

Pasado, Presente y Futuro de los Datos

Un breve panorama



ELIZABETH LEÓN GUZMÁN

Profesora Titular

Departamento de Ingeniería de Sistemas e Industrial

Facultad de Ingeniería

Universidad Nacional de Colombia

- Ingeniera de Sistemas de la Universidad Nacional de Colombia
- Maestría en Ingeniería de Sistemas de la Universidad de Colombia
- Maestría en *Electrical Engineering with emphasis in Computation at The University of Memphis*
- Doctorado en *Computer Engineering and Computer Science at University of Louisville*
- Directora Grupo de Investigación MIDAS – Minería de Datos

¿Qué son los datos?

¿Información?



¿cantidades?

¿Unidad mínima
de información?

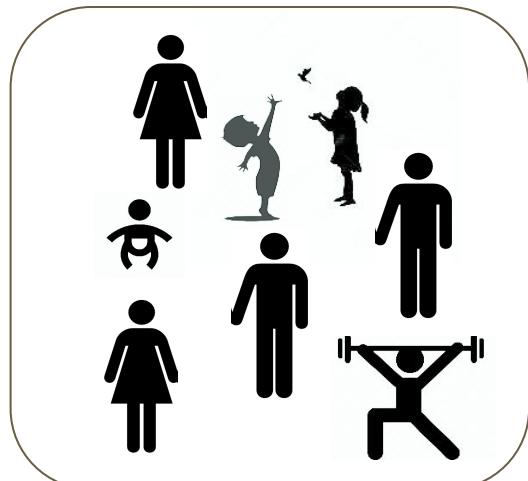
¿valores?

¿medidas?

¿Qué son los datos?

Interés de medición

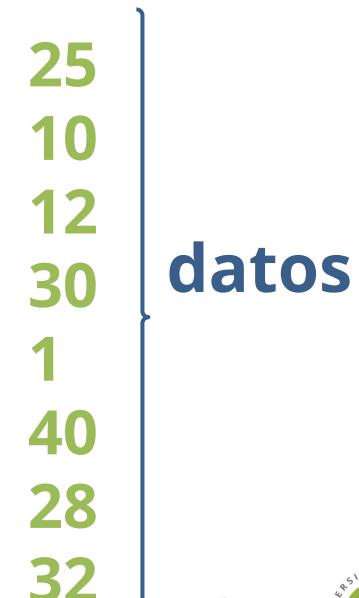
Población: Personas que viven en mi barrio



MIDAS

¿Cuáles características tiene la población?

- Edad
- Género
- Nombre
- Dirección
- Teléfono



¿Qué son los datos?

Un dato
es una representación simbólica de un
atributo o **variable**

Variable es
característica de la población que se desea medir



Ejemplo:

Población: Estudiantes de MINTIC2020

grupo(pepito) = 7

↑
Variable

↑
Dato

La variable es cuantitativa

Ejemplos:

Población: Ciudades de Colombia

Variables: Temperatura

Población

Datos



¿Qué es Información?

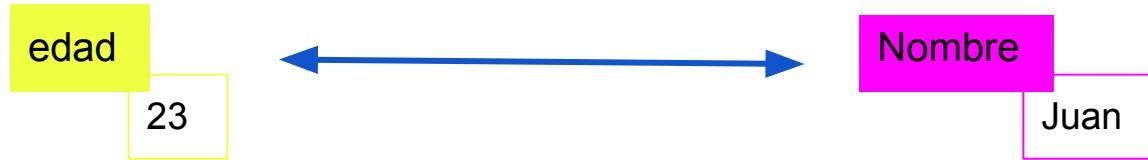
¿Conocimiento?



¿datos?

Información

- Un dato no constituye información.
- Procesamiento de los datos nos proporciona información (relación entre los datos)



Información son el conjunto de datos procesados y relacionados sobre un determinado hecho o evento

Información

Ciudad	Temperatura	Fecha
Bogotá	15	Julio 20 2020
Cartagena	28	Julio 20 2020
Bogotá	17	Julio 20 2021



Datos estructurados y
relacionados
“Información”



Datos almacenados
generalmente en
“bases de datos”

Hechos
QUE CONECTAN

UNIVERSIDAD NACIONAL
DE COLOMBIA

MIDAS

Misión
TIC 2022



Datos almacenados

- Bases de datos
- Archivos (excel, pdf, txt, csv, etc)

Una colección o conjunto de datos relacionados, y una descripción de estos datos, diseñados para cumplir con las necesidades de información de una organización

(Connolly & Begg)



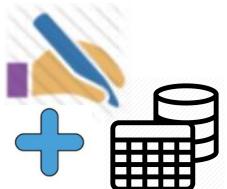
Bases de Datos

Las bases de datos deben permitir realizar operaciones sobre los datos:

