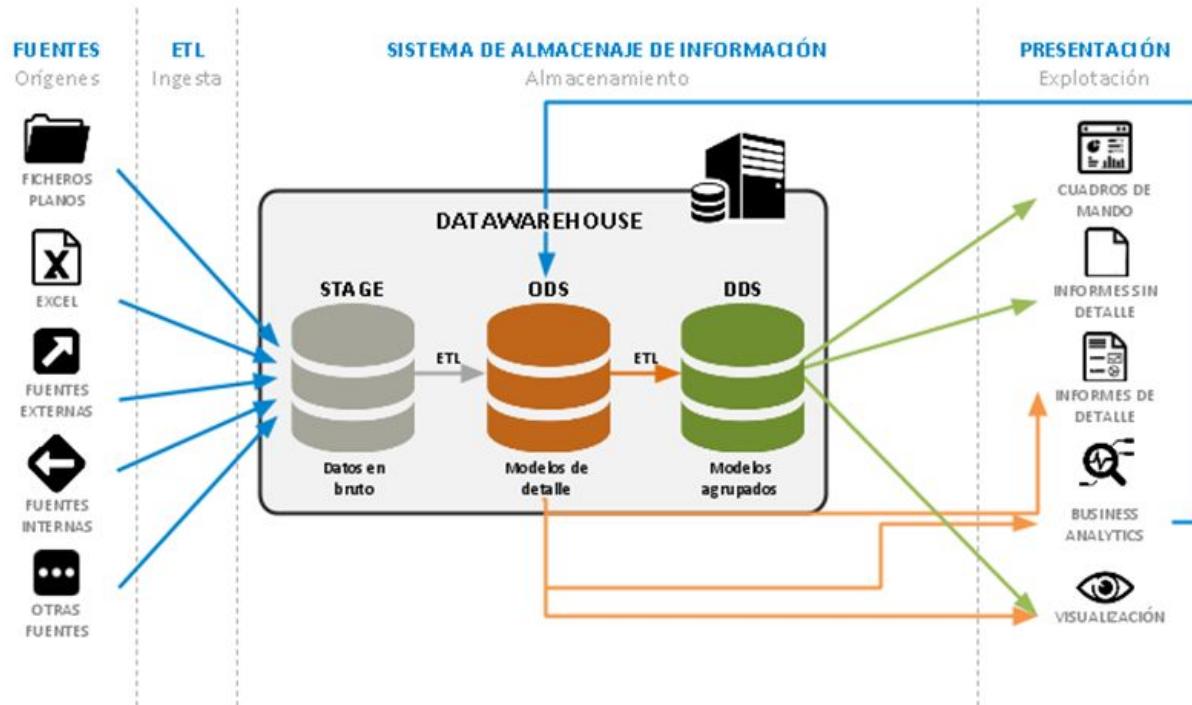
Modelo en estrella



Modelo de copo de nieve



Arquitectura de un Datawarehouse



Un Datawarehouse no es una BBDD relacional!

