

Capítulo 1 - Fundamentos de Big Data

Jaime Veintimilla Reyes

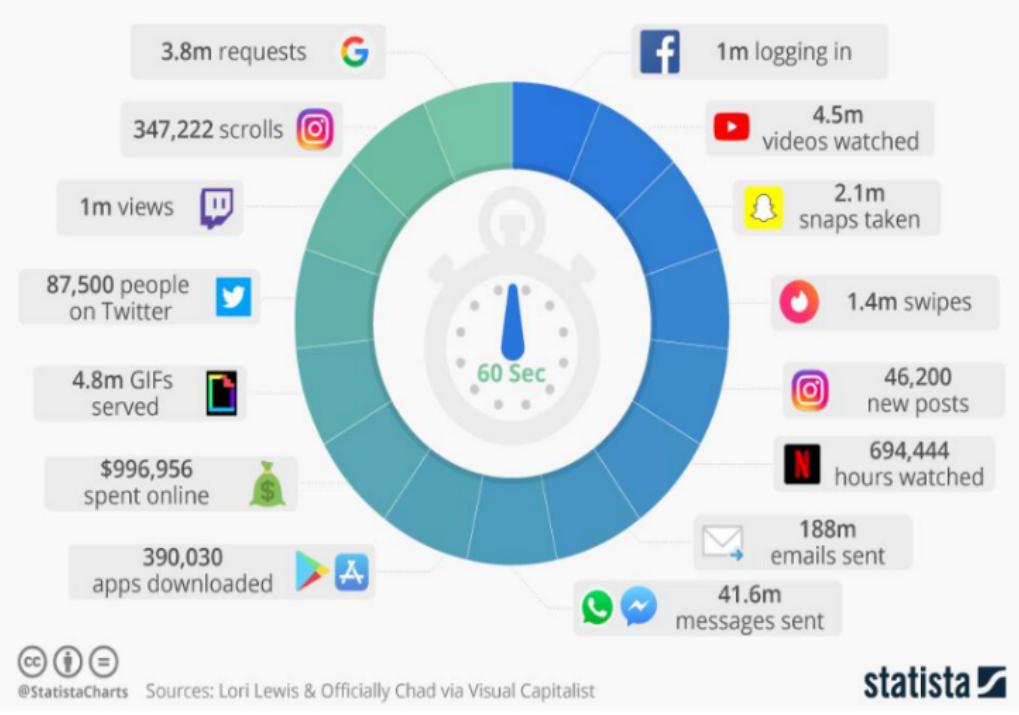
October 29, 2020

- 1 Introducción
- 2 Campo de aplicación de Big Data
- 3 Las 5 Vs de Big Data
- 4 Causas para la adopción de Big Data
- 5 Características deseadas de un sistema de Big Data
- 6 Soluciones de Big Data
- 7 El Teorema CAP
- 8 Preguntas

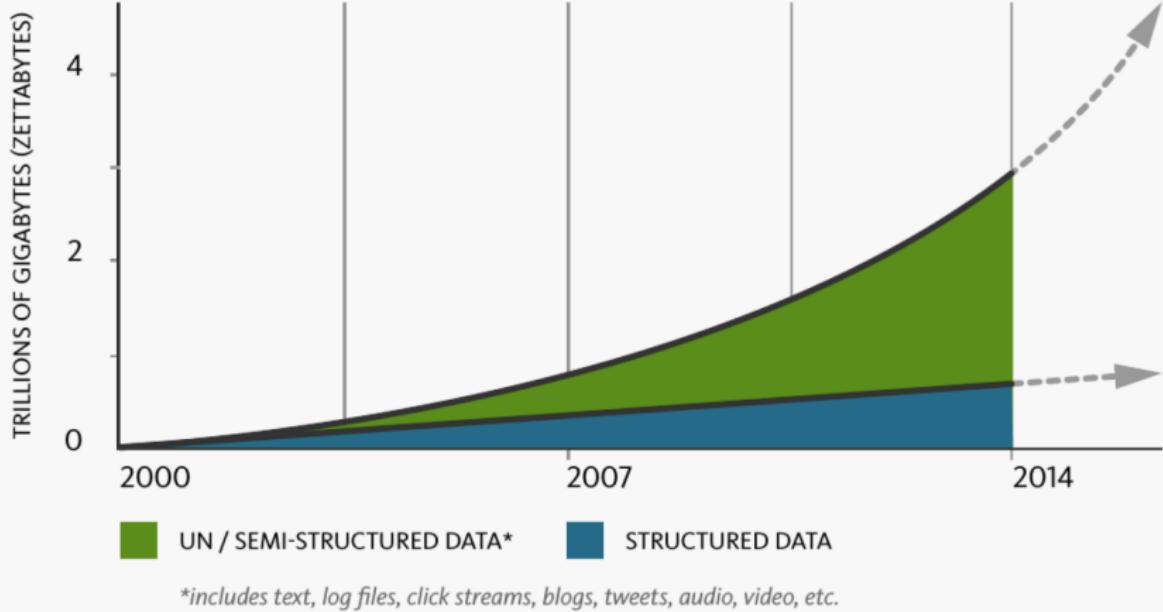
¿Qué es Big Data?

