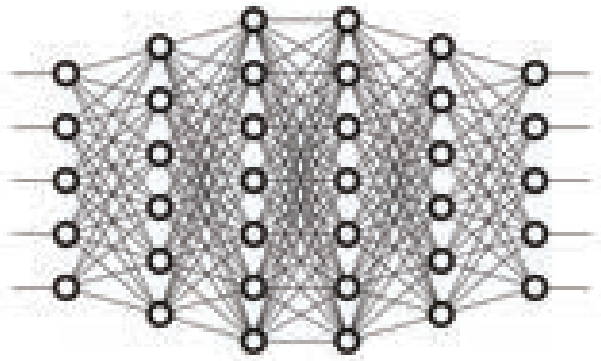
INTRODUCCIÓN

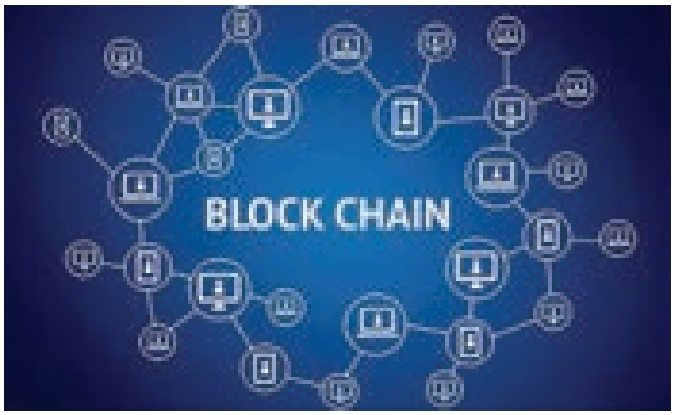
En una economía globalizada e hiperconectada las tecnologías disruptivas aquellas que producen rupturas bruscas causan profundos cambios que han permitido la Integración de grandes volúmenes de datos, hiperconexión e Intercambio de Valor, la información se ha convertido en un activo muy preciado dentro y fuera de las organizaciones. Las nuevas tecnologías juegan un papel crucial y relevante en las organizaciones, entre las que tenemos:

- Big Data. Gestionar grandes volúmenes de información estructurada, Semi Estructurados y No Estructurados

- Deep Learning. Construcción de modelos descriptivos, diagnostico, predictivos y prescriptivos

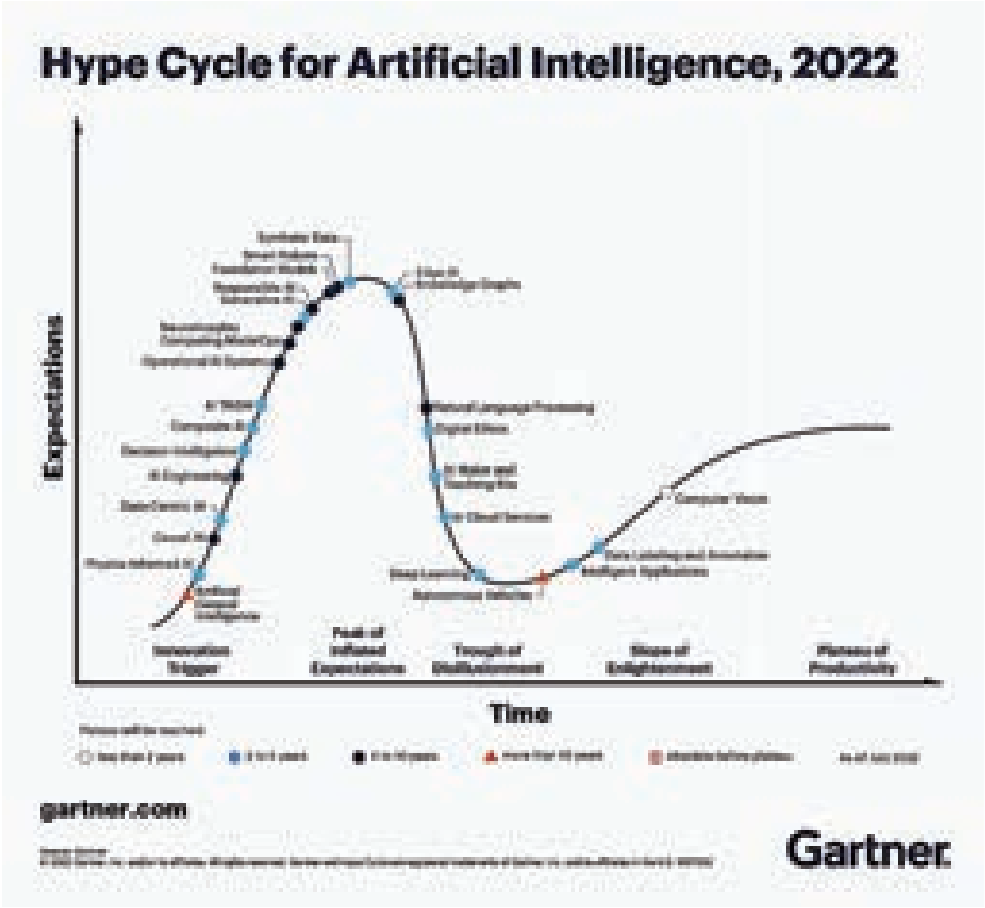


- Intercambio de valor a través del Blockchain



- Ciberseguridad para la protección de los datos y activos

El Hype Cycle de Gartner 2022 para tecnologías disruptivas o emergentes destacan las tecnologías que afectarán significativamente a los negocios, la sociedad y las personas en los próximos 2 a 10 años:



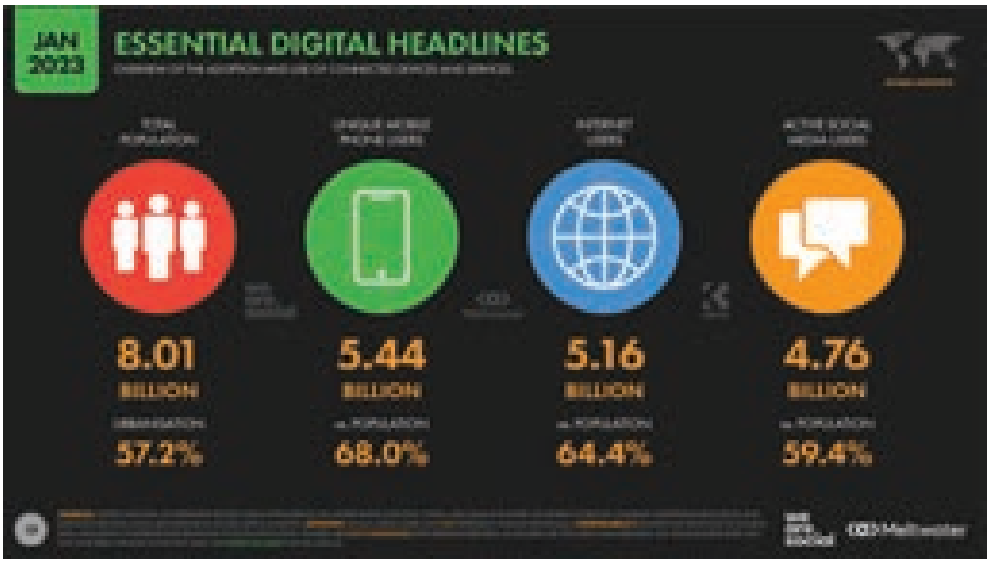
En el presente libro destacamos algunos como Big Data, Deep Learning, Blockchain y Ciberseguridad. Estas tecnologías innovadoras intrínsecamente relacionadas y en crecimiento en la industria.

1

BIG DATA

"Es un término que describe grandes volúmenes de datos estructurados y no estructurados, que debido a su tamaño, complejidad y rapidez para generar nueva información, pueden ser difíciles de recopilar, gestionar procesar y analizar sin el uso de la tecnología apropiada."

1.1 ALGUNAS CIFRAS IMPORTANTES



Fuente: https://datareportal.com/reports/digital-2023-global-overview-report

1.1.1 Cifras del estado digital en el mundo

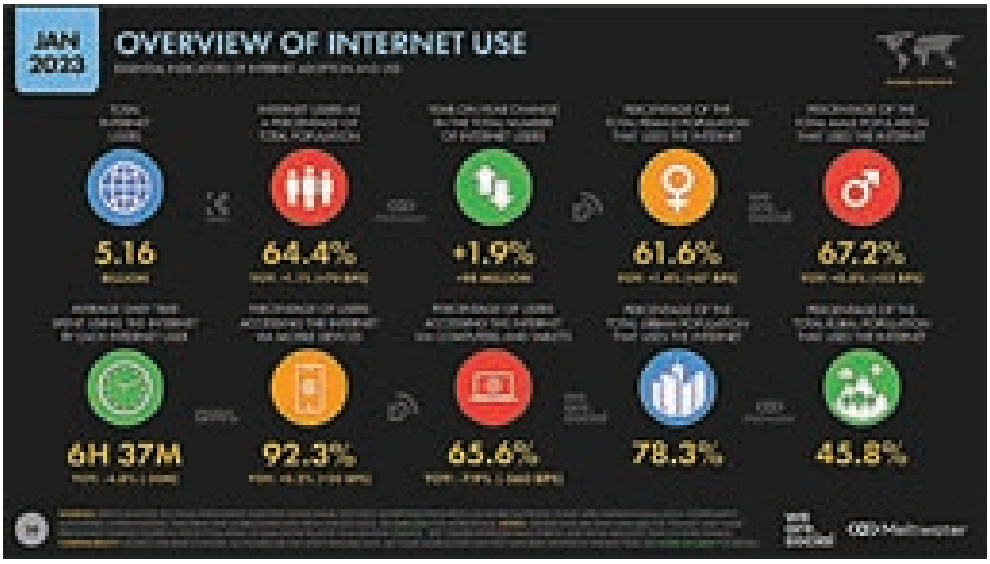
La población mundial superó los 8.000 millones el 15 de noviembre de 2022 y alcanzó los 8.010 millones a principios de 2023. Un poco más del 57 % de la población mundial ahora vive en áreas urbanas.

Un total de 5.440 millones de personas utilizan teléfonos móviles a principios de 2023, lo que equivale al 68 % de la población mundial total. Los usuarios móviles únicos aumentaron un poco más del 3 por ciento durante el año pasado, con 168 millones de nuevos usuarios en los últimos 12 meses.

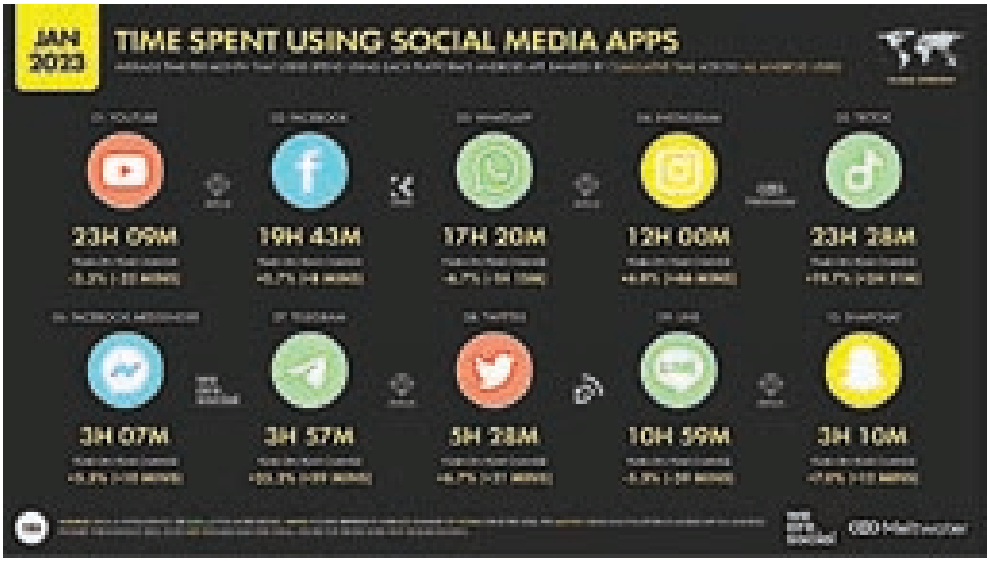
Actualmente hay 5.160 millones de usuarios de Internet en el mundo, lo que representa el 64,4% de la población total del mundo. Los datos muestran que el total de usuarios de Internet en todo el mundo aumentó un 1,9 % en los últimos 12 meses, pero los retrasos en la presentación de datos significan que el crecimiento real probablemente será mayor de lo que sugiere esta cifra.