- Almacenar
- Procesar
- Recuperar o Consultar
- Actualizar
- Eliminar



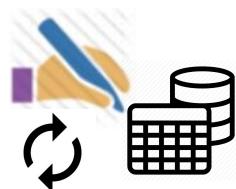
Operaciones CRUD de BD



Creation - Crear- Inserción: Insertar datos en la bd



Read - Lectura: Leer datos (consultas) para obtener información

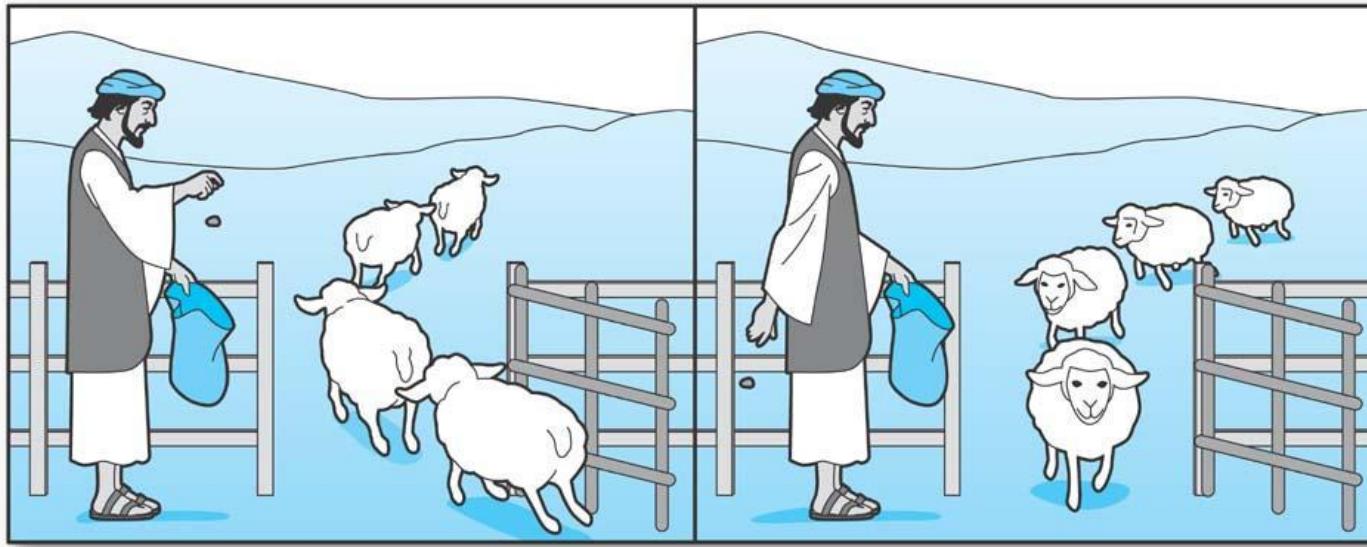


Update - Actualizaciones: Cambiar valor del dato



Delete - Borrado: Eliminar datos

Métodos primitivos para almacenamiento de datos



Tomado de Libro: Guillenson. *Administración de Bases de Datos*. LIMUSA WILEY

Interés en datos por lo menos en los últimos 12.000 años

Historia de los Datos



Fichas de arcilla para llevar registro del contenido de los cargamentos

Cortes y muescas en palos de madera o nudos en cuerdas

Surgimiento de ciudades:
Trueque, uso de moneda para comercio. Necesidad de llevar **registro**: datos para saber la producción. Después registro de calendarios, censos, matrimonios, contribuciones a la iglesia, pago de salarios, etc.



Máquina tabuladora para procesamiento de datos (futuro IBM)



Cinta magnética para guardado de datos

3000 AC

1000 DC

1640

1805

1890

1950

1953

1956

Tablillas de arcilla:
pictogramas, símbolos para describir venta de tierras y transacciones (pan, cerveza, ovejas, ganado y prendas de vestir)

Interés comercial: Grabado de datos en papel (contabilidad por partida doble)



Máquina de Pascal:
sumadora (preursora del odómetro automotriz)



Tarjetas perforadoras. Telar programable



Primeros computadores electrónicos

Dispositivo de almacenamiento en disco

Imágenes tomadas de [1]. Guillenson.
Administración de Bases de Datos.

Inicio de los DBMS

1960s:

- Centrado en la programación de los computadores.
- Necesidad de guardar y controlar los datos (usando cintas y discos). Se inicia la creación de programas que ayuden a esas tareas, lo que da inicio a los Sistemas Gestores o Administradores de Bases de Datos o DBMS (DataBase Management Systems)

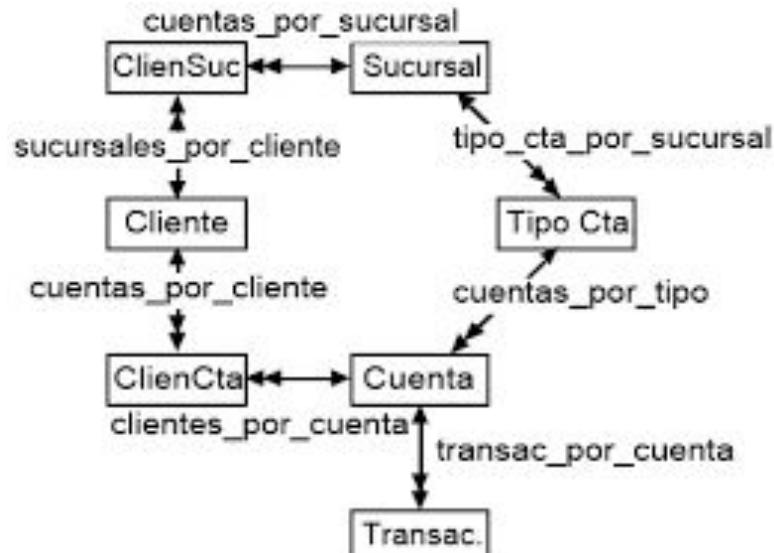
“Conjunto de programas que maneja la estructura de la BD y controla el acceso a los datos guardados en ella”