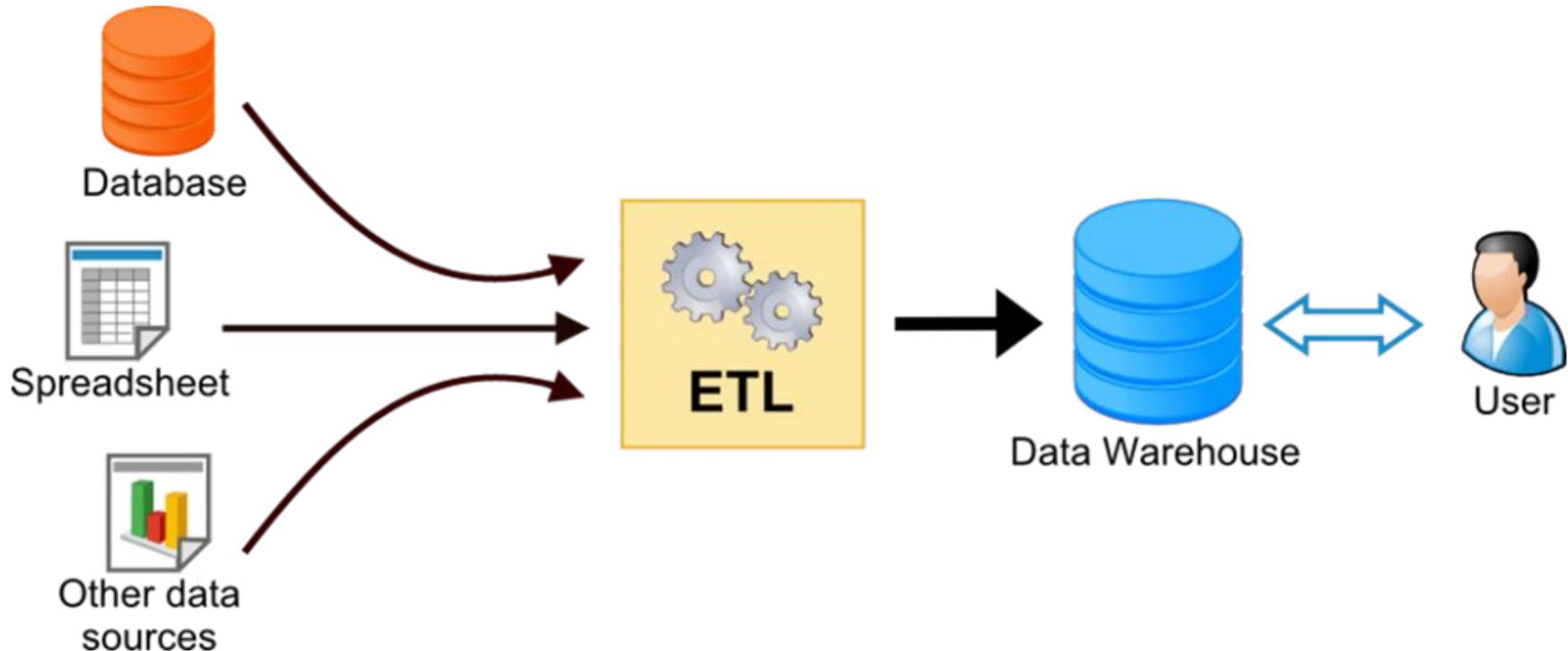
- No podemos tener valores nulos

Tipo de campo	DESCONOCIDO	NO APLICA
Numérico	9999999999999999	9999999999999998
Cadena	DESCONOCIDO	NO APLICA
Fecha	31/12/9999	31/12/9998

- Puede estar desnormalizada



Extract Transform Load



Ten siempre en cuenta que:

- Los procesos de BI son vivos
- La actualización de datos debe automatizarse
- BI debe agilizar los procesos para la toma de decisiones posterior



Big Data



keepcoding.io



cursos@keepcoding.io

|

KeepCoding®. Todos los derechos reservados

Las 4 V's del Big Data (1/2)

- **Volumen:** hoy en día hablamos de procesar, por ejemplo, todo el contenido de Twitter porque quiero saber si en Twitter alguien está hablando bien o mal de mi marca. Estamos hablando de petabytes, que son miles de terabytes, que son millones de gigabytes. Hace muy pocos años este volumen de análisis era inabordable
- **Variedad:** trabajamos de forma masiva con texto escrito en redes sociales, con valores medidos por sensores de cualquier tipo, con imágenes grabadas por cámaras de vigilancia, etc. y esto hablando sólo de datos no estructurados, además de los datos estructurados en bases de datos de toda la vida.

Las 4 V's del Big Data (2/2)

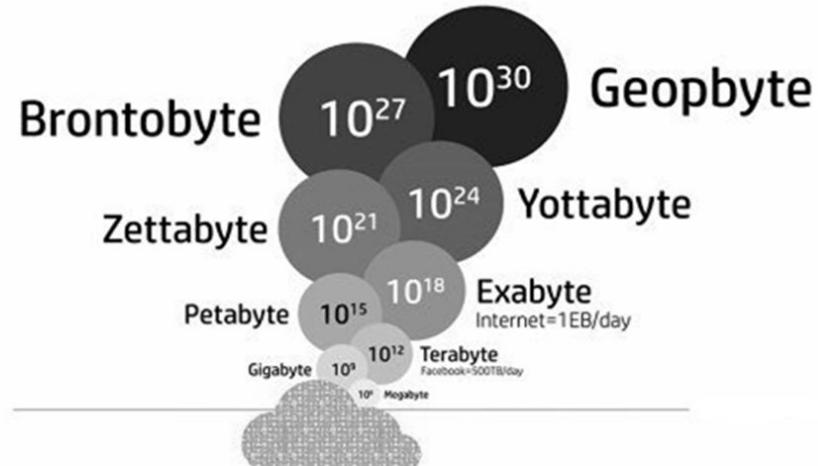
- **Velocidad**: aunque tenga que procesar millones de mensajes en Twitter, por ejemplo, quiero saber “ya” qué están diciendo los usuarios sobre mi empresa, sobre un producto que acabo de lanzar o sobre un acontecimiento que está ocurriendo ahora mismo. O quiero saber en tiempo real si mi cadena de montaje de coches robotizada está funcionando perfectamente, procesando la información de muchos miles de sensores funcionando al mismo tiempo.
- **Variabilidad** o datos “vagos”: no todos son datos concretos como los de un sensor o una medida técnica en unidades concretas. ¿Qué pasa con el texto que escribimos las personas, con sus dobles sentidos, bromas, lenguaje inventado, ironía y chistes? Eso también hay que saber interpretarlo, sobre todo en idiomas como el español; por cierto, los usuarios escriben y hablan en docenas de idiomas distintos, pero lo que quiero saber de ellos es siempre lo mismo, qué opinan de mí o si van a comprar mi producto o el de la competencia.