Ahora hay 4.760 millones de usuarios de redes sociales en todo el mundo, lo que equivale a poco menos del 60% de la población mundial total. Sin embargo, el crecimiento de los usuarios de las redes sociales se ha desacelerado en los últimos

meses, con el incremento de este año de 137 millones de nuevos usuarios, lo que equivale a un crecimiento anual de solo el 3%.



Fuente: https://datareportal.com/reports/digital-2023-global-overview-report



Fuente: https://datareportal.com/reports/digital-2023-global-overview-report

1.1.2 Tiempo en redes sociales por plataforma

El ascenso de TikTok a la cima de estas clasificaciones puede no ser una sorpresa, pero algunos de los datos que impulsan su ascenso aún pueden sorprender a algunos.

Por ejemplo, los propios datos de la plataforma muestran que las publicaciones etiquetadas con #FYP ("para tu página") ahora se han visto un total de 35 billones de veces.

Incluso si cada una de esas vistas solo durara un segundo, eso sumaría más de 1 millón de años de existencia humana combinada y eso es solo para videos etiquetados con #FYP.

Pero en lo que podría ser otra sorpresa si ha estado leyendo los principales medios de comunicación recientemente, Facebook ocupa el tercer lugar en las cifras de data. ai por tiempo promedio invertido por usuario, con casi 20 horas por mes.

Además, la inteligencia de data.ai revela que el tiempo promedio que los usuarios de Facebook pasan usando la aplicación de Android de la plataforma aumentó durante el año pasado, de un promedio de 19,6 horas por mes, por usuario en 2021, a 19,7 horas en 2022.

Y solo para agregar contexto, los datos de Statcounter sugieren que los teléfonos Android representan el 72% de todos los teléfonos inteligentes en uso en la actualidad.

Por otro lado, es interesante notar que el usuario típico de Instagram solo pasa la mitad del tiempo usando la plataforma que los usuarios de TikTok. Pero el uso de Instagram varía significativamente de un país a otro. Por ejemplo, el usuario típico de Instagram en Turquía pasa un promedio de 21,4 horas al mes usando la aplicación Android de la plataforma, pero en Corea del Sur, esa cifra se reduce a solo 6,1 horas al mes.

Preferencias de redes sociales por edad y sexo. Bueno, en lo que podría ser otra sorpresa, los datos de GWI revelan que Instagram sigue siendo la plataforma de redes sociales "favorita" entre los usuarios de Internet de 16 a 24 años. Por contexto, la popularidad de TikTok continúa aumentando, y los datos de GWI revelan que la cantidad de mujeres de 16 a 24 años que identifican el servicio de videos cortos como su plataforma social "favorita" aumentó en más de un tercio durante el último año. Sin embargo, los mismos datos revelan que casi el doble de mujeres en este grupo de edad cita Instagram como su plataforma "favorita" en comparación con TikTok (23,1 por ciento frente a 12,0 por ciento, respectivamente).

Los hombres jóvenes son aún más propensos a elegir Instagram en lugar de TikTok, pero, quizás en la mayor sorpresa en este conjunto de datos, los hombres de 16 a 24 años en realidad son más propensos a citar a Facebook como su plataforma social favorita que a elegir TikTok (10.5 por ciento versus 7.7 por ciento, respectivamente).

También es interesante notar que WhatsApp ocupa el segundo lugar entre este grupo de edad, con mujeres y hombres jóvenes que colocan la plataforma de mensajería favorita del mundo por delante de TikTok.

Fuente: *https://datareportal.com/reports/digital-2023-global-overview-report*

1.1.3 Digital España

En España había 45,12 millones de usuarios de Internet a principios de 2023, cuando la penetración de Internet se situó en el 94,9 por ciento. España albergaba 40.70 millones de usuarios de redes sociales en enero de 2023, lo que equivale al 85,6% de la población total.

Un total de 58,32 millones de conexiones móviles celulares estaban activas en España a principios de 2023, cifra que equivale al 122,7% de la población total.

Fuente: *https://datareportal.com/reports/digital-2023-spain*

1.1.3.1 POBLACIÓN DE ESPAÑA EN 2023

- La población total de España era de 47,54 millones en enero de 2023.
- Los datos muestran que la población de España disminuyó en 35 mil (-0,07%) entre 2022 y 2023.
- El 51% de la población española es femenina, mientras que el 49% de la población es masculina.
- A principios de 2023, el 81.4% de la población española vivía en centros urbanos, mientras que el 18,6% vivía en zonas rurales.

1.1.3.2 POBLACIÓN DE ESPAÑA POR EDAD

- La edad media de la población española es de 44.7 años.
- Así es como se desglosa la población total de España por grupos de edad:
- El 3,8% tiene entre 0 y 4 años.
- El 7,6% tiene entre 5 y 12 años.
- el 5,4% tiene entre 13 y 17 años.
- el 7,3% tiene entre 18 y 24 años.
- el 11,0% tiene entre 25 y 34 años.
- el 14,0% tiene entre 35 y 44 años.
- El 16,4% tiene entre 45 y 54 años.
- El 14,1% tiene entre 55 y 64 años.
- El 20,5% tiene 65 años o más.

1.1.3.3 INTERNET EN ESPAÑA

- En enero de 2023 había 45,12 millones de internautas en España.
- La tasa de penetración de Internet en España se situó en el 94,9% de la población total a principios de 2023.
- El análisis de Kepios indica que los internautas en España disminuyeron en 33 mil (-0,07%) entre 2022 y 2023.
- En perspectiva, estas cifras de usuarios revelan que 2.42 millones de personas en España no usaban Internet a principios de 2023, lo que sugiere que el 5.1% de la población permanecía desconectada a principios de año.

1.1.3.4 VELOCIDADES DE CONEXIÓN A INTERNET EN ESPAÑA EN 2023

Los datos publicados por Ookla indican que los usuarios de Internet en España podrían haber esperado las siguientes velocidades de conexión a Internet a principios de 2023:

- Velocidad media de conexión a internet móvil a través de redes celulares: 36,07 Mbps.
- Velocidad media de conexión a Internet fija: 166,78 Mbps.
- Los datos de Ookla revelan que la velocidad media de conexión a Internet móvil en España aumentó 1,78 Mbps (+5,2%) en los doce meses hasta principios de 2023.
- Mientras tanto, los datos de Ookla muestran que las velocidades de conexión a Internet en España aumentaron en 35,32 Mbps (+26,9%) durante el mismo período.

1.1.3.5 ESTADÍSTICAS DE REDES SOCIALES PARA ESPAÑA EN 2023

- En enero de 2023 había 40,70 millones de usuarios de redes sociales en España.
- Esta cifra puede parecer bastante diferente a los valores que publicamos en años anteriores, pero tenga en cuenta que las fuentes que utilizamos para informar y calcular el número de usuarios de las redes sociales han realizado importantes y amplias revisiones a sus datos en los últimos meses.
- Estos ajustes a los datos de origen significan que nuestras últimas cifras no son comparables con las cifras equivalentes que publicamos en años anteriores, y los lectores no deben considerar ninguna diferencia en estas cifras como un cambio real en el uso de las redes sociales.

- De hecho, nuestro análisis de varios puntos de datos de terceros confiables, como GWI y data.ai, muestra que no ha habido una caída perceptible en el uso general de las redes sociales, y en casi todos los países, el uso de las redes sociales continúa aumentando.
- El número de usuarios de redes sociales en España a principios de 2023 era equivalente al 85,6 % de la población total, pero también es importante tener en cuenta que es posible que los usuarios de redes sociales no representen a personas únicas (consulte nuestras notas detalladas sobre los datos para saber por qué).
- Mientras tanto, los datos publicados en las herramientas de planificación de anuncios de las principales plataformas de redes sociales indican que había 35,60 millones de usuarios mayores de 18 años que usaban las redes sociales en España a principios de 2023, lo que equivalía al 90,1 por ciento de la población total mayor de 18 años. En ese tiempo.
- En términos más generales, el 90,2 por ciento de la base total de usuarios de Internet de España (independientemente de la edad) usó al menos una plataforma de redes sociales en enero de 2023.
- En ese momento, el 51,1 por ciento de los usuarios de las redes sociales en España eran mujeres, mientras que el 48,9 por ciento eran hombres.

1.1.3.6 USUARIOS DE FACEBOOK EN ESPAÑA EN 2023

- Los datos publicados en los recursos publicitarios de Meta indican que Facebook contaba con 19,35 millones de usuarios en España a principios de 2023.
- Sin embargo, Meta ha realizado cambios importantes en la forma en que sus recursos publicitarios informan los datos de alcance de audiencia en los últimos meses, incluida la realización de revisiones significativas a sus datos de audiencia base para Facebook, por lo que las cifras que se muestran aquí pueden no ser directamente comparables con las cifras publicadas en nuestros informes anteriores.
- Cifras publicadas en las propias herramientas de Meta indican que el alcance potencial de los anuncios de Facebook en España disminuyó en 850 mil (-4,2 por ciento) entre 2022 y 2023.
- Para un contexto más reciente, los mismos datos muestran que la cantidad de usuarios a los que los especialistas en marketing podrían llegar con anuncios en Facebook en España disminuyó en 350 mil (-1,8 por ciento) entre octubre de 2022 y enero de 2023.

- Sin embargo, es importante enfatizar que estas cifras de alcance publicitario no son las mismas que las cifras de usuarios activos mensuales que Meta reporta en sus anuncios de ganancias para inversionistas, y no deben interpretarse como tales.
- Como afirma la compañía dentro de sus herramientas de planificación de anuncios, "El tamaño estimado de la audiencia no es un indicador de los usuarios activos mensuales o diarios, ni de la participación. Las estimaciones no están diseñadas para coincidir con la población, las estimaciones del censo u otras fuentes, y pueden diferir según factores como cuántas cuentas en tecnologías Meta tiene una persona, cuántos visitantes temporales hay en una ubicación geográfica particular en un momento dado y Meta datos demográficos informados por los usuarios".
- Pero a pesar de estas advertencias, los datos de alcance de anuncios de Meta aún ofrecen información valiosa sobre cómo está evolucionando el uso de Facebook.
- El alcance de los anuncios de Facebook en España era equivalente al 40,7 por ciento de la población total a principios de 2023.
- Para un contexto adicional, el alcance de los anuncios de Facebook en España fue equivalente al 42,9 por ciento de la base de usuarios de Internet local (independientemente de la edad) en enero de 2023.
- A principios de 2023, el 54,0 por ciento de la audiencia de anuncios de Facebook en España eran mujeres, mientras que el 46,0 por ciento eran hombres.

1.1.3.7 USUARIOS DE YOUTUBE EN ESPAÑA EN 2023

- YouTube contaba con 40.70 millones de usuarios en España a principios de 2023.
- Sin embargo, es importante enfatizar que estas cifras de alcance publicitario no representan necesariamente lo mismo que las cifras mensuales de usuarios activos, y puede haber diferencias significativas entre el tamaño de la audiencia publicitaria de YouTube y su base total de usuarios activos.
- Sin embargo, los propios datos de la compañía sugieren que el alcance de los anuncios de YouTube a principios de 2023 equivalía al 85,6 por ciento de la población total de España a principios de año.
- Para poner esas cifras en perspectiva, los anuncios de YouTube alcanzaron el 90,2 por ciento de la base total de usuarios de Internet de España (sin importar la edad) en enero de 2023.