Inicios de las bases de datos

Charles Bachmann desarrollo el primer DBMS (IDS) con un modelo de red

1960



Tomado de
<https://cursos.aiu.edu/base%20de%20datos%20SOG/Sesi%C3%B3n%204.pdf>

Red: Modelo de datos basado en red

Hechos
QUE CONECTAN

UNIVERSIDAD NACIONAL
DE COLOMBIA

Inicios de las bases de datos

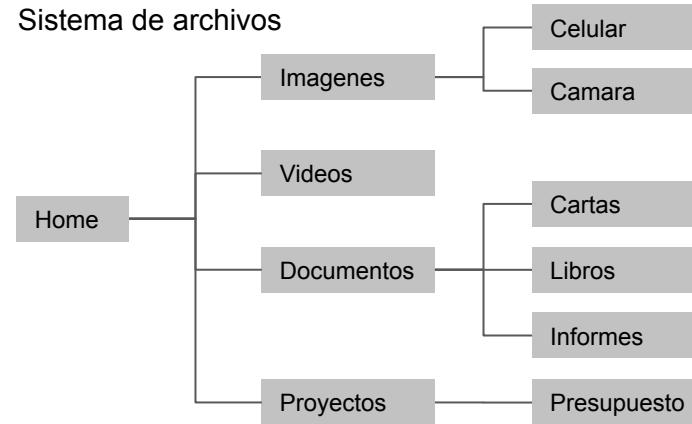
Charles Bachmann desarrollo el primer DBMS (IDS) con un [modelo de red](#)

1960

IBM desarrolló el primer DBMS exitosamente comercial (IMS) [Modelo Jerárquico](#) donde la relación de los datos era en forma de árbol

[Conference On DATA Systems Languages](#) (CODASYL) modelo definido. Modelo de red más estandarizado.

Sistema de archivos



CURSO

código nombre descripción

OFERTA

grupo horario salón

PROFESOR

ID nombre

ESTUDIANTE

ID nombre nota

Tomado de
<http://alucard-base-de-datos.blogspot.com/2012/01/el-modelo-de-base-de-datos-de-red.html>

Jerárquico: Modelo de datos basado en árboles



Inicios de las bases de datos

Charles Bachmann desarrollo el primer DBMS (IDS) con un [modelo de red](#)

1960

IBM desarrolló el primer DBMS exitosamente comercial (IMS) Modelo [Jerárquico](#) donde la relación de los datos era en forma de árbol

Conference On DAta Systems Languages (CODASYL) modelo definido. Modelo de red más estandarizado.

IBM

UNIVAC

HoneyWell

Computadores mainframe
Legacy systems (datos históricos)



Tomada de
<https://sites.google.com/site/miguessite/que-es-la-computadora-mainframe>



Inicios de las bases de datos

Charles Bachmann desarrollo el primer DBMS (IDS) con un [modelo de red](#)

IBM desarrolló el primer DBMS exitosamente comercial (IMS) Modelo [Jerárquico](#) donde la relación de los datos era en forma de árbol

Conference On DATA Systems Languages (CODASYL) modelo definido. Modelo de red más estandarizado.

IBM UNIVAC HoneyWell

Computadores mainframe
Legacy systems (datos históricos)

1960

1970

1980

Ted Codd definió el modelo relacional en el Laboratorio San Jose de IBM

- INGRES en la University of California, Berkeley. Llega a ser comercial. Seguido por POSTGRES que fue incorporado en Informix
- System R en IBM San José Laboratory Llegó a ser DB2

1976: Peter Chen definió el diagrama entidad relación (ER)

Tecnología de bases de datos relacionales es madura

El SQL es estandarizado por la ISO (finales de 1980s)

DBMS relacionales

ACID

Propiedades que garantizan que las transacciones de la base de datos se procesen de manera confiable. ACID garantiza la recuperación de la base de datos de cualquier falla que ocurra durante el procesamiento de una transacción.

Una transacción se puede componer de varias acciones que deben ser ejecutadas todas o ninguna.

Ejemplo:
Transferencia bancaria



MIDAS

Misión
TIC 2022

DBMS relacionales

ACID

Atomicity (Atomicidad): los pasos de una operación se ejecutan todas o ninguna

Consistency (Integridad): garantiza que los datos son correctos e íntegros.

Isolation (Aislamiento): los datos se protegen de una operación mientras se realiza otra. **Ej:** Los datos se aíslan de Lectura mientras se actualización los datos.

Durability (Persistencia): Una vez se realice una acción en la BD esta persiste y no se podrá deshacer.



Relación →
Tabla

Estudiante

Nombre	Apellido	Edad	Género	teléfono
Jorge	Díaz	20	m	3562819
Maria	Martínez	23	f	9873209
Rosa	Gómez	19	f	1743829
Pedro	Suarez	21	m	6386472

Modelo Relacional

Estudiante

Código	Nombre	Apellido	Edad	Género	Teléfono
100	Jorge	Díaz	20	m	3562819
101	María	Martínez	23	f	9873209
102	Rosa	Gómez	19	f	1743829
103	Pedro	Suarez	21	m	6386472

Asignatura

Código	Nombre	Créditos
10	Programación	4
20	Bases de Datos	3
30	Matemáticas	4
40	Software	3

Inscripción

Código_estudiante	Cod_asignatura	Semestre
100	10	2020-I
100	20	2020-I
100	10	2019-II
102	10	2020-I
102	20	2020-I
102	30	2020-I
103	30	2019-I
103	40	2019-I

Datos como recurso Corporativo

- Los datos son indispensables en cada uno de los negocios y organizaciones.
- Los datos provee ventaja competitiva a una compañía.
- Uso de los datos a un nivel más alto haciendo que cada vez sean más importantes como recurso corporativo.