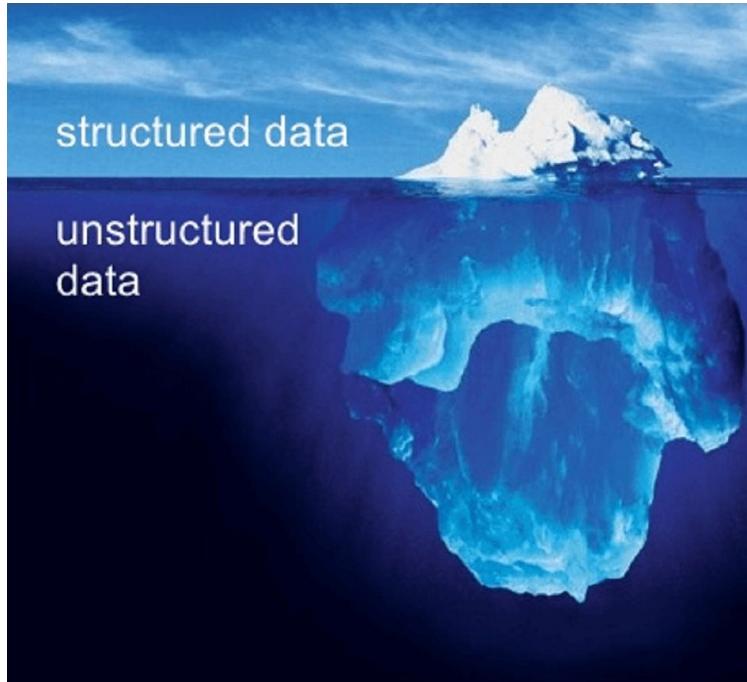
Las V's del Big Data



Datificación

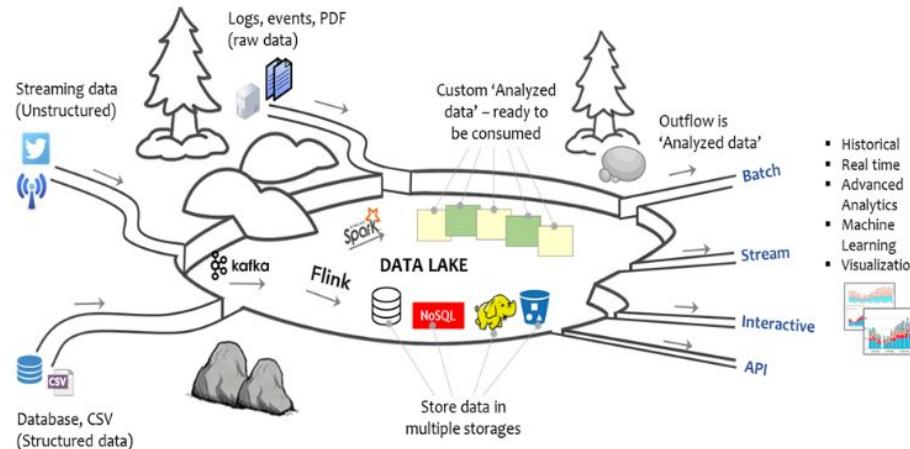


BI vs Big Data



¿Dónde guardar tantos orígenes distintos?

Data Lake

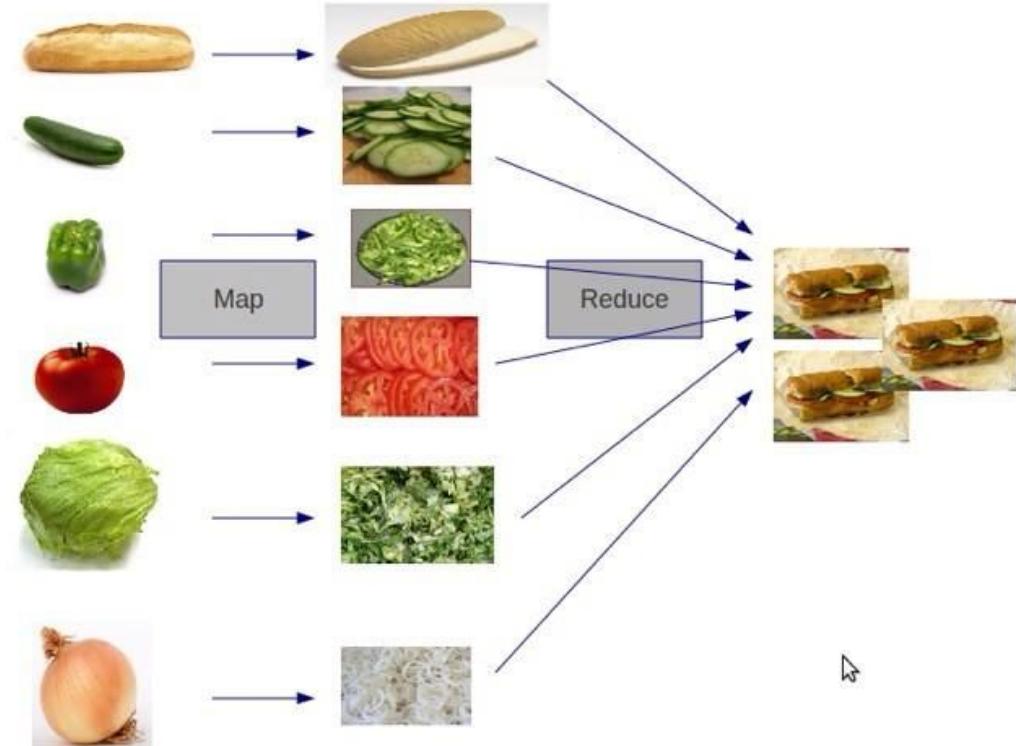


¿Cómo manejar tal cantidad de datos?

En 2004 nace **Hadoop**.

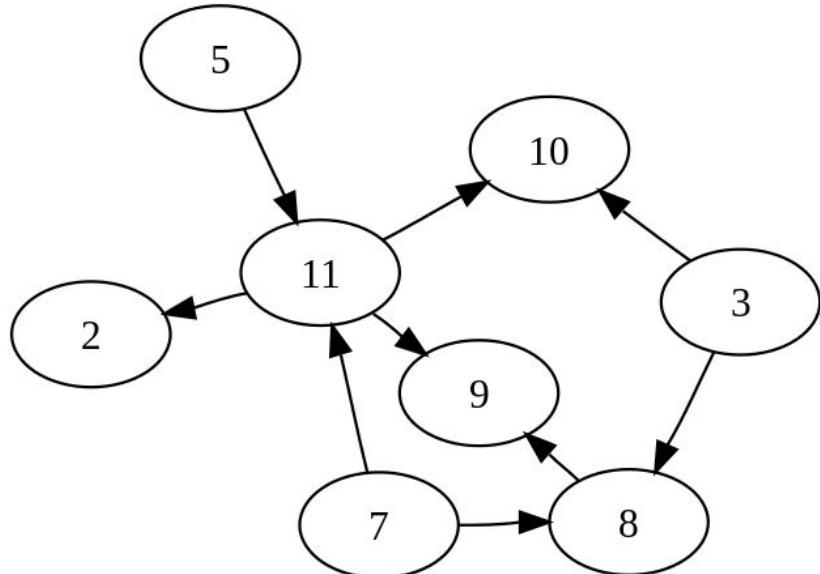
El ingeniero de software Doug Cutting, que trabajaba en Google, se inspiró en MapReduce de Google para describir en un documento técnicas para manejar grandes volúmenes de datos, dividiéndolo en problemas cada vez más pequeños para hacerlos abordables.