Un minuto en Internet 2020

Estimated data created on the internet in one minute

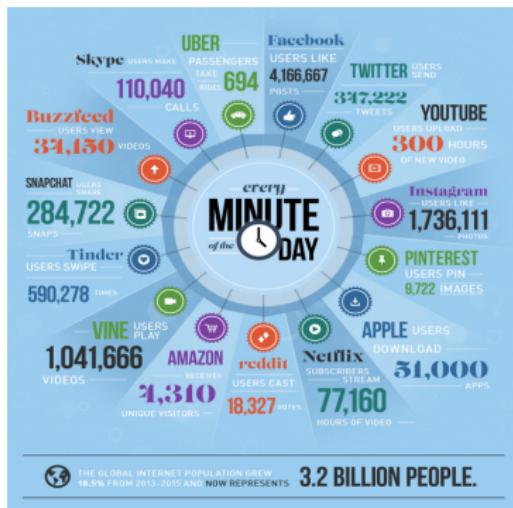


¿Qué es Big Data?



¿Qué es Big Data?

- 30 000 GB se generan cada segundo.
- FB: 1.4 millones usuarios activos cada mes.
- 2.5 billones de contenido
- 500+ terabytes diarios
- 10 000 sensores



¿Qué es Big Data?

- Técnicas tradicionales de análisis, procesamiento y almacenamiento no son suficientes.
- Big Data se dedica al análisis, procesamiento y almacenamiento de una gran cantidad de datos provenientes de fuentes heterogéneas

¿Qué es Big Data?

- **Big data** Consists of datasets that grow so large that they become awkward to work with using on-hand DB Management tools (Wikipedia).
- **Big data** is when the size of the data itself becomes part of the problem (Mike Lukides, O'Reilly Radar)
- It's not just your “**Big Data**” problems, it's all about your BIG “data” Problems (Alexander Stojanovic, Hadoop Manager on Win Azure)