Ejemplos:

- FedEx contrao ventaja competitiva al ser la primera empresa de mensajería en dar acceso a los datos del estado de la entrega de seguimiento en la web
- Bancos permitieron el acceso a las cuentas por la web

Datos como recurso Corporativo

- Los datos son un recurso difíciles de manejar
- Cantidades inmensas de datos
- Es necesario utilizar software que permita ayudar a manejar los datos
- Es necesario hardware cada vez más rápido a medida que el volumen de datos crece.
- Especialistas en administración de datos son necesarios

Inicios de las bases de datos

Expansión de los DBMS relacionales y mejora de su rendimiento

1990

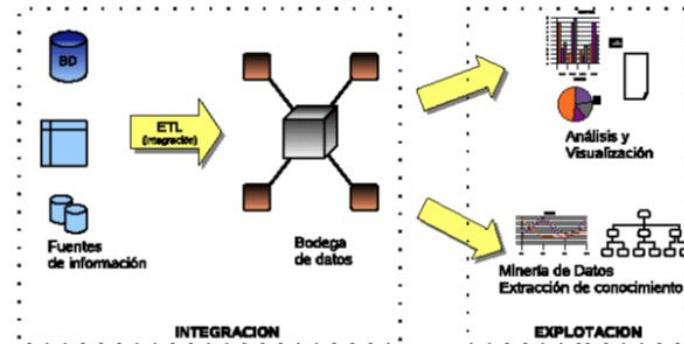
Bases de Datos distribuidas es realidad

Modelo Orientado a Objetos

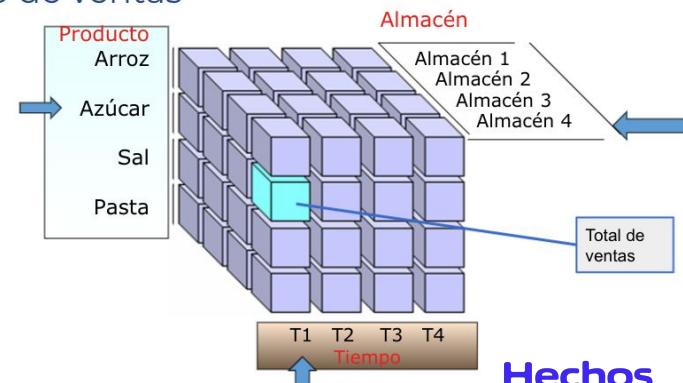
Modelo OO incorporado al modelo relacional

Bodegas de Datos (Data warehouse)
OLAP

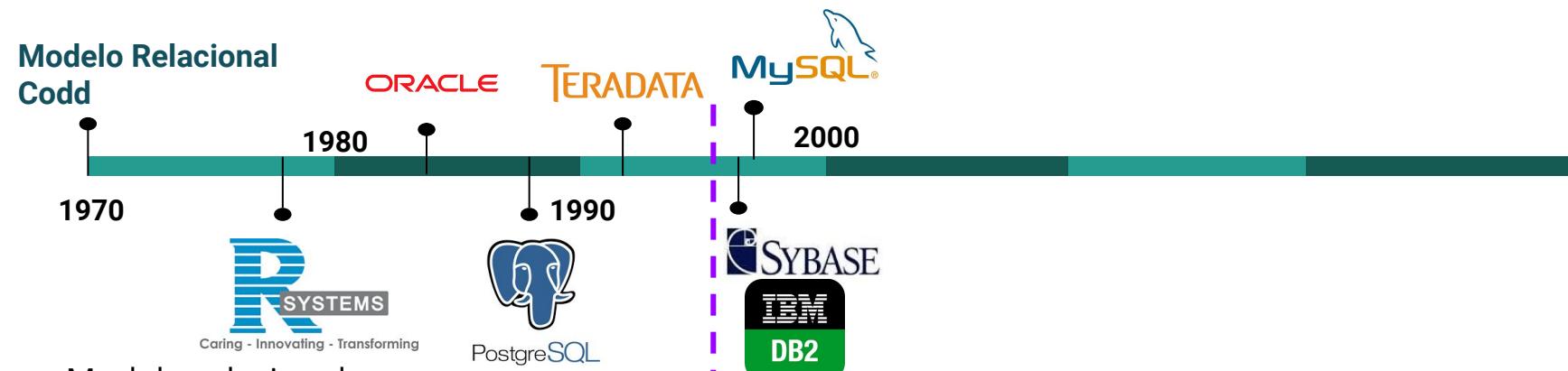
Web, Internet, interés en texto y multimedia.



Cubo de ventas



Historia de las Bases de Datos



Modelo relacional
Modelos robustos,
manejan ACID

La World Wide Web surge
y comienza a popularizarse y a
democratizarse, por lo que el volumen
de datos aumenta de manera
exponencial a medida que pasa el
tiempo.