- En ese momento, el 51,1 por ciento de la audiencia de anuncios de YouTube en España era femenina, mientras que el 48,9 por ciento era masculina.
- Los datos publicados en las propias herramientas de planificación de anuncios de Google muestran que el alcance potencial de los anuncios de YouTube en España se mantuvo sin cambios entre principios de 2022 y principios de 2023.
- Sin embargo, los mismos datos muestran que la cantidad de usuarios a los que los especialistas en marketing podrían llegar con anuncios en YouTube en España aumentó en 400 mil (+1,0 por ciento) entre octubre de 2022 y enero de 2023, lo que sugiere que el alcance puede haber disminuido entre enero de 2022 y octubre de 2022.

1.1.3.8 USUARIOS TIKTOK EN ESPAÑA EN 2023

- Las cifras publicadas en los recursos publicitarios de ByteDance indican que TikTok contaba con 16,63 millones de usuarios de 18 años o más en España a principios de 2023.
- Tenga en cuenta que ByteDance permite a los especialistas en marketing orientar los anuncios de TikTok a usuarios de 13 años o más a través de sus herramientas publicitarias, pero estas herramientas solo muestran datos de audiencia para usuarios de 18 años o más.
- A modo de contexto, las cifras de ByteDance indican que los anuncios de TikTok llegaron al 42,1 % de todos los adultos mayores de 18 años en España a principios de 2023.
- Mientras tanto, el alcance de los anuncios de TikTok en España equivalía al 36,9 por ciento de la base de usuarios locales de Internet a principios de año, independientemente de la edad.
- A principios de 2023, el 57,0 por ciento de la audiencia publicitaria de TikTok en España era femenina, mientras que el 43,0 por ciento era masculina.
- Los datos publicados en las propias herramientas de planificación de anuncios de ByteDance muestran que el alcance potencial de anuncios de TikTok en España aumentó en 2,9 millones (+21,1 por ciento) entre principios de 2022 y principios de 2023.
- Mientras tanto, las cifras indican que el alcance potencial de los anuncios en TikTok en España aumentó en 836 mil (+5,3 por ciento) entre octubre de 2022 y enero de 2023.
- Dicho esto, las audiencias publicitarias a menudo solo representan un subconjunto del total de usuarios de una plataforma, y dado que las

herramientas publicitarias de TikTok solo publican datos para usuarios mayores de 18 años, es importante recordar que estos cambios en el alcance de los anuncios de TikTok pueden no coincidir necesariamente con los cambios. en la base general de usuarios de la plataforma.

1.1.3.9 USUARIOS INSTAGRAM EN ESPAÑA EN 2023

- Cifras publicadas en las herramientas publicitarias de Meta indican que Instagram contaba con 21,90 millones de usuarios en España a principios de 2023.
- Las cifras revisadas recientemente por la compañía sugieren que el alcance de los anuncios de Instagram en España equivalía al 46,1 por ciento de la población total a principios de año.
- Sin embargo, Instagram restringe el uso de su plataforma a personas mayores de 13 años, por lo que es útil saber que el 52,0 por ciento de la audiencia "elegible" en España usa Instagram en 2023.
- También vale la pena señalar que el alcance de los anuncios de Instagram en España a principios de 2023 era equivalente al 48,5 por ciento de la base de usuarios de Internet local (independientemente de la edad).
- A principios de 2023, el 53,8 por ciento de la audiencia de anuncios de Instagram en España eran mujeres, mientras que el 46,2 por ciento eran hombres.
- Los datos publicados en las herramientas de planificación de Meta muestran que el alcance potencial de los anuncios de Instagram en España disminuyó en 950 mil (-4,2 por ciento) entre 2022 y 2023.
- Con carácter trimestral, los datos de la compañía también revelan que el tamaño de la audiencia publicitaria de Instagram en España disminuyó en 750 mil (-3,3 por ciento) entre octubre de 2022 y enero de 2023.

1.1.3.10 USUARIOS DE LINKEDIN EN ESPAÑA EN 2023

- Las cifras publicadas en los recursos publicitarios de LinkedIn indican que LinkedIn contaba con 17 millones de "miembros" en España a principios de 2023.
- Sin embargo, tenga en cuenta que las herramientas publicitarias de LinkedIn publican datos de alcance de audiencia basados en el total de miembros registrados, en lugar de los usuarios activos mensuales que forman la base de las cifras de alcance de anuncios publicadas por la mayoría de las otras plataformas de redes sociales.

- Como resultado, estas cifras de LinkedIn no son directamente comparables con las cifras de otras plataformas de redes sociales publicadas en esta página o en nuestros informes Digital 2023.
- Las cifras de alcance publicitario de la compañía sugieren que la audiencia de LinkedIn en España equivalía al 35,8 por ciento de la población total a principios de 2023.
- Sin embargo, LinkedIn restringe el uso de su plataforma a personas mayores de 18 años, por lo que también es útil saber que el 43,0 por ciento de la audiencia "elegible" en España usa LinkedIn en 2023.
- Como contexto adicional, el alcance de los anuncios de LinkedIn en España fue equivalente al 37,7 por ciento de la base de usuarios locales de Internet (independientemente de la edad) a principios de año.
- A principios de 2023, el 47,1 % de la audiencia de anuncios de LinkedIn en España eran mujeres, mientras que el 52,9 % eran hombres.

ⓘ NOTA

Los recursos publicitarios de LinkedIn solo publican datos de género de la audiencia para usuarios "femeninos" y "masculinos".

- Los datos publicados en las herramientas de planificación de LinkedIn muestran que el alcance potencial de los anuncios de LinkedIn en España aumentó en 3,0 millones (+21,4 por ciento) entre 2022 y 2023.
- Trimestralmente, los datos de la compañía revelan que el tamaño de la audiencia publicitaria de LinkedIn en España aumentó en 1,0 millones (+6,3 por ciento) entre octubre de 2022 y enero de 2023.

1.1.3.11 USUARIOS DE TWITTER EN ESPAÑA EN 2023

- Las cifras publicadas en los recursos publicitarios de Twitter indican que Twitter contaba con 10,85 millones de usuarios en España a principios de 2023.
- Esta cifra significa que el alcance de los anuncios de Twitter en España equivalía al 22,8 por ciento de la población total en ese momento.
- Sin embargo, es importante enfatizar que estas cifras de alcance publicitario no son las mismas que las cifras mensuales de usuarios activos, y puede haber diferencias significativas entre el tamaño de la audiencia publicitaria de Twitter y su base total de usuarios activos.

- También vale la pena señalar que Twitter restringe el uso de su plataforma a personas mayores de 13 años, por lo que estas cifras sugieren que el 25,8 por ciento de la audiencia "elegible" en España usa Twitter en 2023.
- Como contexto adicional, el alcance de los anuncios de Twitter en España fue equivalente al 24,0 por ciento de la base de usuarios locales de Internet (independientemente de la edad) a principios de año.
- A principios de 2023, los propios datos de la compañía indicaban que el 38,4 por ciento de la audiencia de anuncios de Twitter en España era femenina, mientras que el 61,6 por ciento era masculina.
- Sin embargo, vale la pena señalar que Twitter infiere el género de sus usuarios al analizar señales como el nombre que los usuarios ingresan en su perfil y su actividad más amplia en la plataforma.
- Los datos publicados en las propias herramientas de planificación de anuncios de Twitter muestran que el alcance potencial de los anuncios de Twitter en España aumentó en 2,1 millones (+24,0 por ciento) entre principios de 2022 y principios de 2023.
- Mientras tanto, los mismos datos muestran que la cantidad de usuarios a los que los especialistas en marketing podrían llegar con anuncios en Twitter en España disminuyó en 950 mil (-8.1 por ciento) entre octubre de 2022 y enero de 2023.

1.2 DATOS

- Reflejan hechos recogidos en la organización y que están todavía sin procesar. Quedan perfectamente identificados por elementos simbólicos (letras, números, etc.).
- Flujo de hechos en bruto que representan sucesos ocurridos en las organizaciones o en el entorno físico, antes de ser organizados y acomodados de tal forma que las personas puedan entenderlos y usarlos.

- Descripciones básicas de cosas, acontecimientos, actividades y transacciones que se registran, clasifican y almacenan, pero que no se organizan de acuerdo con ningún significado específico.
- Los elementos datos pueden ser numéricos, alfanumérico, figuras, sonidos o imágenes. Una base de datos está compuesta por elementos datos organizados para recuperarse.

1.3 ALGUNAS DEFINICIONES DE INFORMACIÓN

- Se obtienen una vez que los datos se procesan, agregan y presentan de la manera adecuada para que puedan ser útiles a alguien dentro de la organización, por lo que de este modo estos datos procesados y organizados presentan un mayor valor que en su estado original. Son "datos dotados de relevancia y propósito", que permiten reducir la incertidumbre de quien los recibe.
- Datos a los que se les ha dado una forma que tiene sentido y es útil para los humanos.
- La información corresponde a los datos que se han organizado de modo que tengan significado y valor para el receptor. Éste interpreta el significado y obtiene conclusiones e implicaciones. Los datos procesados por un programa de aplicación representan un uso más específico y un valor agregado más alto que la simple recuperación de una base de datos. Una de estas aplicaciones puede ser un sistema de administración de inventarios, un sistema de registro en línea universitario o un sistema de compra y venta de acciones basadas en Internet.

Según Idalberto Chiavenato:

"La Información es un conjunto de datos con un significado, que reduce la incertidumbre o que aumenta el conocimiento de algo. En verdad, la información es un mensaje con significado en un determinado contexto, disponible para uso inmediato y que proporciona orientación a las acciones por el hecho de reducir el margen de incertidumbre con respecto a nuestras decisiones".

Para Ferrell y Hirt:

"La información comprende los datos y conocimientos que se usan en la toma de decisiones"

Según Czinkota y Kotabe

"la información consiste en datos seleccionados y ordenados con un propósito específico"

SegunAlvin y Heidi Toffler, en su libro «La Revolución de la Riqueza» nos brindan la siguiente diferencia (muy entendible) entre lo que son los datos y lo que es información:

"Los datos suelen ser descritos como elementos discretos, huérfanos de contexto: por ejemplo, «300 acciones». Cuando los datos son contextualizados, se convierten en información: por ejemplo, «tenemos 300 acciones de la empresa farmacéutica X»"

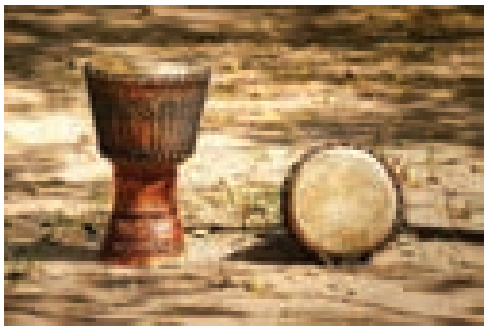
Según Wikipedia:

"Las informaciones "un conjunto organizado de datos procesados, que constituyen un mensaje sobre un determinado ente o fenómeno".

Según IRMA WASSALL (1943):

"Por todo el Continente Negro suenan los tambores que nunca callan: base de toda música, foco de toda danza;

Tambores parlantes, radiotelégrafo de la jungla inexplorada."



Desde hace más de 500 años los pobladores del África subsahariana contaban con una tecnología de información que sería la envidia de cualquier gobernante europeo. Eran los tambores. Sus rítmicas melodías eran en realidad un código tonal capaz de transmitir información detallada sobre distintos aspectos de una manera amplia, rápida y eficiente.