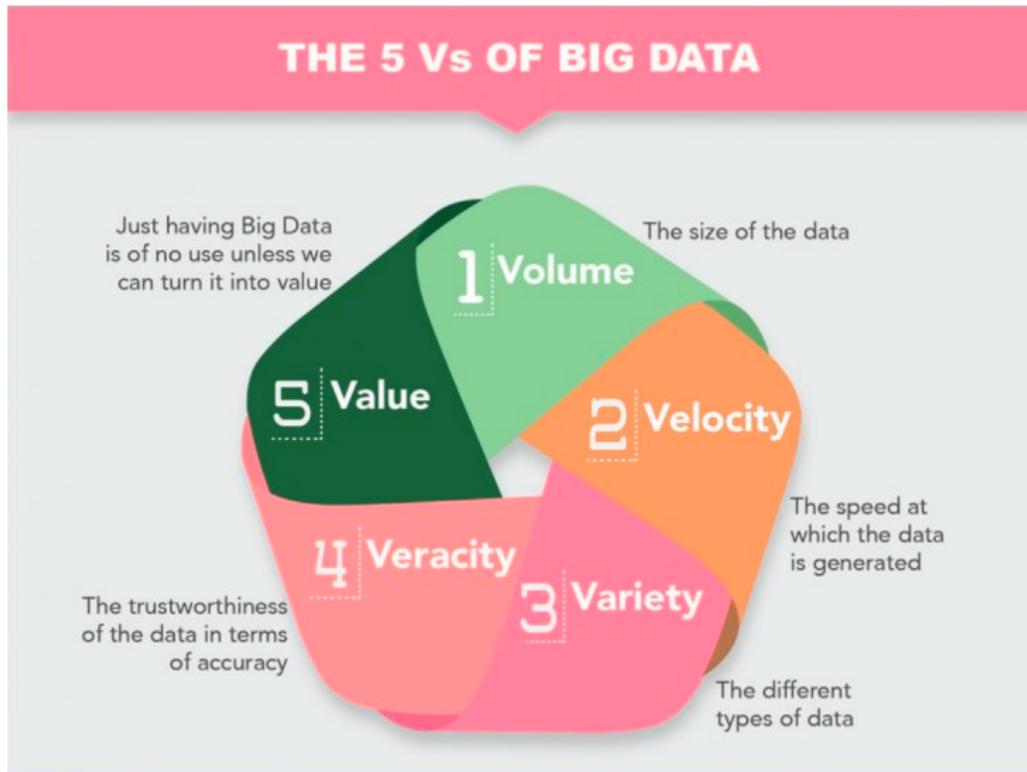
- 1 Introducción
- 2 Campo de aplicación de Big Data
- 3 Las 5 Vs de Big Data
- 4 Causas para la adopción de Big Data
- 5 Características deseadas de un sistema de Big Data
- 6 Soluciones de Big Data
- 7 El Teorema CAP
- 8 Preguntas

Campo de aplicación de Big Data

- Optimización de operaciones
- Identificación de nuevos mercados
- Predicción (Clima, desastres, bolsa de valores)
- Detección de fraudes
- Soporte a la toma de decisiones
- Descubrimientos científicos

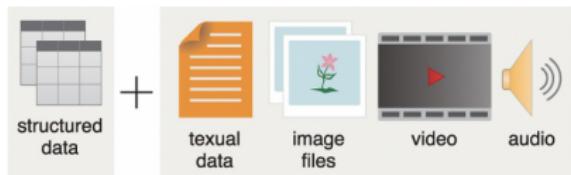
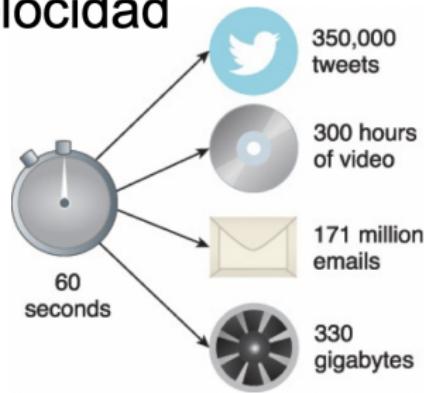
- 1 Introducción
- 2 Campo de aplicación de Big Data
- 3 Las 5 Vs de Big Data
- 4 Causas para la adopción de Big Data
- 5 Características deseadas de un sistema de Big Data
- 6 Soluciones de Big Data
- 7 El Teorema CAP
- 8 Preguntas

Las 5 Vs de Big Data

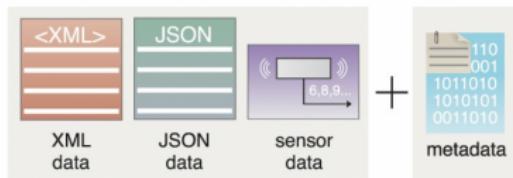


Velocidad - Variedad

Velocidad



+



Variedad

Veracidad??

misskimmy4 about an hour ago

#Latergram Last night's @wolfhomeny X @artmarkit Pop-Up Shop launch party was a success! Thanks to @hannahbronfman and @vanityprojects 🎉 PR #SocialMediaManagement #HappyClientsHappyLife

TecnoEstudios @tecnostudios

Televisión Inteligente: Interactiva y Social (TV 2.0) is.gd/iyVgRo #SocialMediaManagement

22 Aug 9:09am

BGH Business Network @BCHBusiness

A lesson in **#socialmediamanagement** from the **#NewYorkTimes**. nyti.ms/1vnVRTo #socialmedia #entrepreneur pic.twitter.com/mVtYXoOdyp