Not SQL

Only

2009

Dynamo de Amazon
Bigtable de Google



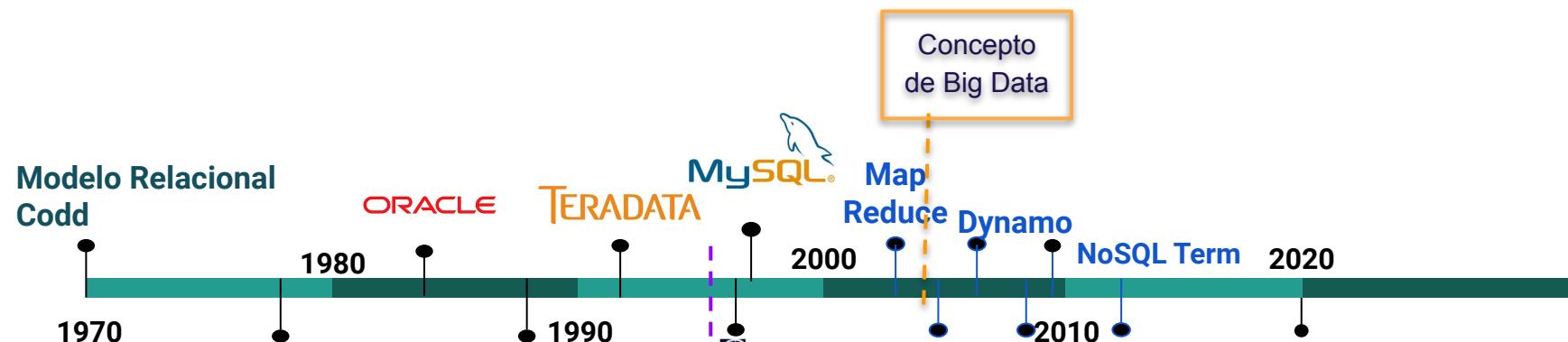
Hechos
QUE CONECTAN

UNIVERSIDAD NACIONAL
DE COLOMBIA

MIDAS

Misión
TIC 2022

Historia de las Bases de Datos



Modelo relacional
Modelos robustos,
manejan ACID



Caring - Innovating - Transforming

La World Wide Web surge
y comienza a popularizarse y a
democratizarse, por lo que el volumen de
datos aumenta de manera exponencial a
medida que pasa el tiempo.

NoSQL Código abierto
Modelos más flexibles para
almacenar, organizar y capturar
mayor cantidad de datos.



Información - Almacenamiento

Bases de Datos Relacionales

Tradicionales,
orientadas a lo
transaccional



Bodegas de Datos

Almacenar datos
históricos,
agregados

Datos
estructurados

Bases de Datos NoSQL

Orientadas a
consulta
Formatos JSON, XML



Sistema de Archivos

HDFS

Datos semi o NO
estructurados

Características de Big Data

Generación de datos

Redes sociales
Correos
Comercio electrónico
Gobierno electrónico



Tomado de documental “¿El fin de la memoria?” de Vincent



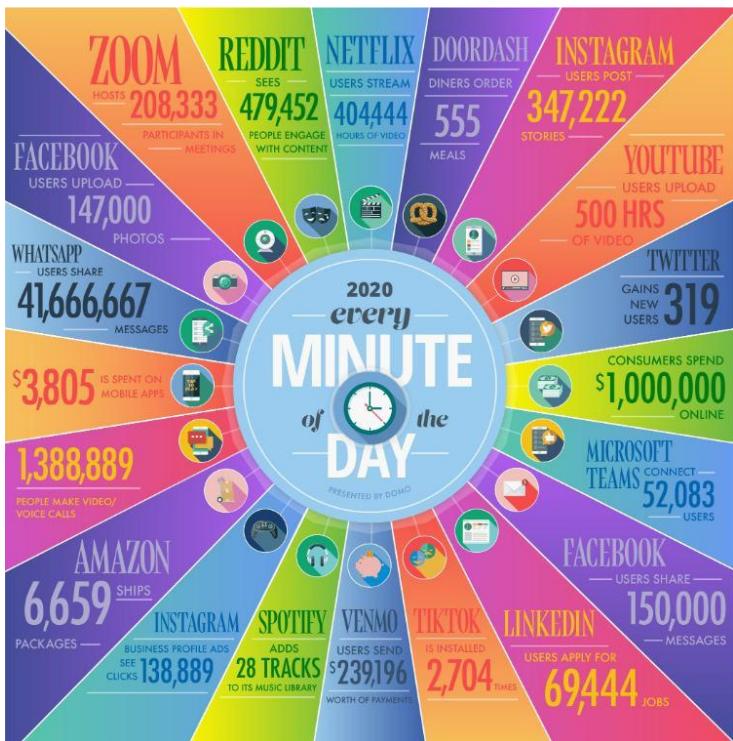
Tomado de web 123rf.com icono de comercio electrónico

Computación en la nube
Consolas de Juego
Teléfonos Inteligentes
Tabletas
Plataformas de Streaming



Características de Big Data

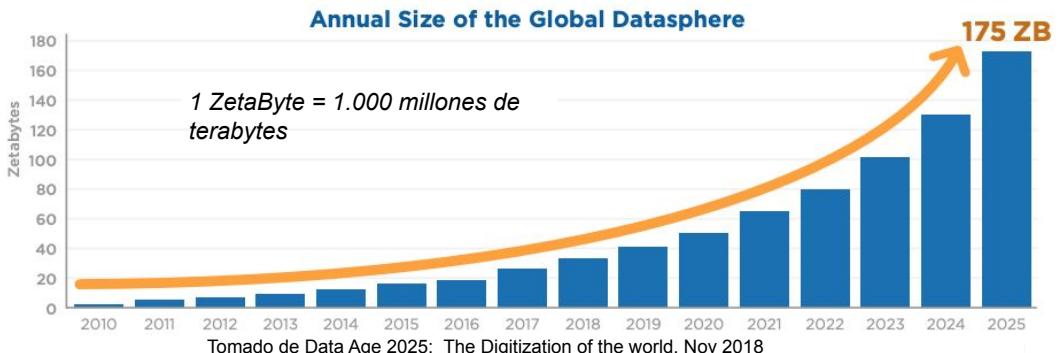
Grandes volúmenes de información en cortos períodos de tiempo