1.4 CONOCIMIENTO

- El conocimiento está compuesto por datos o información que se ha organizado y procesado para llevar entendimiento, experiencia, aprendizaje acumulado y pericia cuando se aplican a un problema o actividad presente. Los datos que se procesan para obtener implicaciones críticas y reflejar experiencia y pericia pasadas brindan al receptor conocimiento organizacional, de muy alto valor potencial.
- Una colección de información no es conocimiento.
- Mientras que la información entrega las asociaciones necesarias para entender los datos, el conocimiento provee el fundamento de cómo cambian (en el caso que lo hagan). Esto claramente puede verse como patrones de comportamiento contextuados, es decir una relación de relaciones. Representa el Como.
- Son las creencias cognitivas, confirmadas, experimentadas y contextuadas del conocedor sobre el objeto, las cuales estarán condicionadas por el entorno, y serán potenciadas y sistematizadas por las capacidades del conocedor, las cuales establecen las bases para la acción objetiva y la generación de valor.

1.5 SABIDURÍA

La sabiduría abarca los principios fundacionales responsables de los patrones que representan el conocimiento. Representa el porqué. Estos tres términos, en especial datos e información, se utilizan con frecuencia de modo intercambiable. Es posible que los datos, la información y el conocimiento sean entradas para un SI, aunque también pueden ser las salidas. Por ejemplo, los datos acerca de los empleados, sus salarios y el tiempo trabajado se procesan para producir la información de la nómina. Esta misma puede usarse después como una entrada en un sistema que elabora un presupuesto o asesora a la administración acerca de aumentos salariales.

1.6 LA IMPORTANCIA DE LA INFORMACIÓN EN LAS ORGANIZACIONES

Siempre se ha dicho que la información es poder. Pero, ¿por qué es tan importante la información? El éxito de una empresa no depende sólo de cómo maneja sus recursos materiales (trabajo, capital, energía, etc.), sino de cómo aprovecha sus "activos intangibles" (know-how, conocimiento del mercado, imagen de marca, fidelidad de los clientes, etc.), y que el correcto desarrollo de estos últimos depende de que exista un adecuado flujo de información entre la empresa y su entorno, por un lado, y entre las distintas unidades de la empresa, por otro.

1.7 QUE ES BIG DATA

"Big Data es un campo del conocimiento dedicado al análisis, procesamiento y almacenamiento de grandes colecciones de datos que con frecuencia se originan en fuentes dispares."

"El término 'Big Data' proviene originalmente del ámbito de las ciencias de la computación y ha sido típicamente empleado para referirse a sets de datos cuyo tamaño excede al que puede manejar el software y hardware estándares disponible para capturar, almacenar y analizarlos".

*"Big Data: es el conjunto de tecnologías, técnicas y herramientas que hacen posible la recogida, procesamiento y análisis de volúmenes masivos de datos, y también la visualización de los resultados. El propósito es convertir la información hallada en esos grandes conjuntos de datos en algo útil como estadísticas, patrones de comportamiento, análisis de rendimiento, etc.***"**

"Big data es información de alto volumen, alta velocidad y /o alta variedad de activos que demandan formas innovadoras y rentables de procesamiento de información que permitan una mejor comprensión, toma de decisiones y automatización de procesos."

"Herramientas y técnicas de información de alto volumen, alta velocidad y /o alta variedad de activos, que permitirán capturar datos, almacenar, analizar, y obtener estadísticas, patrones de comportamiento, análisis de rendimiento, que permitirán una mejor comprensión, toma de decisiones y automatización de procesos."

1.8 OTROS CONCEPTOS DE BIG DATA

Wikipedia: "Big Data es un término general para colecciones de datos tan grandes y complejas que son difíciles de procesar con el uso de herramientas de procesamiento de datos tradicionales.", (Wikipedia, 2009).

Microsoft: "Big Data es un término cada vez más utilizado para describir el proceso de aplicación de alta potencia de cómputo, machine Learning y de inteligencia artificial a información masiva y a menudo de gran complejidad.", (Microsoft, 2012).

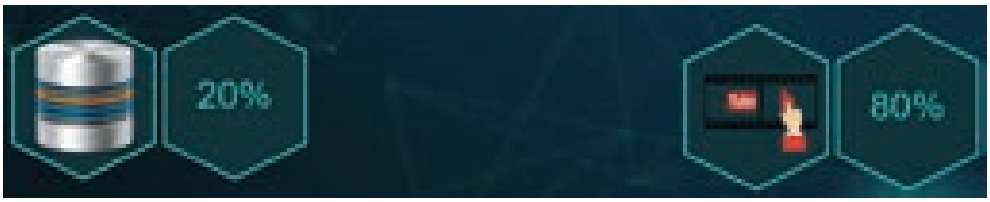
Mayer-Converger & Tucker: "Big Data se refiere a nuestra capacidad creciente de hacer cálculos a vastas colecciones de información, analizarla instantáneamente y sacar conclusiones profundas de ellas.", (Víctor Mayer-Converger, 2013).

IBM: "Big Data está siendo generado por todo lo que nos rodea en cada momento. Cada proceso digital e intercambio de medios sociales lo produce. Sistemas, sensores y dispositivos móviles lo transmiten. Big Data está llegando desde múltiples fuentes a una velocidad, volumen y variedad.", (IBM, 2014).

1.9 GENERACIÓN DE DATOS EN INTERNET EN TIEMPO REAL

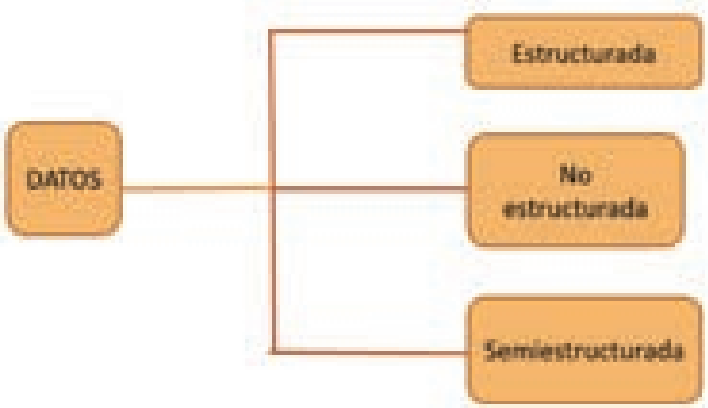
https://pdm.com.co/BlogPDM/wp-content/uploads/2014/06/Internet-en-tiempo-real.gif?x81790

1.10 TIPOS DE DATOS

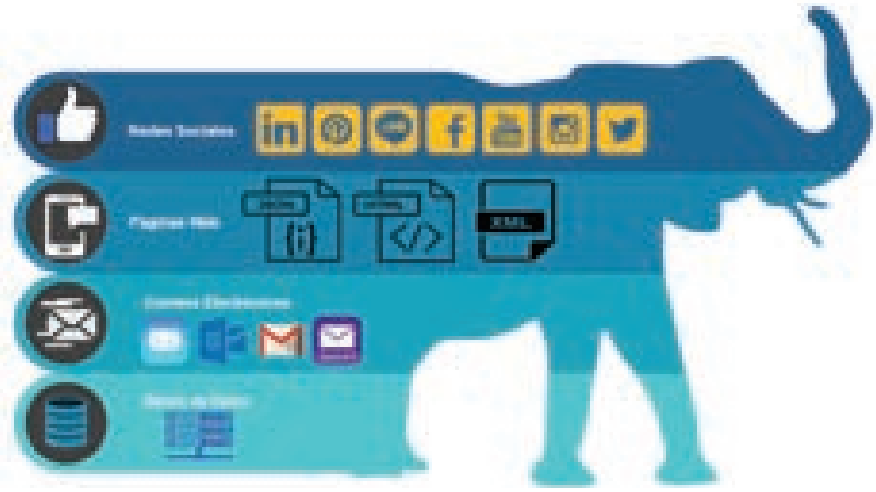


Es frecuente indicar la siguiente clasificación de los tipos de datos: estructurados (datos tradicionales) y no estructurados (datos Big Data). Sin embargo, las nuevas herramientas de manipulación de Big Data han originado nuevas categorías dentro de los tipos de datos no estructurados: datos semiestructurados y datos no estructurados propiamente dichos, que a continuación definimos:

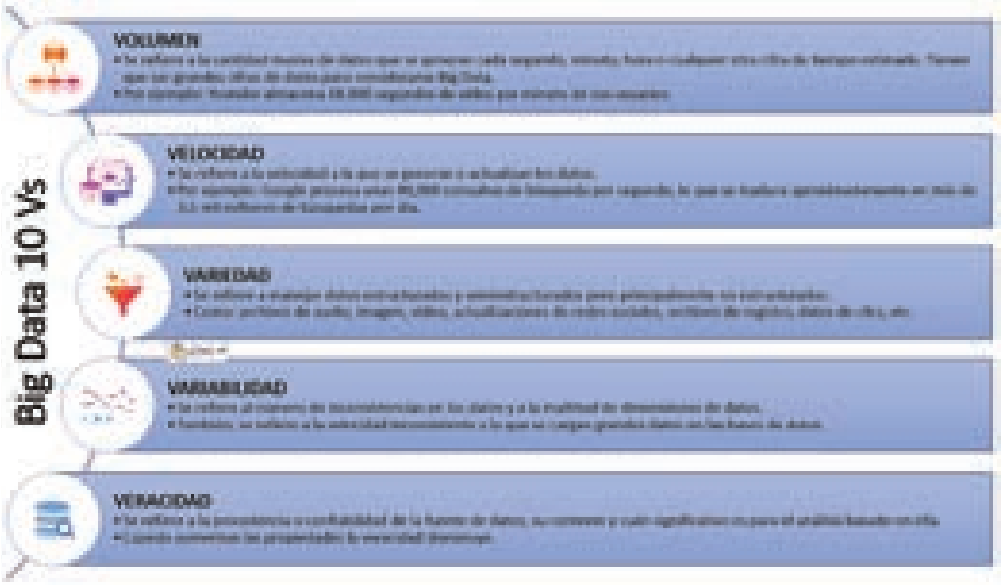
- Datos estructurados (Structured Data): datos que tienen bien definidos su longitud y su formato, como las fechas, los números o las cadenas de caracteres. Se almacenan en tablas. Un ejemplo son las bases de datos relacionales y las hojas de cálculo.
- Datos no estructurados (Unstructured Data): datos en el formato tal y como se recolectaron, carecen de un formato específico. No se pueden almacenar dentro de una tabla ya que no se puede desgranar su información a tipos básicos de datos. Algunos ejemplos son los PDF, documentos multimedia, e-mails o documentos de texto.
- Datos semiestructurados (Semistructured Data): datos que no se limitan a campos determinados, pero que contiene marcadores para separar los diferentes elementos.