Trabaja en disco



¿Cómo manejar tal cantidad de datos?

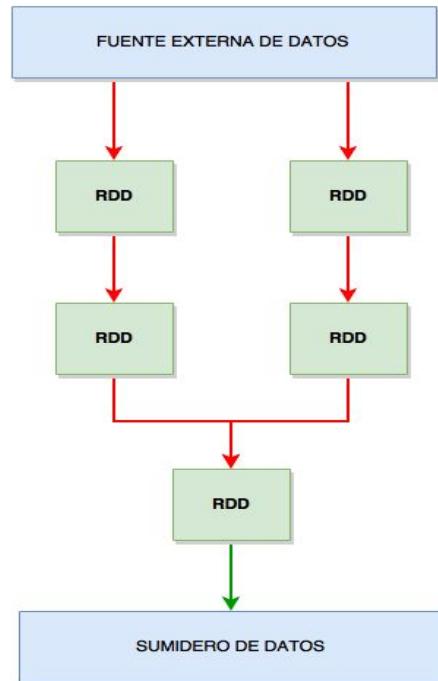
- En 2009 nace **Spark**.
- Trabaja en memoria = mayor velocidad de procesamiento
- Permite trabajar en disco: Si la cantidad de información no cabe en memoria, la herramienta permite almacenar parte en disco, aunque pierde velocidad.
- La memoria es más cara que el disco
- Permite procesamiento en tiempo real (módulo Spark Streaming)
- Usa un objeto “especial”: RDDs (tolerante a fallos y permite ejecución en paralelo)
- Usa la evaluación perezosa: guarda las transformaciones sobre los RDDs que tiene que hacer en un grafo acíclico dirigido (DAG) y sólo las ejecuta cuando no le queda más remedio



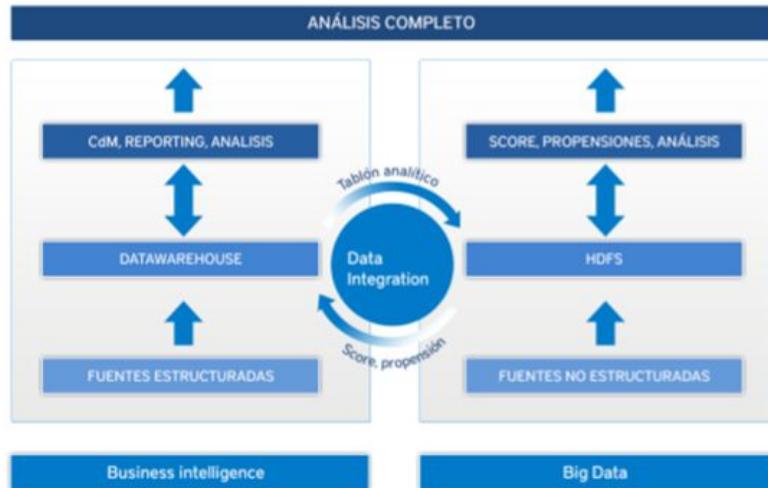
¿Cómo manejar tal cantidad de datos?

Un programa típico se organiza de la siguiente manera:

- A partir de una variable de entorno llamada context se crea un objeto RDD leyendo datos de ficheros, bases de datos o lo que sea.
- Una vez creado el RDD inicial se realizan transformaciones para crear más objetos RDD a partir del primero. Dichas transformaciones no eliminan el RDD original, sino que crean uno nuevo.
- Tras realizar las acciones y transformaciones necesarias sobre los datos, los objetos RDD deben converger para crear el RDD final. Este RDD puede ser almacenado.