22 Aug 8:57am

Dingo Integrated Marketing

about 2 hours ago

President LGM @LGMAdAgency

Social Media services at localgeniusmarketing.com/socialmedia ... #socialmedia #SocialMediaMarketing #SocialMediaManagement

22 Aug 9:01am

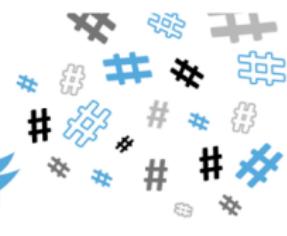
YKMD Visual Communication about 2 hours ago

#flashbackfriday #fbf

GRANT Amazingly Healthy Sour Sop, a #YKMDClient, sells excellent quality dried #Soursop Leaves.

GRANT Amazingly #Healthy has presented the Soursop leaf in organic teabags for your #health and convenience.

You may enjoy it day or ...



VALOR!



- 1 Introducción
- 2 Campo de aplicación de Big Data
- 3 Las 5 Vs de Big Data
- 4 Causas para la adopción de Big Data
- 5 Características deseadas de un sistema de Big Data
- 6 Soluciones de Big Data
- 7 El Teorema CAP
- 8 Preguntas

Causas para la adopción de Big Data

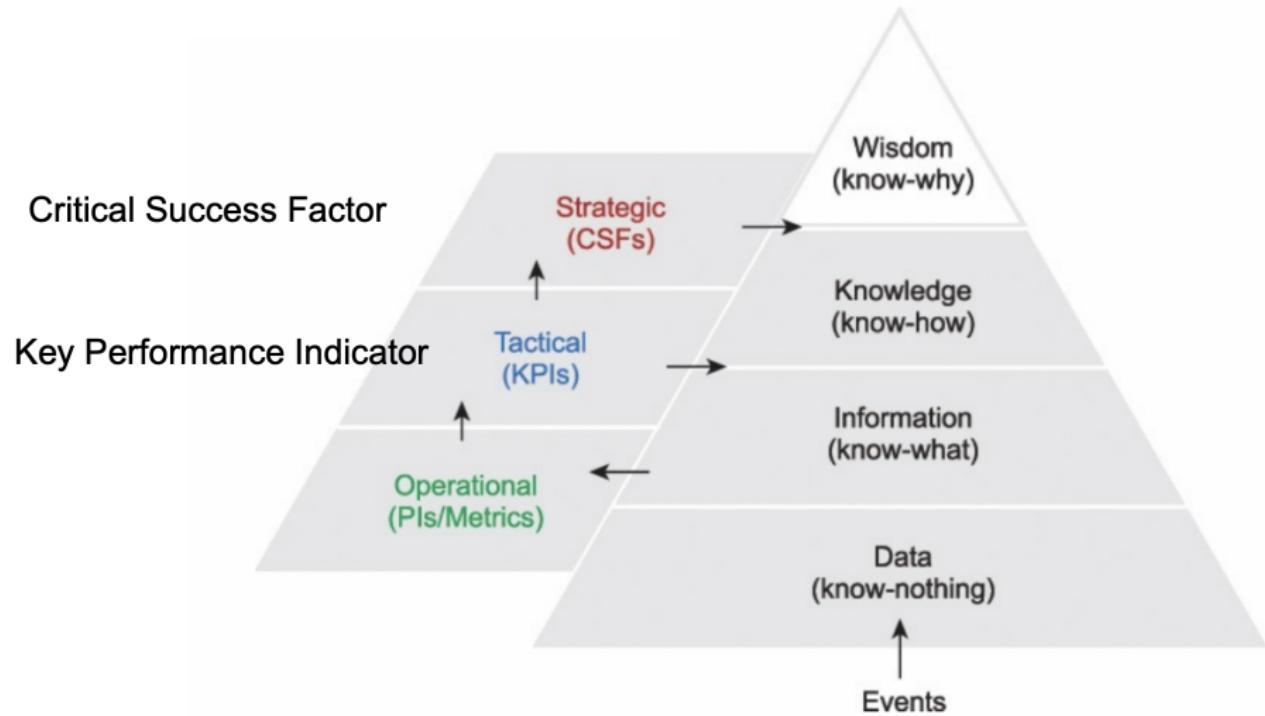
- Nueva dinámica del mercado
- Arquitectura del negocio
- Manejo de procesos de negocios
- TICs
- Internet of Everything (IoE)