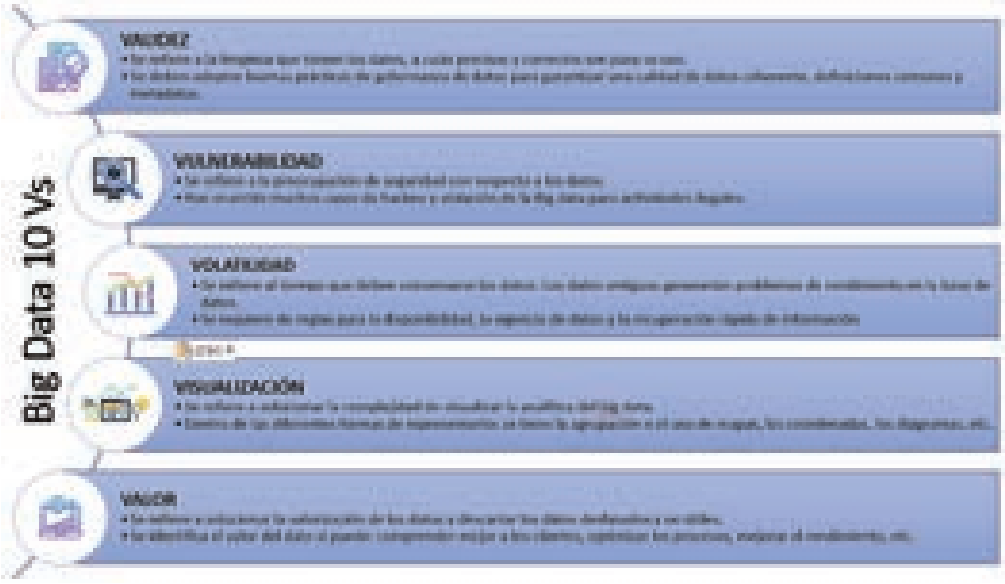


1.11 FUENTES DEL BIG DATA



1.12 LAS VS EN BIG DATA


Big Data 10 Vs
VOLUMEN
VELOCIDAD
VARIEDAD
VARIABILIDAD
VERACIDAD


Big Data 10 Vs
VALIDEZ
VULNERABILIDAD
VOLATILIDAD
VISUALIZACIÓN
VALOR

1.13 LAS VS A LO LARGO DEL TIEMPO

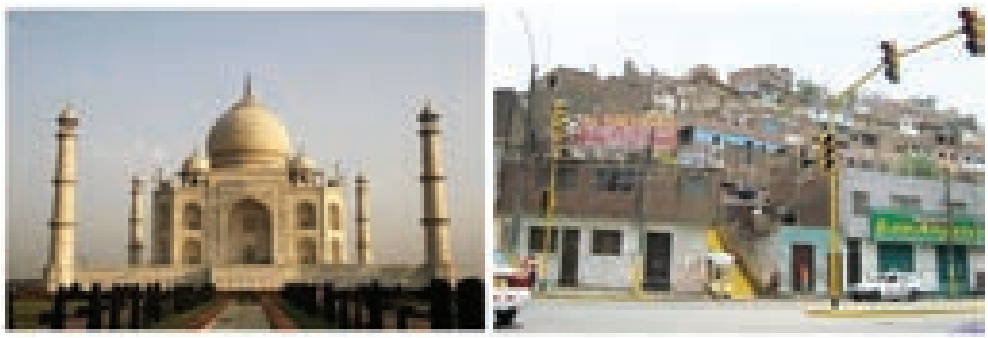
1.14 APLICACIONES DEL BIG DATA EN GENERAL

1.15 PATRONES DE ARQUITECTURA DE SOFTWARE

"La arquitectura de software de un programa o sistema, es la estructura o estructuras del sistema, lo cual incluye los elementos de software, las propiedades visibles externamente de esos elementos y las relaciones entre ellos."

Bass, Clement, Kazman: Software Architecture in Practice

1.15.1 Qué podría decir acerca de



1.15.2 Algunas Afirmaciones Importantes

- Los arquitectos de construcciones y los arquitectos de software enfrentan retos similares." Por lo mismo, es necesario que los arquitectos cuenten con una serie de conocimientos, habilidades y conocimientos especiales".
- Según Glenn Murcutt Arquitecto australiano famoso por ganar muchos premios y por ser presidente fundador de la Asociación de Arquitectura de Australia. "Necesitamos soluciones para problemas reales y no inventar problemas para justificar malas soluciones".

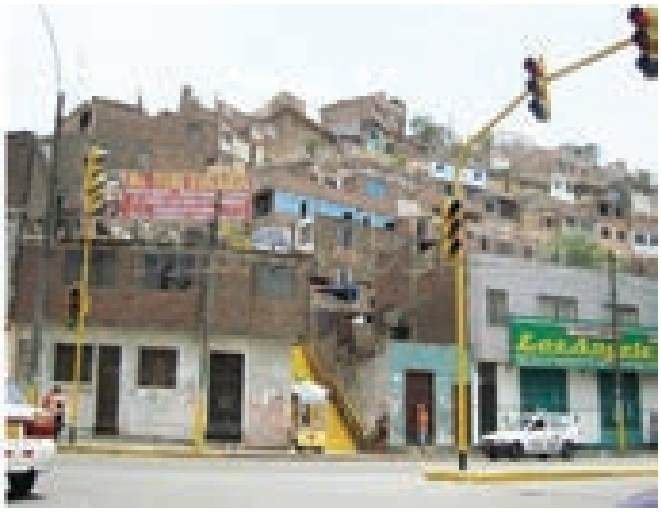
1.15.3 Conclusiones

El Taj Mahal es un monumento de amor.

Comprende un conjunto de bellos edificios planificados y exquisitamente construidos entre 1631 y 1654 en la ciudad de Agra, estado de Uttar Pradesh, India, a orillas del río Yamuna, por el emperador musulmán Shah Jahan de la dinastía mogola.

Se construyó en honor a la fallecida esposa favorita del emperador, Arjumand Bano Begum más conocida como Mumtaz Mahal.

Es considerada una de las maravillas del mundo



El Agustino es una zona populosa de Lima llena de construcciones realizadas en el cerro.

Muchas de ellas están hechas en adobe y sin acabados.

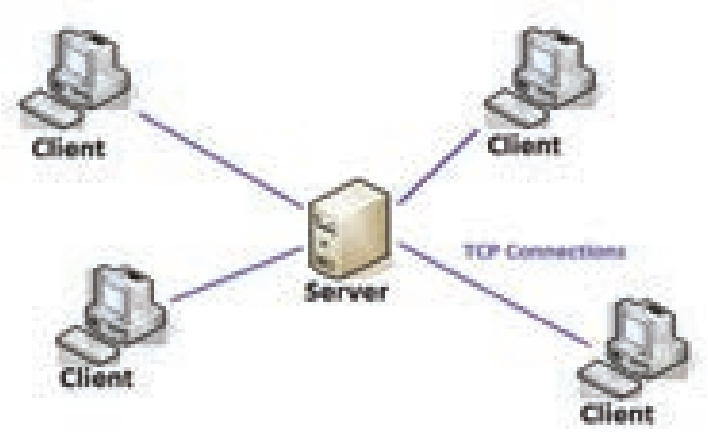
Es una de las muchas zonas de Lima que ha crecido sin planificación urbana.

- La arquitectura de software representa el estudio de alto nivel del sistema.
- Se representa a través de muchas vistas o estructuras del sistema.
- Cada vista se compone de elementos de software y de las relaciones entre ellos.
- Cada elemento de software tiene atributos visibles interna o externamente.
- Cada elemento debe estar claramente documentado para que se conozca su naturaleza.
- Forma parte de la disciplina de análisis y diseño.

1.15.4 Aplicaciones monolíticas

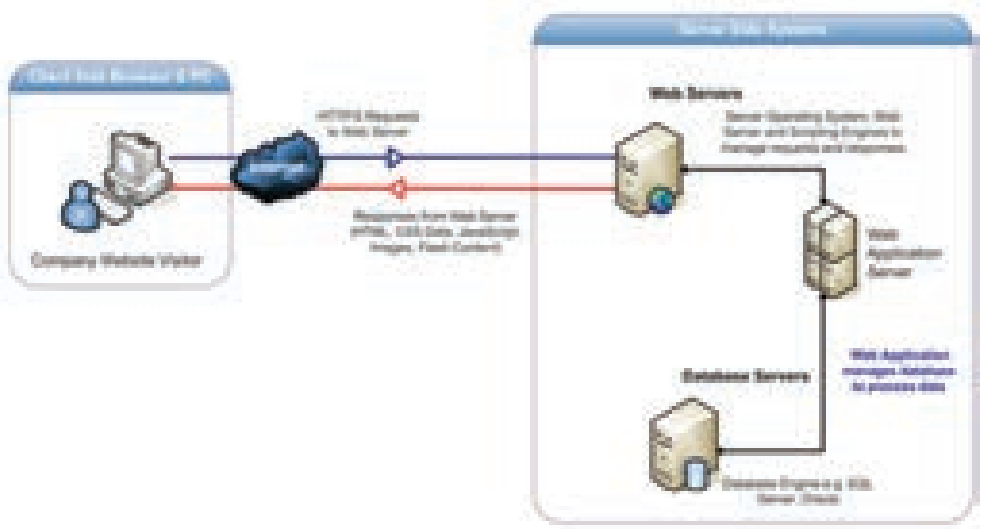
- Diseñadas para un usuario único en una sola máquina.
- Explotan la capacidad de la máquina al máximo.
- La seguridad no es una prioridad.
- Existe una dificultad para mejorar la aplicación porque igualmente debe actualizarse cada máquina.
- Aplicaciones tipo: editores de texto, hojas de cálculo, sistemas embebidos en máquinas especializadas como por ejemplo calculadoras, máquinas expendedoras de bebidas y alimentos.
- Lenguajes de programación usados: C/C++

1.15.5 Aplicaciones Cliente Servidor



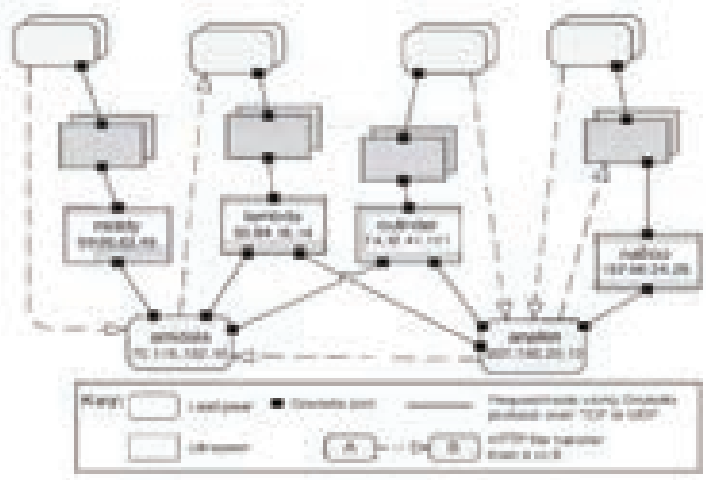
- El servidor se diseñó para múltiples usuarios. Cada cliente representa a un usuario único.
- Es necesario tener un protocolo propietario para la comunicación entre el cliente y el servidor.
- La seguridad en estas aplicaciones es una prioridad.
- En sus versiones iniciales, el servidor era fácil de mejorar, pero el upgrade de un cliente no, debido a que cada máquina tenía que modernizarse.
- Es necesaria la administración de una sesión de trabajo. El cliente es el responsable de administrar el inicio y fin de la sesión.
- Lenguajes de programación usados: C/C++/VB, SQL (Stored procedures).

1.15.6 Aplicaciones Web



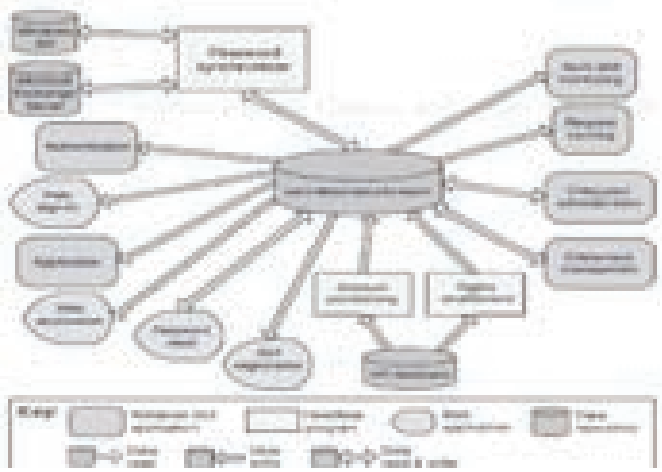
- Tanto el cliente como el servidor se diseñaron para múltiples usuarios.
- La seguridad es de alta prioridad.
- Se usa un protocolo estándar para la comunicación entre el cliente y el servidor.
- El manejo de la sesión es complejo porque afecta el desempeño del sistema. No se puede operar sobre muchos objetos al mismo tiempo y la administración de la sesión la hace el servidor con muy poca ayuda del cliente (que usa cookies).
- Se incrementa la importancia del middleware.
- Lenguajes de programación usados: HTML, JavaScript, VB.Net or Java y SQL para la comunicación con la base de datos.