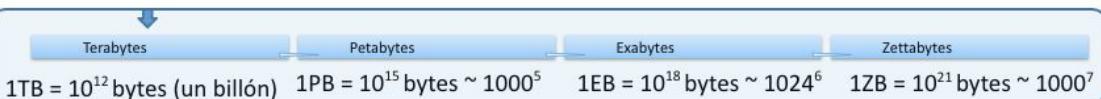
Tomado de
[https://www.trecebits.com/2020/08/16/que-ocurre-en-u-n-minuto-en-internet-en-2020-infografia/](https://www.trecebits.com/2020/08/16/que-ocurre-en-un-minuto-en-internet-en-2020-infografia/)



En los últimos dos años se ha generado el 90% de los datos



Almacenamiento promedio de las compañías (ahora)



Características de Big Data

Generación de datos

Internet de las Cosas (IoT)

Dispositivos pequeños y baratos, GPS, todos los aparatos podrán tener su propia IP

- Redes de sensores
- Redes de dispositivos



Tomado de
<https://www.muycomputerpro.com/zona-transformacion-digital/iot-industrial-optimizacion-transformacion/>

Características de Big Data

Análisis - Extracción de Conocimiento



Características de Big Data

Veracidad y Valor de los Datos



Tomado de
<https://datos.gob.es/es/noticia/transformando-los-datos-abiertos-en-valor-socioeconomico>

- Datos “Correctos”
- Calidad de los datos
- Ruido, inconsistencias, vagos, errores, etc.
- Noticias falsas
- Spam



Apoyo a toma de
decisiones

Video CERN



["¿El fin de la memoria?"](#) de Vincent Amouroux
Minutos del 18:10 al 21:20

Hechos
QUE CONECTAN

UNIVERSIDAD NACIONAL
DE COLOMBIA

MIDAS

Misión
TIC 2022

Características de Big Data

5 V's

Big Data



Volumen

Variedad

Velocidad

Veracidad

Valor

Big Data y sus retos

“Big Data” son datos cuyo volumen, diversidad y complejidad requieren nueva arquitectura, técnicas, algoritmos y análisis para gestionar y extraer valor y conocimiento oculto en ellos

El progreso y la innovación no se ven obstaculizados solo por la capacidad de recopilar datos, sino por la capacidad de **gestionar**, analizar, sintetizar, visualizar, y descubrir el conocimiento de los datos recopilados de manera oportuna y en una forma escalable



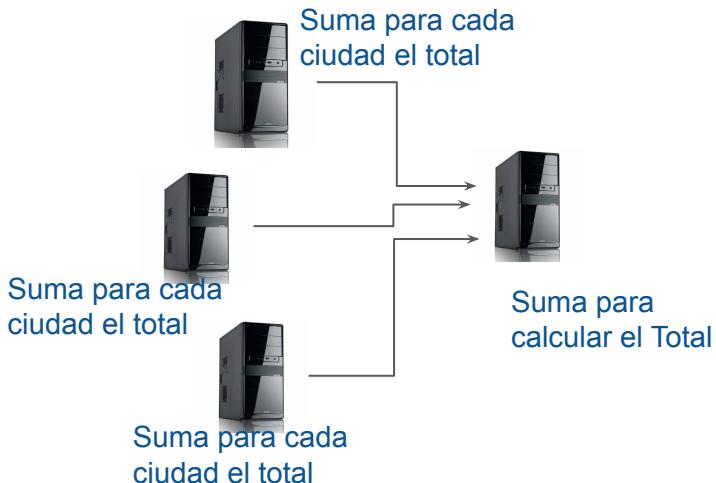
Arquitecturas Distribuidas

Computación en la Nube



Ejemplo: Map Reduce

Contar las personas que viven en Colombia de los datos de un Censo por ciudades. Los datos están distribuidos en varias máquinas



Google

Hechos
QUE CONECTAN

UNIVERSIDAD NACIONAL
DE COLOMBIA

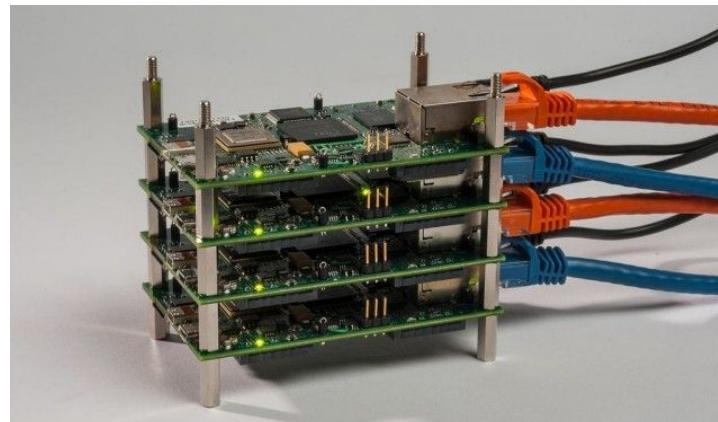
Arquitecturas Distribuidas

clusters en la nube



Tomada de
https://es.wikipedia.org/wiki/Computaci%C3%B3n_en_la_nube

Propio Cluster con Raspberries PI



Tomada de
<https://arstechnica.com/information-technology/2013/07/creating-a-99-parallel-computing-machine-is-just-as-hard-as-it-sounds/>