Causas para la adopción de Big Data

- Nueva dinámica del mercado
 - Dot-com bubble, Recesión 2008 (Empresas en Internet)
 - Mantener la rentabilidad y reducir costos
 - Encontrar nuevos clientes y mantener los existentes
 - Ofrecer nuevos productos y servicios
 - Otorgar valor agregado al cliente

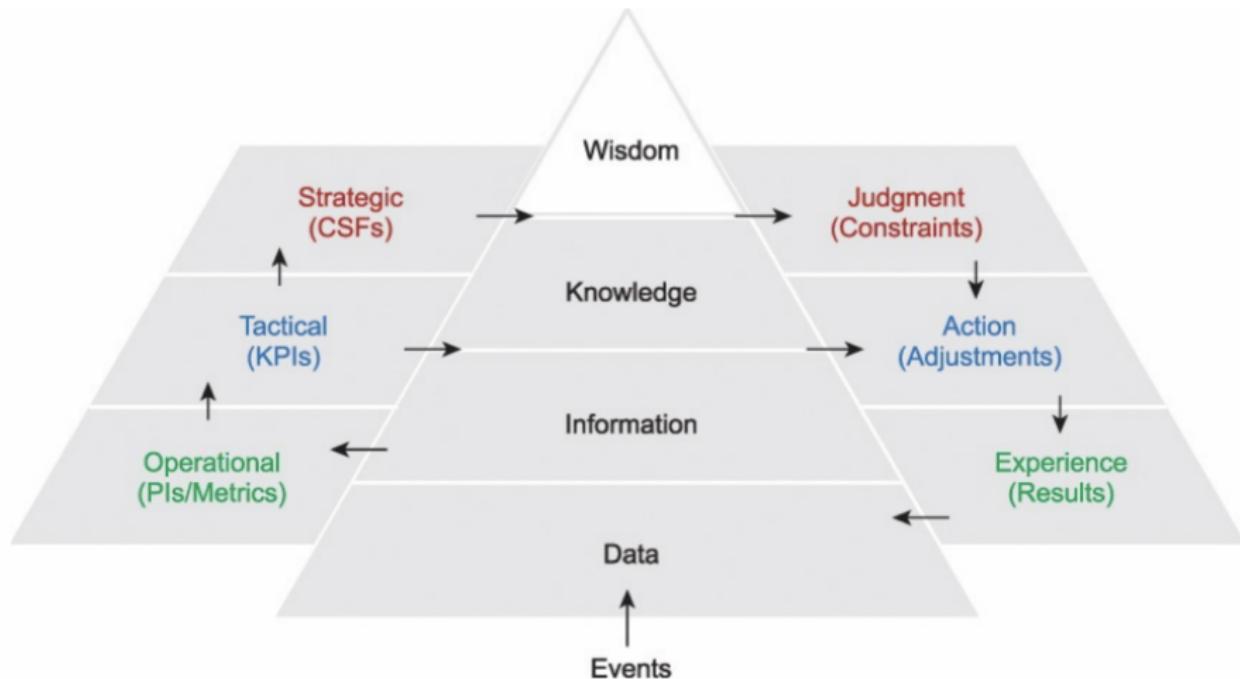
Causas para la adopción de Big Data

Arquitectura de Negocio



Causas para la adopción de Big Data

Arquitectura de Negocio



Causas para la adopción de Big Data

Procesos de Negocio

- Descripción de cómo se realiza el trabajo.
- Actividades del negocio y las relaciones con los actores responsables de ejecutarlas.
- Procesos alineados a los objetivos del negocio

Causas para la adopción de Big Data

TICs

- Data analytics and data science
- Digitization
- Affordable technology and commodity hardware
- Social media
- Hyper-connected communities and devices
- Cloud computing

Causas para la adopción de Big Data

Internet of Everything

- 14 billones
- 2020: 32 billones



- 1 Introducción
- 2 Campo de aplicación de Big Data
- 3 Las 5 Vs de Big Data
- 4 Causas para la adopción de Big Data
- 5 Características deseadas de un sistema de Big Data
- 6 Soluciones de Big Data
- 7 El Teorema CAP
- 8 Preguntas