BI & Big Data coexistiendo





Visuals



keepcoding.io

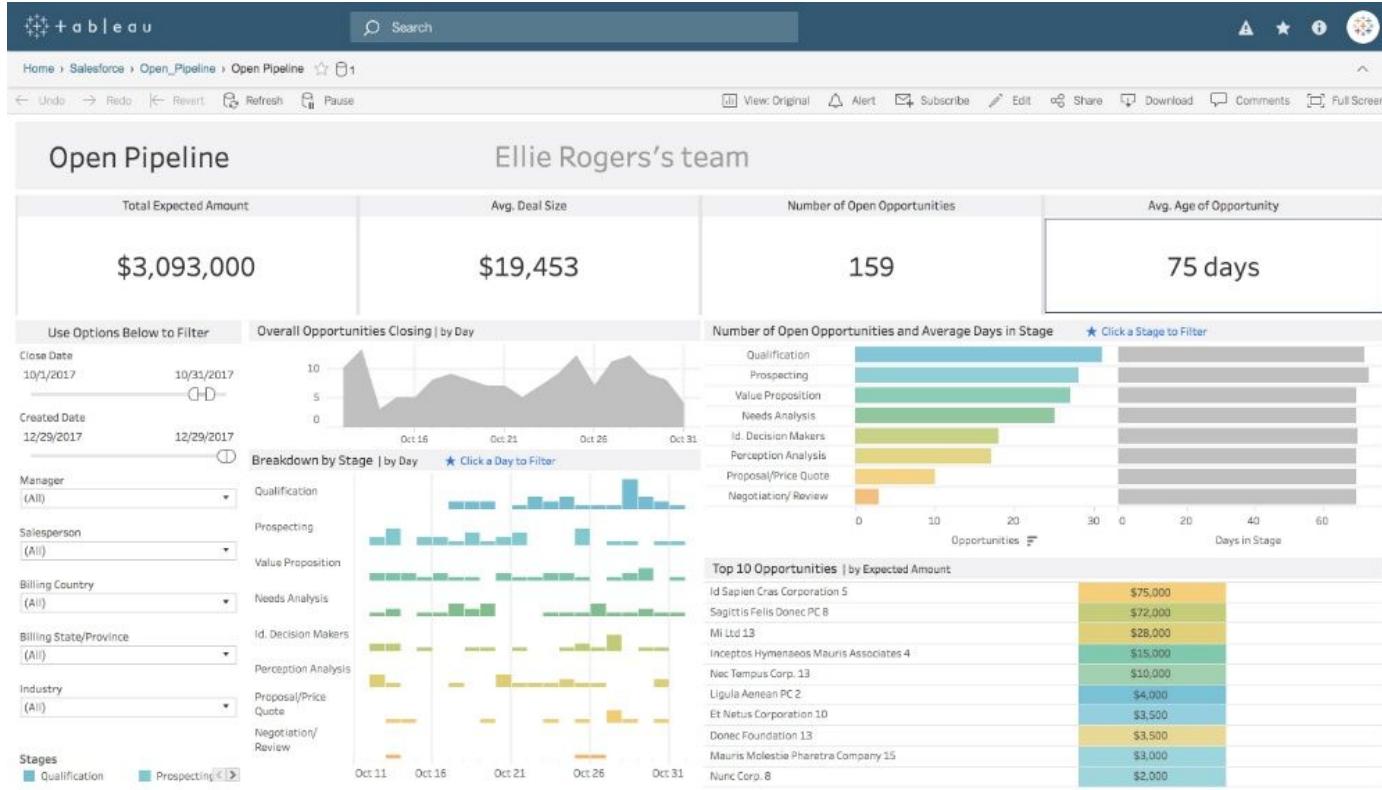


cursos@keepcoding.io

|

KeepCoding®. Todos los derechos reservados

Tableau



Power BI

Microsoft Power BI New Retail Analysis Retail Analysis Sample | Data updated 8/2/21

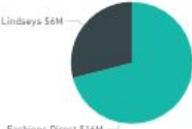
Search ...

Pages Overview District Monthly Sales New Stores District Sales Report

File Export Share Chat in Teams ...

Store Sales Overview

This Year Sales by Chain



Lindsay's \$6M
Fashions Direct \$16M

10 New Stores

104 Total Stores

This Year Sales by PostalCode and Store Type



Store Type: New Store, Same Store

UNITED STATES

MINNESOTA WISCONSIN IOWA SOUTH DAKOTA NEBRASKA KANSAS OREGON ARKANSAS TEXAS FLORIDA ALABAMA GEORGIA MISSISSIPPI LOUISIANA MICHIGAN OHIO TENNESSEE ILLINOIS INDIANA VICTORIA CANADA NS NB PEI NF

© 2021 Jamison. © 2021 Microsoft Corporation. All rights reserved.

Total Sales Variance by Fiscal Month and District Manager



District Manager: Alan Gurnet, Andrew Ma, Annelise Zubar, Brad Sutton, Carlos Grillo, Chris Gray, Chris McGurk, Tina Lassila, Valery Ushakov

Jan Feb Mar Apr May Jun Jul Aug

Total Sales Variance %, Sales Per Sq Ft and This Year Sales by District and District