Video NUBE



[“¿El fin de la memoria?” de Vincent Amouroux](#)
Minutos del 34:15 al 36:52

Hechos
QUE CONECTAN ✓

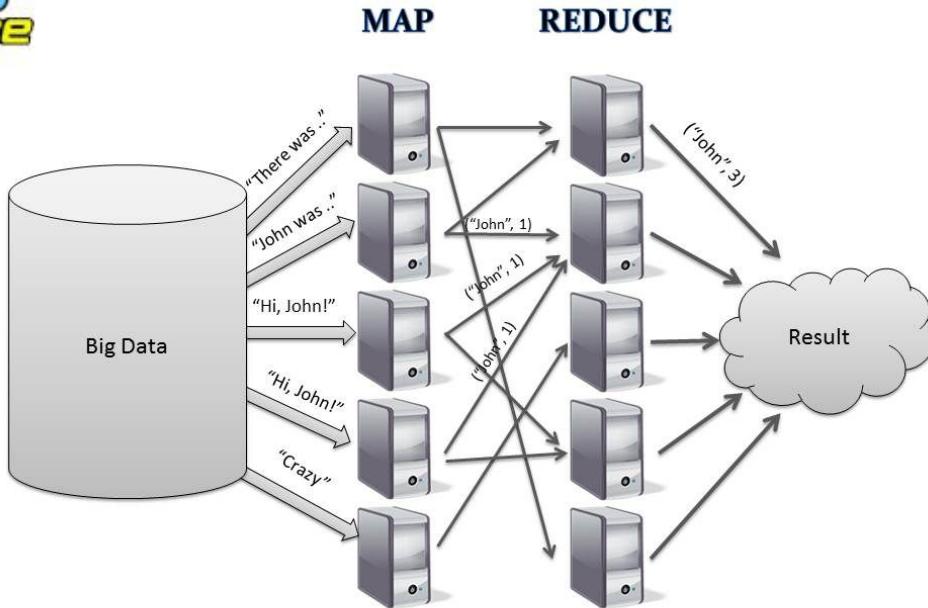
UNIVERSIDAD NACIONAL
DE COLOMBIA ✓



- Framework de código abierto diseñado para almacenar y procesar grandes conjuntos de datos en clúster de máquinas.
- Creado por Doug Cutting y Mike Cafarella en 2005. El nombre proviene del nombre del elefante (juguete) de un hijo de Cutting.
- Permite a los desarrolladores utilizar múltiples máquinas para una única tarea
- 2009: Yahoo usó Hadoop para ordenar 1TB de datos en 62 segundos
- 2013: Hadoop es utilizado por cientos de compañías

Tecnología para la Gestión de “Big Data”

Gestión de datos
e información



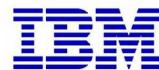
Hechos
QUE CONECTAN ✓

UNIVERSIDAD NACIONAL
DE COLOMBIA

MIDAS

Tecnología para la Gestión de “Big Data”

Hadoop - ¿Quienes lo utilizan?



The New York Times



YAHOO!



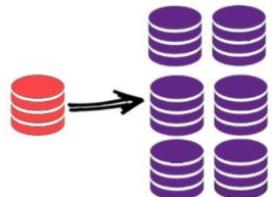
Tecnología para la Gestión de “Big Data”

DBMS - ¿Por qué NoSQL?

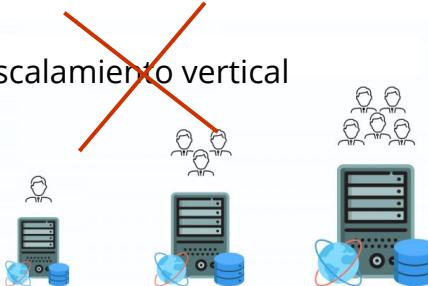
Gestión de datos
e información

Arquitecturas Distribuidas

Bases de Datos NoSQL



Escalamiento vertical



Escalamiento horizontal



- Evitar la complejidad innecesaria.
- Alto rendimiento. Orientadas a consulta
- Escalabilidad Horizontal a bajo costo

Tecnología para la Gestión de “Big Data”

DBMS - ¿Por qué NoSQL?

Gestión de datos
e información

- Aplicaciones web ejecutan miles de transacciones por segundo:
 - Publicidad (Ad Serving)
 - Servicio de email (Email Services)
 - Carritos de compra (Shopping Carts)
 - Operaciones financieras (Financial Trades)
- Bases de datos tradicionales no capacitadas para el manejo de esas “tasas” de generación de datos.



- Modelo flexible
- Evita gastos innecesarios en transacciones
- Evita la complejidad de consultas SQL

Permite:

- Cambios en la BD de manera fácil y frecuente.
- Desarrollo rápido
- Manejar grandes volúmenes de datos (ejemplo: Google)
- No definir esquema



NoSQL



Amazon DynamoDB



48

Cuatro grandes modelos

- Clave-Valor (Key-Value)
- Documento
- Familias de Columnas
- Grafos

NoSQL - Modelo Clave Valor



- Más popular y sencilla
- Datos son asociados por claves a valores. Los valores no conforman estructuras particulares.
- Estructura de “diccionario”
- Datos no requieren formato.

árbol :

Un árbol es una planta, de tallo leñoso, que se ramifica a cierta altura del suelo.