1.15.7 Aplicaciones Peer-to-Peer



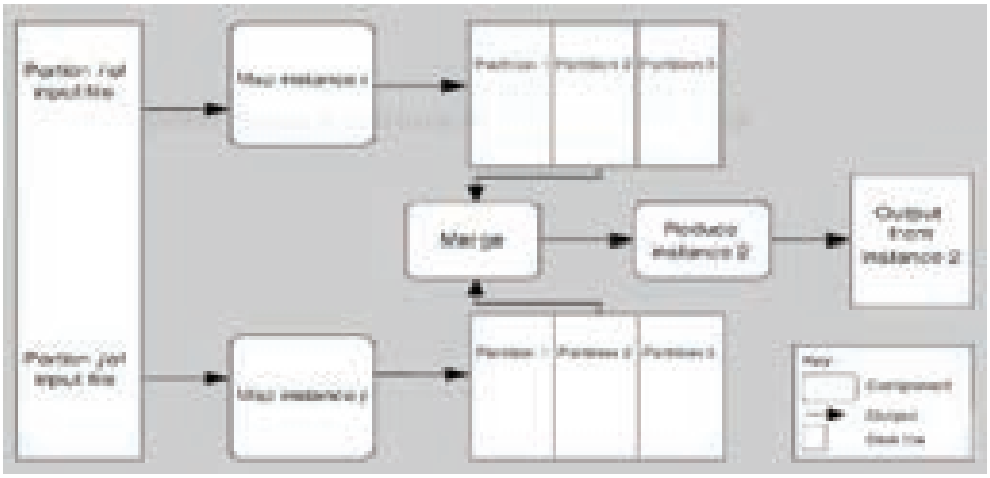
- Los componentes interactúan como peers y son todos "iguales".
- Estilo arquitectural petición- respuesta sin asimetrías.
- Proveer una conexión bidireccional.
- Algunas arquitecturas pueden tener super-nodos (indexación o routear).
- Los peers pueden agregarse o retirados sin gran impacto en el sistema en general. Por tanto, se logra escalabilidad.
- Bitorrent, Skype, VoIP.

1.15.8 Aplicaciones de Data Compartida



- Está formado por múltiples "data accesors" y al menos un "shared-data store".
- El acceso puede ser de lectura.
- Escritura o ambos.
- Gestiona la concurrencia.
- Podría generar cuellos de botella, y de contarse con un solo "shared-data store", ser riesgoso en términos de alta disponibilidad.
- Restricción: los "shared-data store" no puede interactuar directamente.

1.15.9 Aplicaciones Map-Reduce

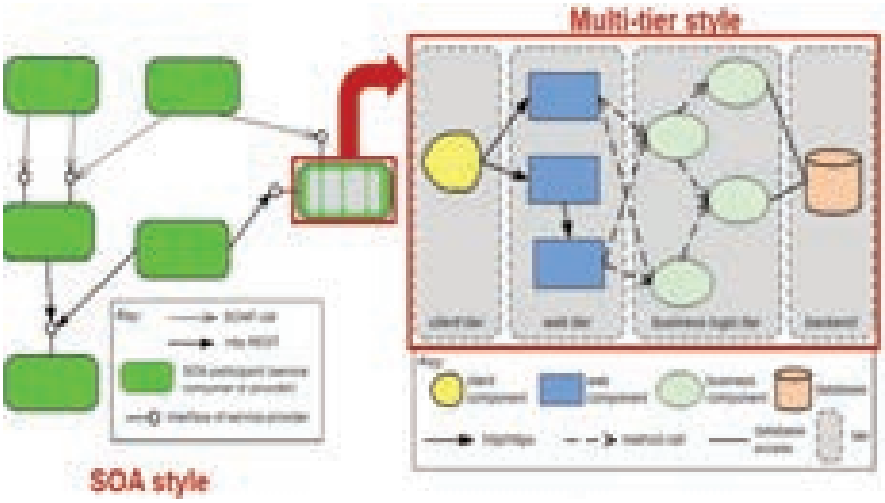


Se compone de tres partes:

- Infraestructura: donde se deploya el map y se reduce las instancias. Puede ser un nodo o multiple nodos en programación paralela.
- Map: una función con múltiples instancias deployado en la infraestructura.
- Reduce: función para procesar la porción de extraer-transformar-cargar.

Bueno para mejorar la eficiencia en una arquitectura distribuida.

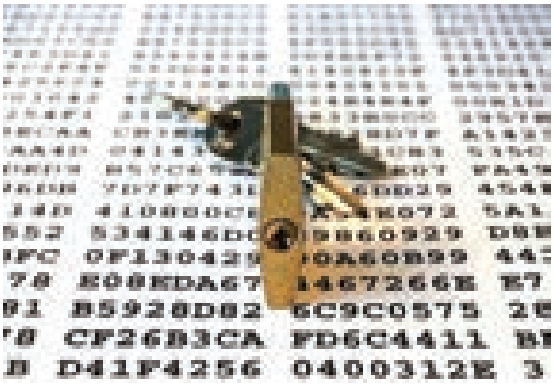
1.15.10 Aplicaciones heterogéneas



1.16 RETOS DE SEGURIDAD EN BIG DATA

- Falta de seguridad en el diseño de la solución: debido a las diferentes plataformas utilizadas, no siempre cuentan con un cifrado de datos, gestión de políticas.
- Anominización: debemos proteger la información personal y evitar que alguien pueda hacer un análisis profundo.
- Complejidad y diversidad de datos: debido a la variedad de fuentes que posee, se debe de proporcionar protección a cada tipo de fuente distinta con una característica adaptada.
- Poca inversión en seguridad: muchas empresas por ahorro de costos, evitan implementar seguridad a estos volúmenes de información.
- Falta de habilidades: al ser nuevas tecnologías, se debe de capacitar a los empleados para poder manejar de forma óptima estos grandes proyectos.
- Ruptura de datos: debemos verificar que la información que poseemos sea confiable ya que al definir una toma de decisiones es fundamental que la información sea cierta.
- Información sensible: muchas empresas recopilan informaciones sensibles del usuario como las búsquedas realizadas, ubicación geográfica, relaciones de amistades.

1.17 MEDIDAS DE SEGURIDAD BÁSICAS



El proceso de anonimizar los datos consiste en eliminar todo rastro de información personal identificable de un usuario. Aunque las organizaciones deben eliminar todo tipo de información identificable de usuario, esto es una tarea difícil.

Cifrado de datos

El manejo de grandes volúmenes de datos requiere que las organizaciones pongan en práctica todas las medidas que sean necesarias para asegurar su confidencialidad.

Control de acceso y monitorización

Otro de los aspectos claves en Big Data reside en la aplicación de mecanismos de seguridad que controlen el acceso a la información manipulada en los sistemas Big Data.

1.18 USO DEL BIG DATA EN LA CIBERSEGURIDAD

La digitalización ha transformado la estructura de la información global y no solo la periodística, sino cualquier fuente de datos y documentación. La proliferación de plataformas, dispositivos y contenidos ha multiplicado la oferta de contenidos disponibles, al mismo tiempo que la digitalización ha abaratado los costes de creación, producción, distribución y almacenamiento de información periodística, propaganda y contenido de cualquier naturaleza. Oferta y demanda se han retroalimentado hasta el punto de crear un mercado de la propaganda rentable por su audiencia global y la participación activa de bots, trols y otros autómatas.

Bradshaw y Howard

Las empresas al ser vulnerables y constantemente atacadas, ha sido necesario implementar herramientas que analicen y gestiones en tiempo real. Gracias al big data y la información recolectada por los ataques, es posible predecir o saber cómo responder a un ataque en tiempo real, analizando los patrones o tendencias de comportamiento.

Tanto el big data y la ciberseguridad se complementan:

1. Monitorización en tiempo real.
2. Ofensiva.
3. Precursores para el futuro.
4. Protección y seguridad para todos.
5. Detección de intrusión física en grandes espacios o infraestructuras abiertas.
6. Computación sobre información cifrada.
7. Análisis automático de vulnerabilidades de red (máquinas-tráfico de datos).
8. Criminología computacional.
9. Uso fraudulento de recursos corporativos y/o sensibles.
10. Análisis de video en tiempo real / Búsqueda y recuperación rápida en librerías de video.
11. Inteligencia visual en máquinas.
12. Identificación de anomalías, patrones y comportamiento en grandes volúmenes de datos.
13. Análisis de texto (estructurado y no estructurado) como apoyo a la toma de decisión en tiempo real en entornos intensivos en datos.
14. Consciencia situacional.
15. Traducción automática a gran escala (en número de idiomas y en volumen).
16. Predicción de eventos.

1.19 METODOLOGÍAS DE PROCESAMIENTO DE GRANDES VOLÚMENES DE DATOS

La metodología de procesamiento de grandes volúmenes de datos, que permite transformar los datos en conocimiento posee ocho fases que a continuación describiremos:

1.19.1 Entendimiento del Negocio

Debemos conocer el negocio si queremos resolver un problema, alcanzar un objetivo, etc. Por ejemplo, deseamos entender por qué mis clientes dejaron de comprar nuestros productos para ello deberemos realizar las siguientes preguntas:

- Pasado: ¿Qué ocurrió para que ello pasara-Por qué ocurrió?
- Presente: ¿Qué ocurre actualmente-Qué debo de hacer?
- Futuro: ¿Qué ocurrirá-Qué debería hacer para mitigarlo?

Luego es necesario también identificar a los stakeholder o involucrados que podrían ser de las siguientes áreas:

- Tecnología.
- Analítica.
- Negocio.

Al final es necesario definir el tipo de analítica que vamos a utilizar:

- Analítica descriptiva. Es aquella que describe los hechos ocurridos en el pasado.
- Analítica diagnóstica. Es aquella que contrasta los datos históricos con otros datos para responder a las interrogantes.

- Analítica predictiva. Determinar datos en el futuro a través de datos históricos.
- Analítica prescriptiva. Es una evolución de la analítica predictiva cuyo propósito es determinar cuáles acciones se deben tomar para sacar el mayor provecho a una tendencia favorable o evitar un problema a futuro.

Para entender el negocio es recomendable diseñar un modelo de dominio en la cual representaremos los flujos de datos en la cual se caracterizará por:

- Representación del negocio.
- Jerárquica.
- Realidad.
- Sin atributos o características.
- Sin métodos.
- Punto de partida para la elicitación de requerimientos.