Características deseadas de un sistema de Big Data

- Robustez y tolerancia a fallos
- Latencia baja en lecturas y escrituras
- Escalabilidad
- Generalizable
- Extendible
- Ad hoc queries
- Mantenimiento Mínimo
- Depurable

Robustez y tolerancia a fallos

- El sistema necesita comportarse correctamente así sea que algunos computadores se han caído.
- Compleja semántica y consistencia en base de datos distribuidas.
- Los sistemas deben ser "human-fault-tolerant"

Latencia Baja en lecturas y escrituras

- La mayoría de sistemas requieren leer mucha información en muy pocos segundos.
- Algunas aplicaciones requieren tiempo para propagar las actualizaciones en sus sistemas.
- Se requiere leer rápidamente información sin comprometer la robustez del sistema

Escalabilidad

- Capacidad de agregar nuevos datos o recursos sin comprometer el desempeño del sistema.
- La **arquitectura Lambda** es horizontalmente escalable a través de cada capa.
- Se logra al añadir varias computadoras.

Generalizable

- Puede soportar un número grande de aplicaciones.
- La arquitectura Lambda está basada en función de todos los datos.
- Los datos pueden ser de diferente tipo: financieros, social media, aplicaciones científicas.

- No se tiene que reinventar la rueda cada vez que se quiera agregar una característica.
- Agregar una funcionalidad requiere un costo mínimo de esfuerzo.
- A veces la inclusión de una nueva funcionalidad requiere de la migración de datos viejos en un nuevo formato.
- Capáz de migrar grandes cantidades de datos rápida y fácilmente.

Ad hoc queries

- La posibilidad de crear consultas específicas.
- La capacidad de hacer consultas puede permitir información interesante.

Hyper-connected communities and devices

- Se relaciona directo con la Internet de las cosas (IoT)
- Permitir obtener información de todos los dispositivos posibles.
- Utilizar como fuente de datos comunidades de información.

Cloud computing

- Capacidad de utilizar la nube para acceder a los datos
- Utilizar la capacidad de procesamiento existente en aplicaciones almacenadas en la nube.

Cloud computing

- Capacidad de utilizar la nube para acceder a los datos
- Utilizar la capacidad de procesamiento existente en aplicaciones almacenadas en la nube.

- 1 Introducción
- 2 Campo de aplicación de Big Data
- 3 Las 5 Vs de Big Data
- 4 Causas para la adopción de Big Data
- 5 Características deseadas de un sistema de Big Data
- 6 Soluciones de Big Data
- 7 El Teorema CAP
- 8 Preguntas

Soluciones de Big Data

- MapReduce
- NoSQL Database
- Algoritmos genéticos
- Recaptcha
- Reconocimiento de patrones
- NUI (Natural User Interface)

MapReduce

- Framework de hardware distribuido (cluster o grids) que divide los problemas en subproblemas (Map) y luego se recopila las mini-respuestas (Reduce) para generar conclusiones.
- La solución más común es Hadoop.
- Este modelo fue creado y promovido por Google.



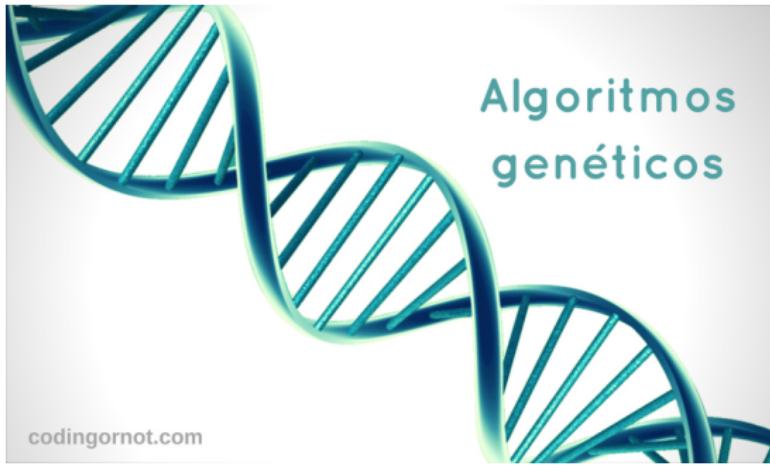
NoSQL Database

- Amplia clase de sistemas de gestión de bases de datos que difieren del modelo clásico del sistema de gestión de bases de datos relacionales (RDBMS) en múltiples aspectos:
 - No usan SQL como el principal lenguaje de consultas
 - Los datos almacenados no requieren estructuras fijas como tablas
 - No garantizan ACID (atomicidad, consistencia, aislamiento y durabilidad)
 - Escalan bien horizontalmente (ej: MongoDB, Cassandra, BigTable)