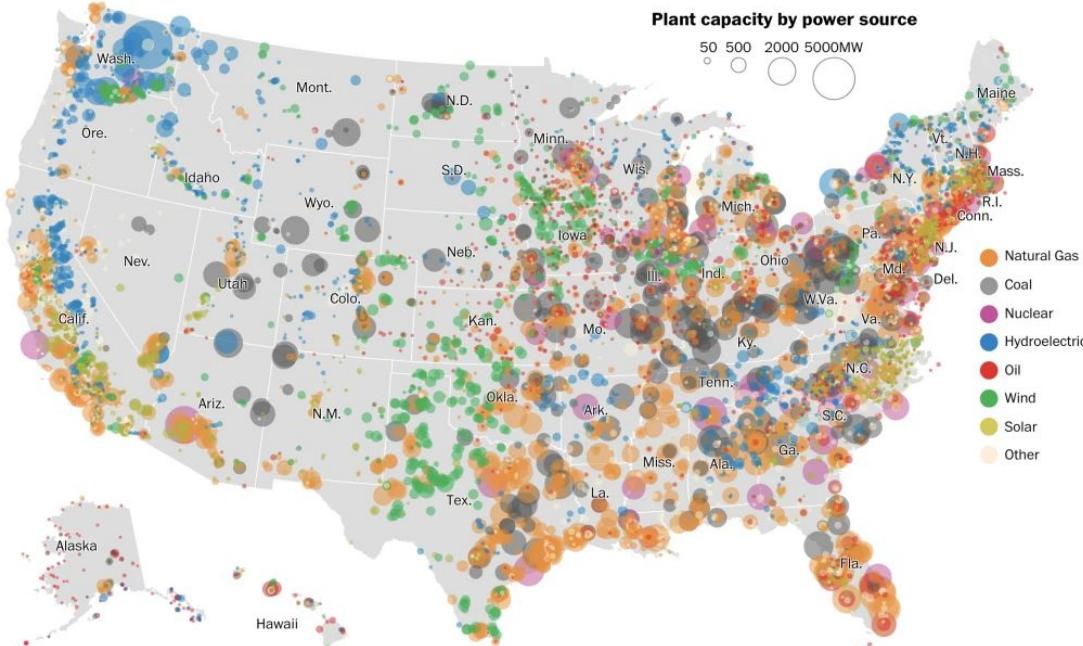


Sales Per Sq Ft

Total Sales Variance %

FO - 01
LI - 01
LI - 02
FD - 03
LI - 04
FD - 04
LI - 03
LI - 05
FO - 02

D3.js





Data Science



keepcoding.io



cursos@keepcoding.io

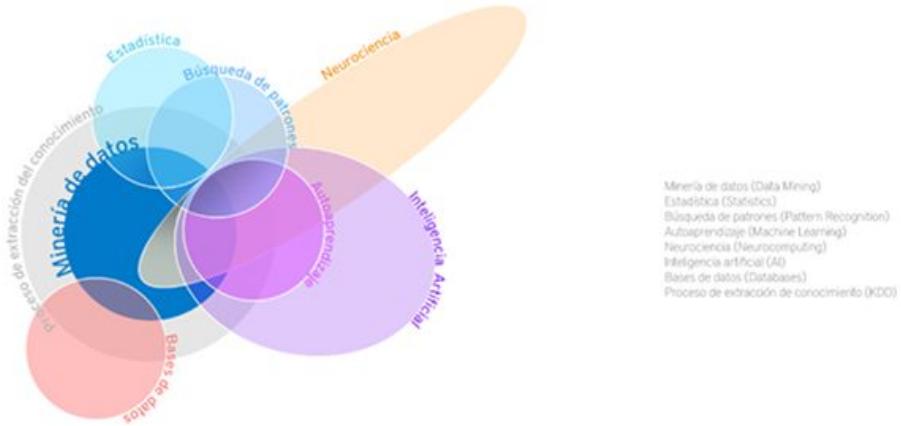


KeepCoding®. Todos los derechos reservados

Enfoque multidisciplinario - La tercera variable



Disciplinas científicas



(C) Reference SAIC Institute

Se trata de muchas disciplinas que se entrecruzan entre sí y al final, del uso de algunas de ellas y sobre todo del uso de todas ellas como conjunto, obtenemos información de gran valor para la empresa

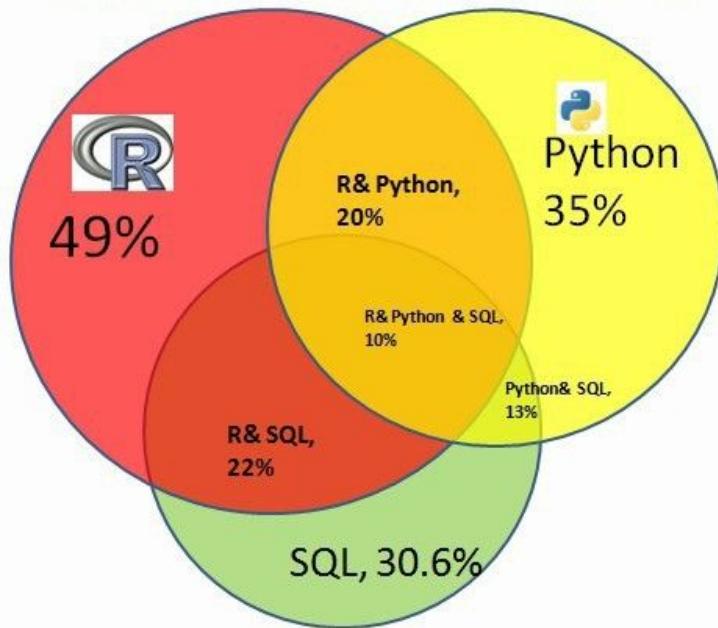
El científico de datos

- Popularmente: “Es un estadístico que trabaja en San Francisco”
- Josh Wills: “Persona que sabe más de estadística que cualquier programador y que a la vez sabe más de programación que cualquier estadístico”
- Sea como sea, es un profesional dedicado a analizar e interpretar grandes bases de datos, algo muy importante en la era digital en la que estamos inmersos
- Según Burtch Works:
 - El 32% de los científicos de datos en activo vienen del mundo de las matemáticas y la estadística
 - El 19% de la ingeniería informática
 - El 16% de otras ingenierías



Lenguaje de los data Scientists

KDnuggets 2014 Poll: Languages used for Analytics/Data Mining



Las matemáticas...

- El Big Data tiene un lado “desagradecido”, el cómo almacenar y tratar los datos.
 - La fuerza del Big Data está en optimizar procesos muy complejos, cuanto mayor es el sistema, mayor valor podrá aportar.
 - Hay que construir la infraestructura flexible y escalable para que vaya creciendo a nuestra medida.
- Su lado más “agradecido”, y a la vez más peligroso, está basado en las matemáticas. Son la matemáticas de la efectividad, ya que si tenemos muchos y diversos datos y los combinamos con las reglas adecuadas, podremos encontrar los patrones, los valores atípicos que nos ayudarán a gestionar de forma potente nuestra estrategia de negocio.
- ¡Big Data entendido como que no se trata de los datos sino del impacto que tiene en el negocio!

¿Cómo lo hacemos?