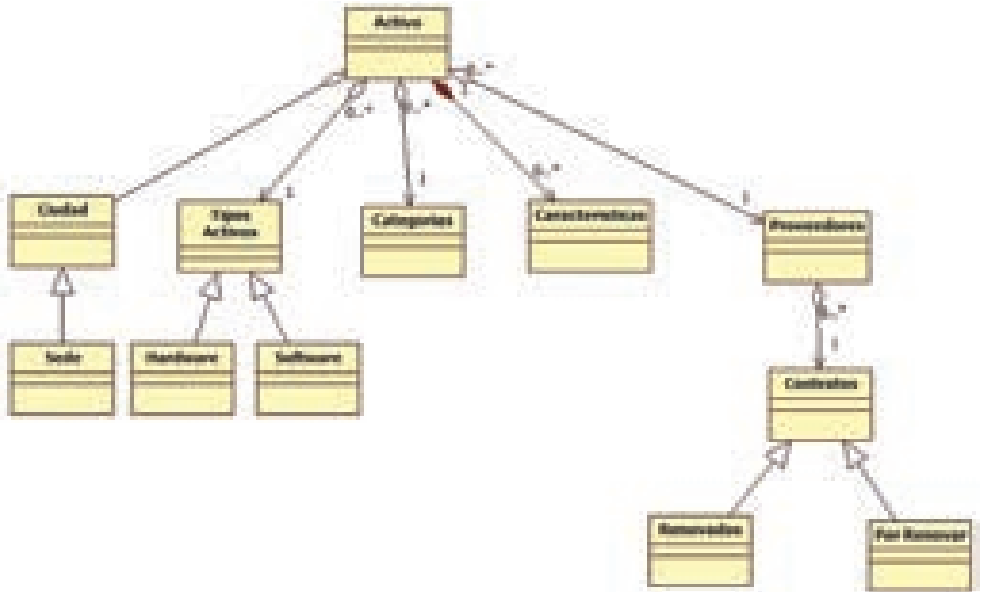
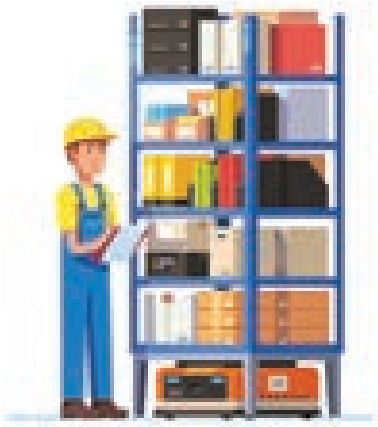


Figura. Mapa de dominio de datos

1.19.2 Comprensión de datos

En esta fase debemos identificar las fuentes de información y su conveniencia necesaria de las mismas para su posterior captura y almacenamiento, las etapas de la comprensión son:

- Inventario de información. Cuál es la información adecuada que necesitamos. Para ello debemos hacer un listado de la información que debemos tener para ello es necesario apoyarnos de la gente de negocio.



- Fuentes de información. Es importante saber dónde se encuentra esa información o donde podríamos localizarla. Se trata de fuentes Internas o fuentes Externas.

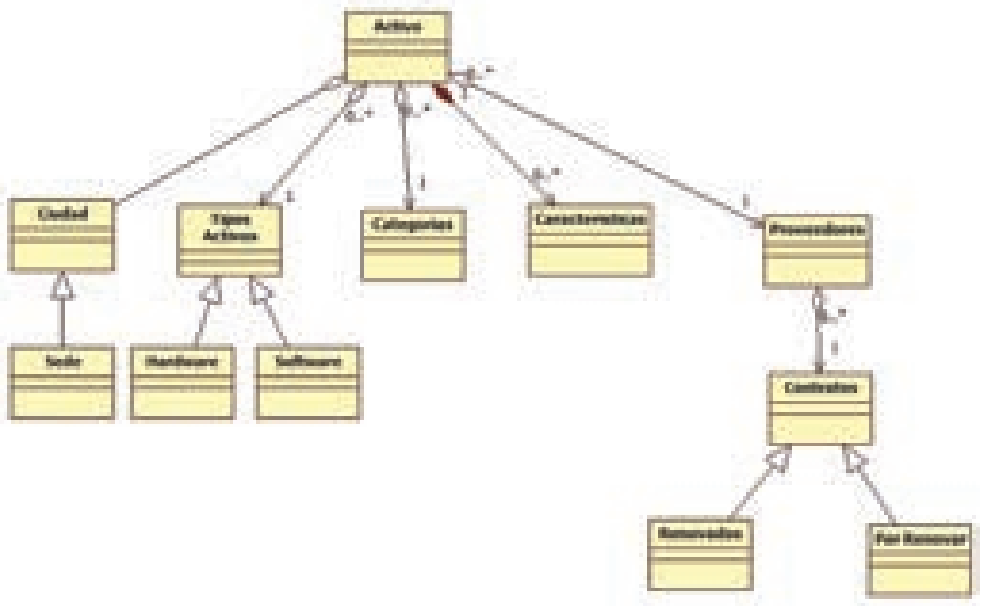
- Disponibilidad. La fuente identificada se encuentra disponible, podemos capturarla y almacenarla o si podemos adquirir o acceder a esas fuentes de datos.

Si podemos capturarla y almacenarla o si ya la tenemos capturada y almacenada perfecto, si no podemos actualmente por problemas técnicos, tecnológicos o

por alguna otra causa lo que tendremos que hacer es un plan de adquisición de fuentes para que a futuro tengamos disponible toda esa información para enriquecer nuestros análisis.

- Relaciones. Como podemos relacionar todos estos datos sea fuentes estructuradas, no estructuradas y Semi estructurada.

- Representación funcional de los datos. Como se relacionan estos datos (mapa de dominio).



1.19.3 Tecnología

Qué tecnología debemos utilizar para tratar nuestros datos en consecuencia para la construcción de nuestro modelo analítico, esta fase contiene tres etapas:

Diseño de la arquitectura. Definir los componentes necesarios para brindar soporte a la construcción del modelo analítico a construir.

Figura. Plataforma Tecnológica big data landscape.

Ecosistema de big data 2022:

Fuente: https://modern-cdo.medium.com/2022-technology-landscape-to-democratize-data-272b8228b6cb

Implementación. Es importante definir la estrategia de implantación. Para ello debemos determinar si queremos tener todo el control de toda la infraestructura, aplicaciones, datos, o queremos utilizar soluciones Cloud que nos permitan externalizar y que un proveedor nos puede dar prácticamente todos los servicios. Entonces, tenemos un amplio rango de posibilidades, desde On-premise hasta SaaS, "software as a service", pasando por infraestructura como servicio o plataforma como servicio.

Fuente de la imagen: bmc.com

1.19.4 Tratamiento de datos

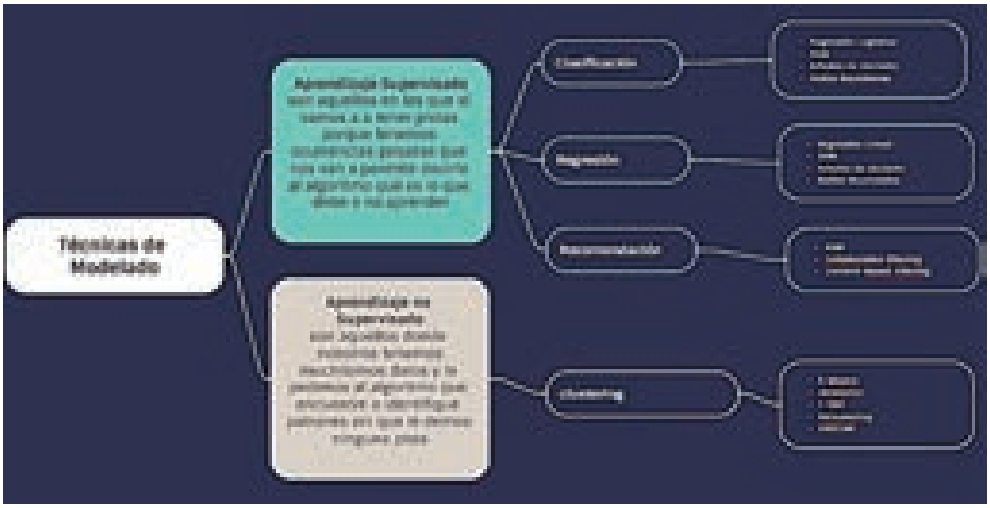
Esta etapa es una de las más importantes y es en la que el científico de datos ocupa la mayor cantidad de tiempo tratando la información, porque a diferencia de la información estructurada está en formato de tablas y es la que normalmente solemos trabajar pero actualmente el 80% de la información viene desestructurada, puede ser :vídeo, texto, voz y esa información además no está relacionada y normalmente la calidad no suele ser la esperada por lo que requerimos un trabajo y un esfuerzo muy grande en ordenar toda esa información, darle un formato adecuado para que esté lista.

Esta etapa la podemos dividir en:

- Adquisición y Registros.
- Metadatos.
- Exploración y Análisis.
- Calidad de Datos y Limpieza.

1.19.5 Modelización

En esta etapa construiremos un modelo analítico, para ello debemos elegir que técnica de modelado vamos a utilizar:



1.19.6 Presentación

Esta fase consiste en trasladar el conocimiento al resto de stakeholders. Es muy importante elegir la manera en la cual se presentan estos conocimientos, entre los que podemos utilizar:

- Informes.
- Reportes.
- Visualizaciones.
- Infografías.
- Cuadros de mando.

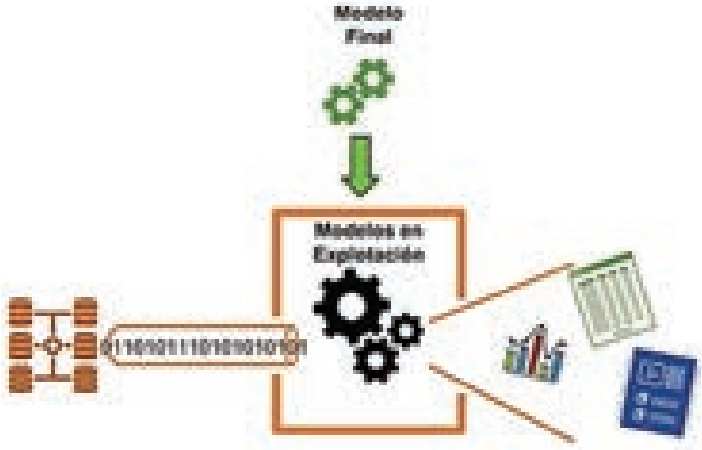
NOTA

Esta fase se refiere al proceso de creación de representaciones gráficas de información, generalmente mediante el uso de una o más herramientas de visualización.

1.19.7 Despliegue

Ahora debemos desplegar el modelo en la plataforma tecnológica. Esta fase consiste en:

- Integración en la plataforma. Lo primero es integrar en la arquitectura la tecnología de la organización, para eso nosotros tendremos un proceso donde tendremos los datos que se van ingestando, habrá modelos en explotación que nos van dando los resultados que pueden ser informes, gráficos, predicciones, etcétera.



- Planificación temporal. Lo siguiente es establecer la planificación temporal para explotar este modelo necesitamos datos y esos datos tienen una cadencia de producción o una cadencia de captura y almacenamiento probablemente distinta, tendremos que ver :
  - Cuándo se capturan los datos.
  - Cómo se capturan los datos.
  - Cómo vamos a utilizar este modelo con aplicaciones.

- Integración con las aplicaciones. Hay que pensar muy bien cómo vamos a integrar estos modelos en las aplicaciones y eso normalmente requerirá desarrollos en otros lenguajes de programación o en otras plataformas.

1.19.8 Puesta en Valor

Ahora debemos integrar el modelo dentro de las operaciones para ello podemos utilizarlos en:

- Tomar decisiones. Las decisiones se toman para resolver problemas. Al resolver un problema, la persona encargada de hacerlo puede tomar muchas decisiones. Las decisiones son cursos de acción que se toman para:
  - evitar o reducir los efectos negativos,
  - para aprovechar oportunidades,
  - eliminar las debilidades, potenciar y usar las fortalezas.

NOTA

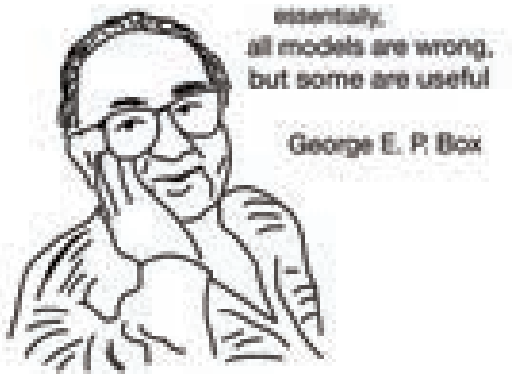
Según Herbert Simon las decisiones existen en un continuo, con las decisiones programadas en un extremo y las no programadas en el otro:

- Las decisiones programadas son "repetitivas (recurrentes) y rutinarias, en la medida en que se ha establecido un procedimiento definido para manejarlas, y así no tener que tratarlas de nuevo (como nuevas) cada vez que ocurren".
- Las decisiones no programadas son "novedosas (no recurrentes), no estructuradas e inusitadamente importantes. No existe un método de recetario para manejar el problema porque no ha surgido antes, o porque su naturaleza y estructura precisas son elusivas o complejas, o porque es tan importante que requiere un tratamiento a la medida".
- Campañas periódicas. Es decir, ese conocimiento que vamos obteniendo de forma periódica conforme tenemos más datos, podemos generar una serie de acciones proactivas por nuestra parte, basadas en ese conocimiento, y después medir cuáles han sido sus resultados, cuáles han funcionado mejor, cuáles han funcionado peor, para que nos sirva de input para ir optimizando esas acciones o esas campañas que nos permitan mejorar u optimizar nuestros procesos.