Algoritmos genéticos

- Un algoritmo es una serie de pasos organizados que describen el proceso que se debe seguir, para dar solución a un problema específico.
- Con la inteligencia artificial, surgieron los algoritmos genéticos, inspirados en la evolución biológica
- Evolucionan sometidos a mutaciones y recombinaciones genéticas.



ReCaptcha

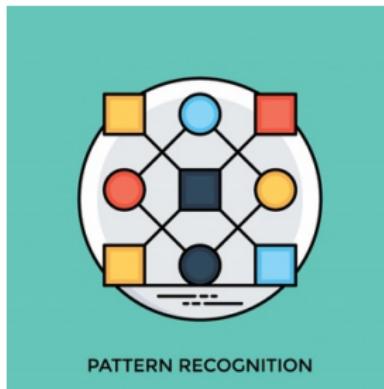
- reCAPTCHA es una extensión de la prueba CAPTCHA.
- Se utiliza para reconocer texto presente en imágenes.
- Se usa para determinar si el usuario es o no humano.
- Mejorar la digitalización de textos.
- Google compró Recaptcha.



reCAPTCHA

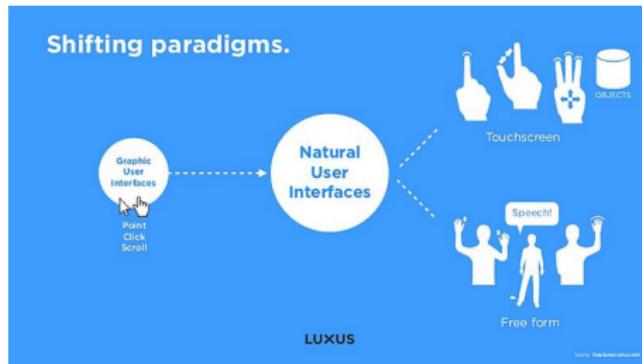
Reconocimiento de patrones

- El reconocimiento de patrones es la ciencia que se ocupa de los procesos sobre:
 - Ingeniería, computación y matemáticas
 - Relacionados con objetos físicos o abstractos
 - con el propósito de extraer información que permita establecer propiedades de entre conjuntos de dichos objetos.



NUI (Natural User Interface)

- Interfaz que permite interactuar con un sistema sin utilizar sistemas de mando o dispositivos de entrada de las GUI (ratón, teclado...)
- En su lugar, se hace uso de movimientos gestuales.
- Un ejemplo es Kinect. Reconocimiento de gestos y movimientos.