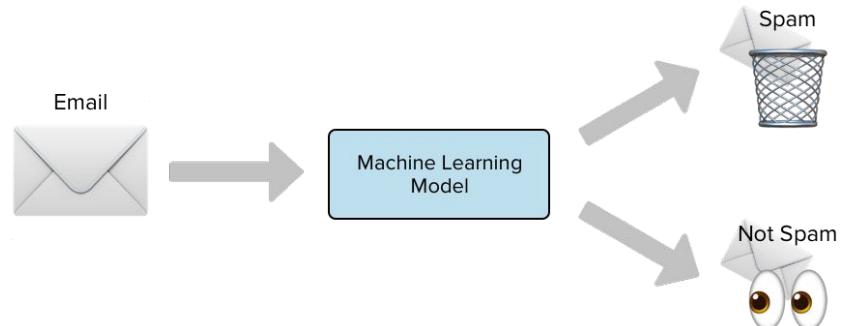
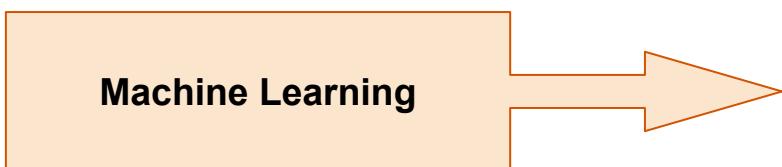
- Un primer paso del análisis puede ser usar datos pasados para predecir la posibilidad de qué va a pasar después. También el utilizar los datos del presente por medio de realtime para poder hacer un análisis predictivo potente en base a lo que ya conocemos.
- La idea es tener un sistema que aprenda a la vez que nosotros
- Se trata de encontrar los patrones, los perfiles, los modelos, las relaciones dentro de nuestros datos que nos permitan optimizar al máximo nuestro negocio.
- La anticipación y la pronta alerta será más y más importante para ganar ventaja competitiva. Estamos en un mundo en el que la ganancia marginal es la clave que nos diferencia de la competencia.
- El problema es que se hacen los análisis predictivos pero no se acompañan de las hipótesis para tomar acción. Se utilizan para entender comportamientos, para entender cómo funcionan nuestros esfuerzos, pero no como base de las acciones.



Data Mining vs Machine Learning

El Data Mining sirve para extraer las reglas de grandes cantidades de datos, mientras que el Machine Learning le enseña a una computadora cómo aprender y comprender los parámetros dados.

Comprados juntos habitualmente



Deep Learning

El Deep Learning es un subconjunto del Machine Learning, basado en redes neuronales complejas, inspiradas en las redes neuronales humanas.
Se busca que el ordenador aprenda por sí solo

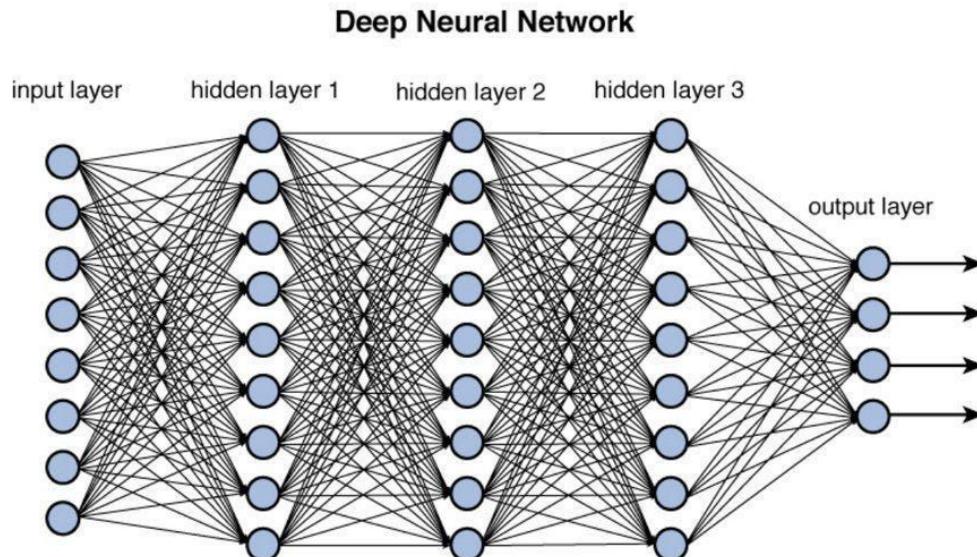
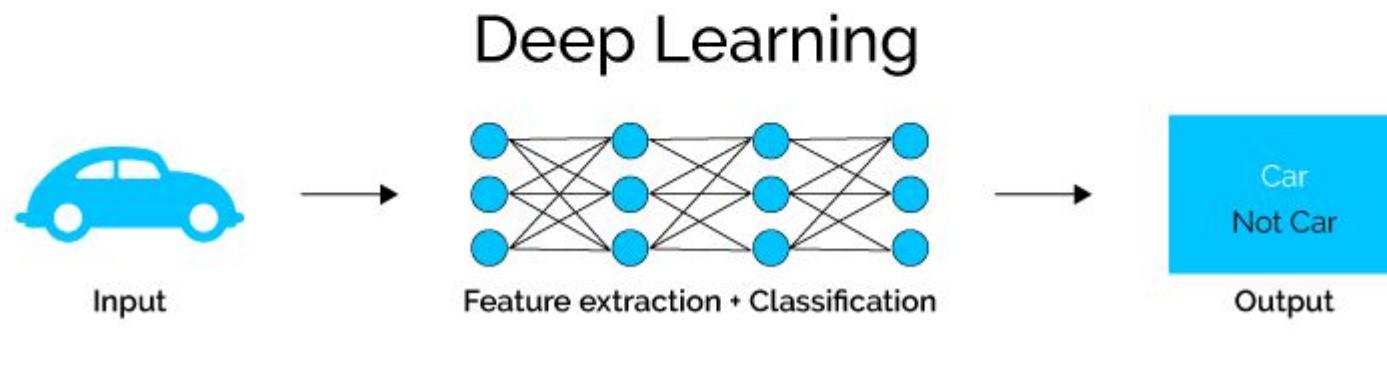
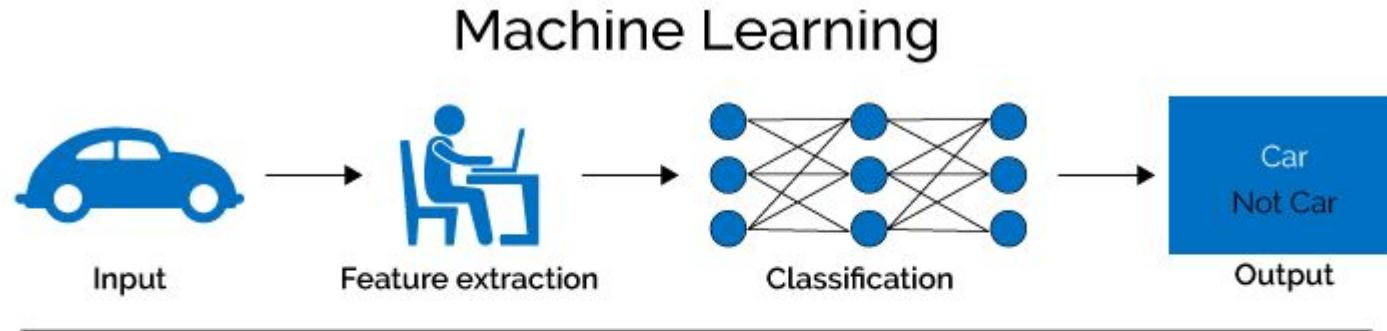