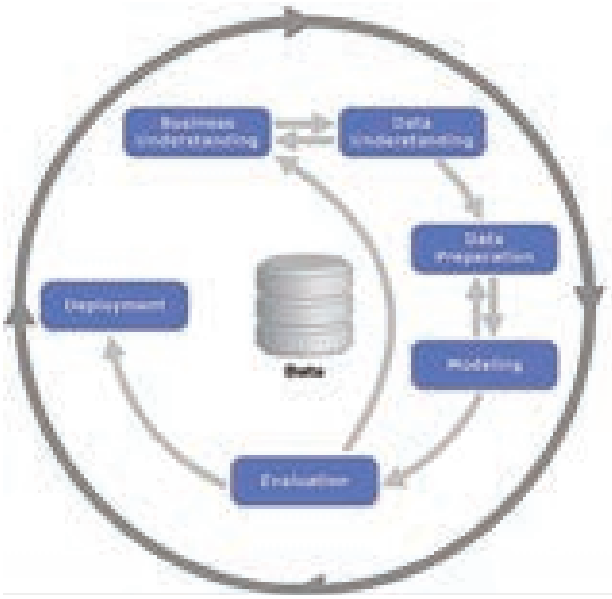
- Decisiones autónomo. Nosotros podemos utilizar el modelo y en base a ese modelo y unos criterios que establezcamos, podemos hacer que tome esas decisiones autónomas.

Es importante resaltar lo que expresa George Box:

> *"Todos los modelos son erróneos, pero algunos son útiles. Es importante tener claro que los modelos no aciertan siempre. Los modelos tienen incertidumbre, porque son predicciones a futuro y el futuro es incierto".*



1.20 METODOLOGÍAS CRISP-DM



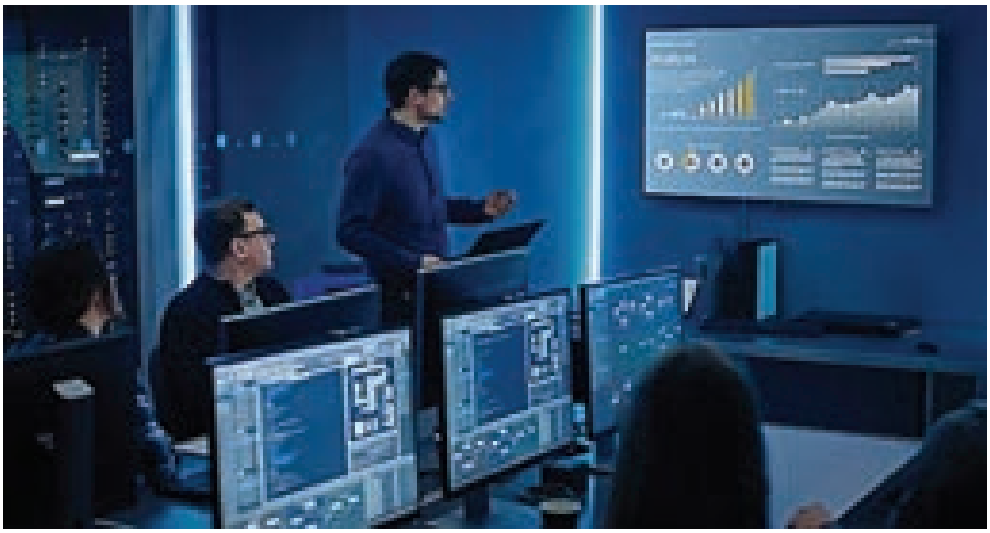
La metodología CRISP-DM (Cross-Industria Standard Procesos foro Data Mining) integra todas las tareas necesarias en los proyectos de minería de datos, desde la fase de comprensión del problema hasta la puesta en producción de sistemas automatizados analíticos, predictivos y/o prospectivos.

CRISP-DM está compuesta por seis fases:

1.20.1 Comprensión del negocio

- Entendimiento de los objetivos y requerimientos del proyecto.
- Definición del problema de Minería de Datos.

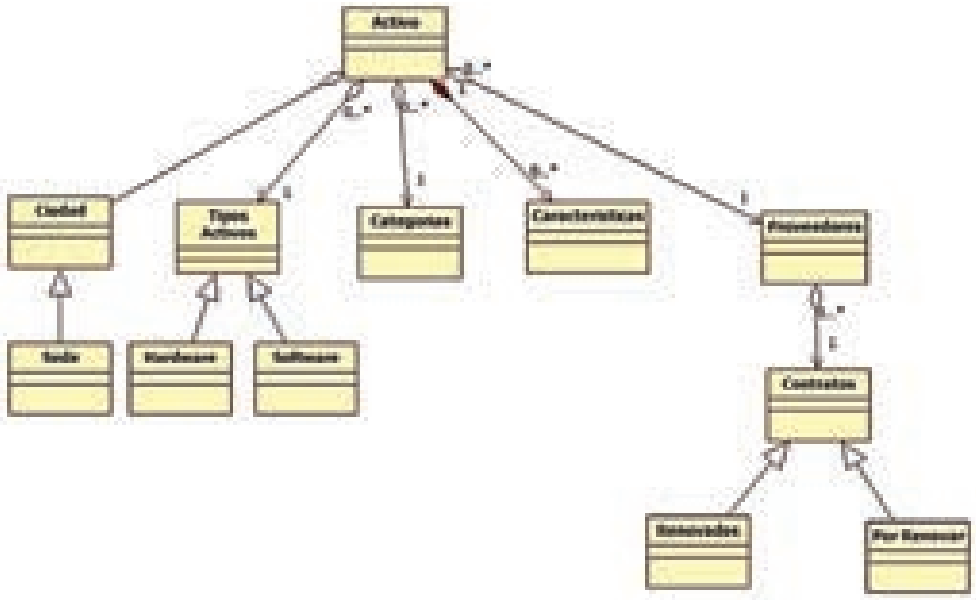
"Definir si el proyecto empieza en relación a la data master o caso en uso describa el problema a resolver"



1.20.2 Comprensión de los datos

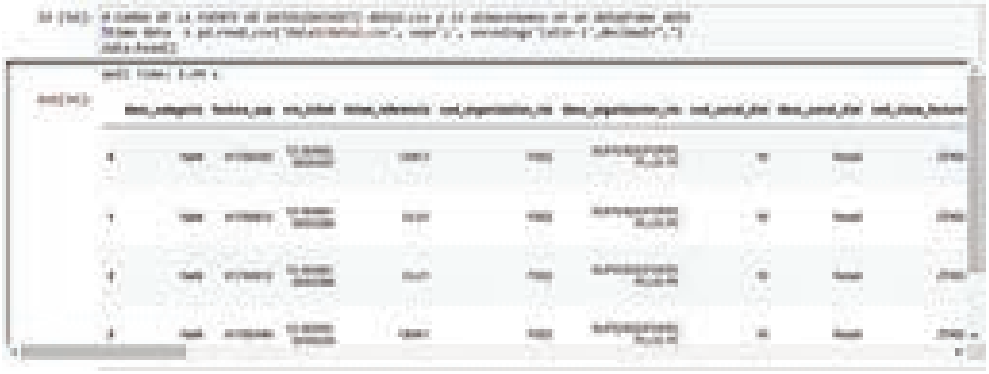
- Obtención conjunto inicial de datos.
- Exploración del conjunto de datos.
- Identificar las características de calidad de los datos.
- Identificar los resultados iniciales obvios.

"Comprensión de los datos, Recomendamos crear un modelo de dominio de datos"



1.20.3 Preparación de Datos

- Selección de datos.
- Limpieza de datos.



```
#deshacerse de los valores infinitos
data.replace([np.inf, -np.inf], np.nan, inplace=True)
#deshdgase de los valores
data.fillna(999, inplace=True)
```

NOTA

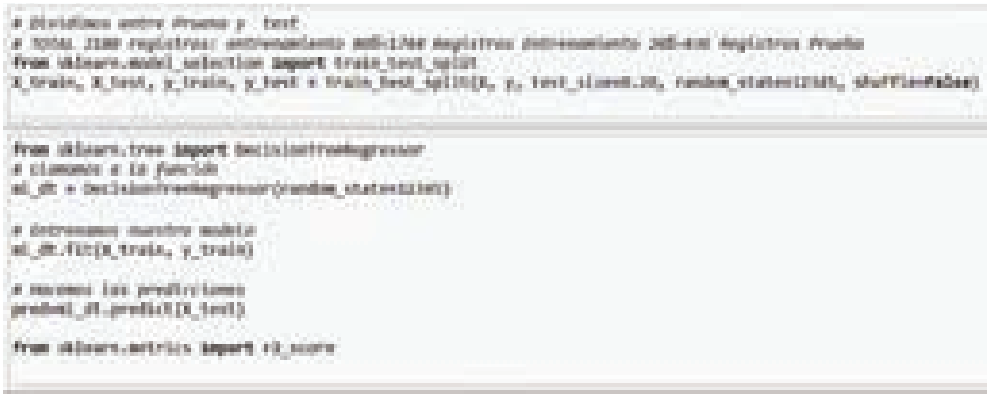
Formatos para comparar. Vamos a considerar los siguientes formatos para almacenar nuestros datos.

- CSV de texto sin formato: un buen amigo de un científico de datos.
- Pickle: una forma de Python de serializar cosas.
- MessagePack: es como JSON pero rápido y pequeño.
- HDF5: un formato de archivo diseñado para almacenar y organizar grandes cantidades de datos.
- Feather: un formato de archivo binario rápido, liviano y fácil de usar para almacenar marcos de datos.
- Parquet: formato de almacenamiento en columnas de Apache Hadoop.

1.20.4 Modelamiento

- Implementación en herramientas de Minería de Datos.

"Aquí se aplican técnicas de aprendizaje automático para crear un modelo que permita responder a las preguntas planteadas en la fase de comprensión del problema."



1.20.5 Evaluación

- Determinar si los resultados coinciden con los objetivos del negocio.
- Identificar los temas de negocio que deberían haberse abordado.

"en esta fase se evalúa la calidad del modelo y se comprueba si responde a las preguntas planteadas en la fase de comprensión del problema"

```
predexp = exp(pred)
ytestexp = exp(y_test)

test2=pd.DataFrame(ytest)
test2["pred"]=pred

test2["diferencia"]=abs(test2["valor_neto_total"]-test2["pred"])

test2
```

fec_factura	valor_neto_total	pred	diferencia
2021-12-10	116058.60	120794.44	4236.84
2021-12-09	114258.35	120794.44	6539.09
2021-12-01	107552.33	100966.55	6585.78
2021-12-02	101122.04	118629.11	17507.07
2021-12-03	108100.70	108629.11	528.41
...	...	...	...
2023-10-02	17567.63	17567.99	0.36
2023-11-01	23494.45	23503.77	36.32
2023-11-02	34548.44	35278.52	326.18
2023-12-01	40169.98	40152.46	57.40

1.20.6 Despliegue

- Instalar los modelos resultantes en la práctica.
- Configuración para minería de datos de forma repetida o continua.

"en la fase final se implementa el modelo en el entorno de producción y se proporciona a los usuarios las herramientas necesarias para utilizarlo de forma efectiva o para hacer que lleguen los datos."

1.21 HADOOP