1

Introducción

2

Campo de aplicación de Big Data

3

Las 5 Vs de Big Data

4

Causas para la adopción de Big Data

5

Características deseadas de un sistema de Big Data

6

Soluciones de Big Data

7

El Teorema CAP

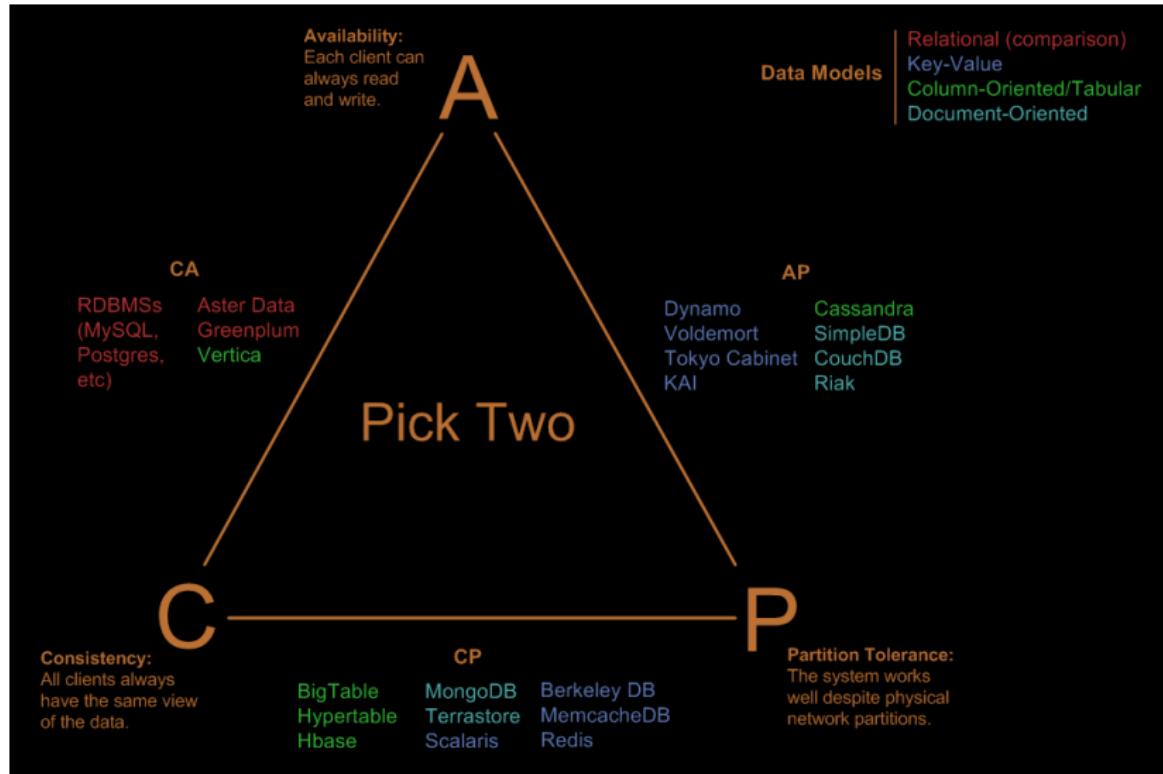
8

Preguntas

El Teorema CAP

- Teorema que nos dice que en un sistema distribuido de almacenamiento de datos no podemos garantizar consistencia y disponibilidad (para actualizaciones)
- Cuando el sistema sufre una partición (queda separado en dos o más islas).
- Este es el Teorema CAP, por Consistency (Consistencia), Availability (Disponibilidad) y Partition Tolerance (Tolerancia al Particionamiento).
- Se debe tener en cuenta las exigencias del proyecto para saber qué atributos de calidad necesitamos
- Y así elegir el tipo de base de datos que necesitaremos.

El Teorema CAP



El teorema es que solo puedes garantizar dos de estos tres atributos:

- CP (Consistencia y Tolerancia al particionamiento):
 - No se garantiza la disponibilidad
 - Hay clientes que por ejemplo requieren que el sistema esté disponible 100% del tiempo o muy cerca
 - Con bases de datos que cumplan con CP no es posible garantizar esto
 - Se puede lograr en cierto nivel, pero el sistema está enfocado en aplicar los cambios de forma consistente aunque se pierda comunicación con algunos nodos.

El Teorema CAP

- AP (Disponibilidad y Tolerancia al particionamiento):
 - En este caso no se garantiza que los datos sean iguales en todos los nodos todo el tiempo
 - En este caso el sistema siempre estará disponible para las peticiones aunque se pierda la comunicación entre los nodos.

- CA (Consistencia y disponibilidad):

- En este caso no se puede permitir el particionado de los datos, porque se garantiza que los datos siempre son iguales y el sistema estará disponible respondiendo todas las peticiones.
- Por ejemplo, los sistemas de bases de datos relacionales (SQL de toda la vida) son CA porque todas las escrituras y lecturas se hacen sobre la misma copia de los datos.

- 1 Introducción
- 2 Campo de aplicación de Big Data
- 3 Las 5 Vs de Big Data
- 4 Causas para la adopción de Big Data
- 5 Características deseadas de un sistema de Big Data
- 6 Soluciones de Big Data
- 7 El Teorema CAP
- 8 Preguntas

Preguntas

- ¿Alguna Pregunta?.