Figure 12.2 Deep network architecture with multiple layers.

Deep Learning vs Machine Learning



Natural Language Processing

El procesamiento del lenguaje natural (NLP) se ocupa de cómo las computadoras entienden y traducen el lenguaje humano.

Con NLP, las máquinas pueden dar sentido al texto escrito o hablado y realizar tareas como traducción, extracción de palabras clave, clasificación de temas y más.

Pero para automatizar estos procesos y brindar respuestas precisas, necesitará el Machine Learning. El aprendizaje automático es el proceso de aplicar algoritmos que enseñan a las máquinas cómo aprender y mejorar automáticamente a partir de la experiencia sin ser programadas explícitamente.



Artificial Intelligence

La Inteligencia Artificial son aquellas técnicas que llevan a las computadoras a adquirir ciertas habilidades de la inteligencia humana, como son:

- Entender las situaciones y los contextos
- Identificar objetos y reconocer sus significados
- Analizar y resolver problemas
- Aprender a realizar nuevas tareas
- Comprender el lenguaje natural (NLP)
- Reconocer imágenes (Computer Vision)

Artificial Intelligence vs Deep Learning vs Machine Learning vs NLP

- INTELIGENCIA ARTIFICIAL:
 - Una máquina es capaz de aprender por sí mismo (imita el razonamiento humano)
- MACHINE LEARNING:
 - Es un subconjunto de la IA
 - Las personas entran a las máquinas para reconocer patrones y extraer sus propias conclusiones
- DEEP LEARNING:
 - Un subconjunto de ML
 - Las máquinas son capaces de razonar y sacar sus propias conclusiones, aprendiendo por sí misma
- NLP
 - Es un subconjunto de la IA
 - Las máquinas son capaces de entender el lenguaje





Ciclo de vida del Big Data



keepcoding.io



cursos@keepcoding.io



KeepCoding®. Todos los derechos reservados

Ciclo de vida del Big Data





Smart Data



keepcoding.io



cursos@keepcoding.io

|

KeepCoding®. Todos los derechos reservados

Smart Data

Con el Big Data, las empresas empezaron a guardar todos los datos disponibles a su alcance y luego trabajar sobre esos datos para extraer información.

El **Smart Data** es dar un paso más, es una evolución del Big Data. Ya no se trata de almacenar todos los datos disponibles, sino sólo aquéllos que seleccionados y bien procesados aporten el valor que la empresa necesita para llevar a cabo sus objetivos en el mínimo tiempo posible.



Smart Visual Data

Tras una selección inteligente de los datos de valor, lo interesante es poder mostrarlos de una forma visual y comprensible, como se viene haciendo desde siempre con los datos que ha habido en cada momento.

El Smart Visual Data siguen siendo dashboards o paneles de datos en tiempo real, como ocurría con el Big Data, y desde mucho antes con el Business Intelligence, pero ahora enfocado a los datos de valor, aunque con el mismo objetivo: que las personas que los manejen entiendan e interpreten mejor y más rápido los resultados que se obtienen.

Sin embargo, no nos equivoquemos pensando en los típicos informes tediosos y aburridos que se hacen desde hace años. Ahora hablamos de informes gráficos, con llamativos diseños y formados principalmente de gráficos / imágenes que son mucho más simples de entender a pesar de que tienen la misma información.



Para terminar...dejemos volar la imaginación



Aún falta...¿pero cuánto?

https://www.youtube.com/watch?v=gSOY1I_Nr2Q&t=



keepcoding.io



cursos@keepcoding.io

|

KeepCoding®. Todos los derechos reservados



KEEPCODING

Tech School

Madrid | Barcelona | Bogotá