Clave

Valor

La clave debe ser única

NoSQL - Modelo Clave Valor



Clave (key)	Valor (Value)
101	La Ciudad y los Perros de Mario Vargas Llosa
102	Cien Años de Soledad de Gabriel García Márquez
201	La Divina Comedia de Dante Alighieri
202	Don Quijote de La Mancha de Miguel de Cervantes

No estructura

```
get ("102") -> "Cien Años de Soledad de Gabriel García Márquez"  
put ("902", "La Iliada de Homero")
```

NoSQL - Modelo Clave Valor



<https://docs.riak.com/>



amazon
DynamoDB



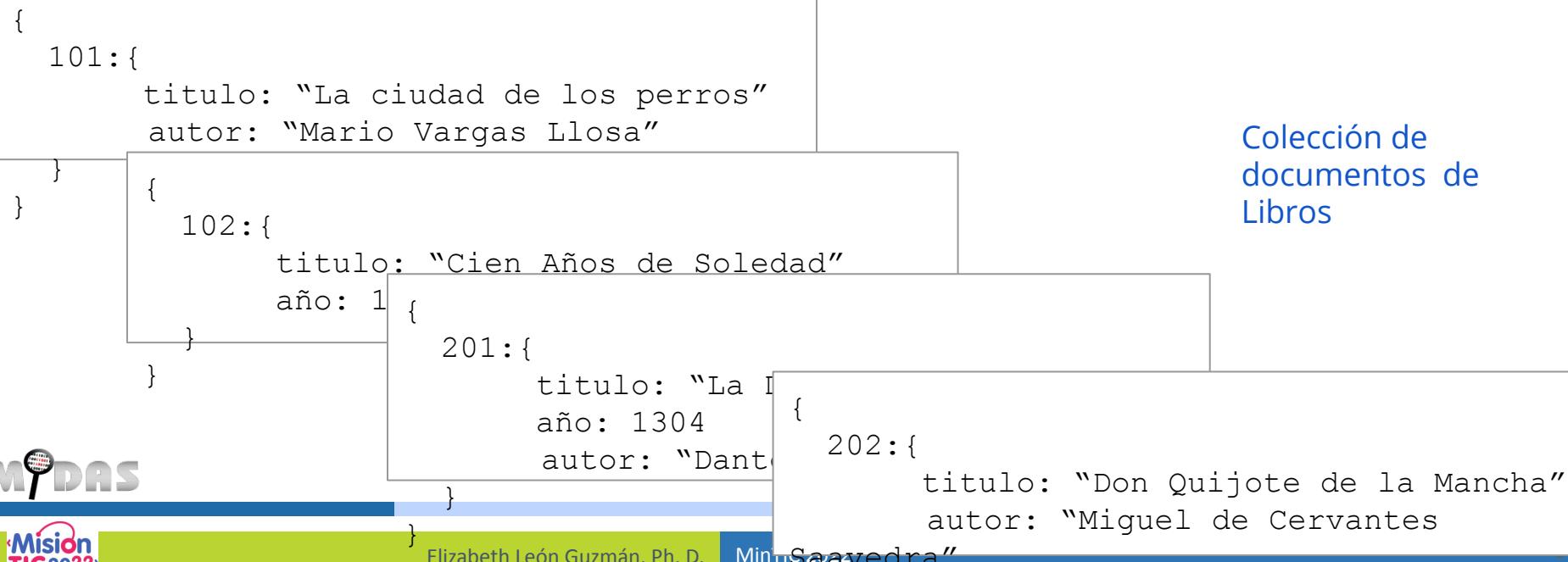
Compañías que usan bd clave-valor:

- GitHub
- At&T
- BestBuy



NoSQL - Clave Documento

clave	titulo	año	autor
101	La Ciudad y los Perros		Mario Vargas Llosa
102	Cien Años de Soledad	1967	
201	La Divina Comedia	1304	Dante Alighieri
202	Don Quijote de La Mancha		Miguel de Cervantes Saavedra



Colección de libros en JSON

```
{ "Libros" : [ { _id: 101, titulo: "La ciudad de los perros", autor: "Mario Vargas Llosa"}, { _id: 102, titulo: "Cien años de Soledad", año: 1967}, { _id: 201, titulo: "La Divina Comedia", año: 1304 autor: "Dante Alighieri"}, { _id: 202, titulo: "Don Quijote de la Mancha", autor: "Miguel de Cervantes Saavedra"} ] }
```



NoSQL - Clave Documento



- Internet de las Cosas.
- Comercio electrónico.
- Procesamiento de información y análisis en tiempo real.
- Administración de contenido.

Compañías que usan bd clave-documentos:

- Ubuntu
- BBC
- Aplicaciones para android:
SpreadLyrics
- Uber
- Ebay
- Facebook
- Expedia
- McAfee

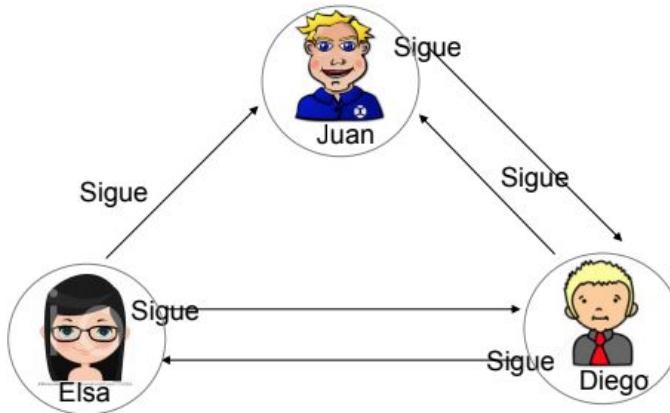


Bases de Datos con modelo en Grafos

Bases de datos que usan como estructura un grafo para guardar los datos

Estructura de datos, que se compone de:

- Nodos → Entes (personas, lugares, cosas, etc)
- Propiedades → Información de los nodos y enlaces
- Enlace → Relaciones



Bases de Datos con modelo en Grafos



AllegroGraph



Cytoscape



neo4j

- Armada de USA
- Caterpillar
- Volvo
- NASA

Google Knowledge Graph



<https://youtu.be/mmQI6VGvX-c>



Hechos
QUE CONECTAN ✓



MIDAS

Misión
TIC 2022

Tecnología para la Gestión de “Big Data”

Blockchain

Gestión de datos
e información

Arquitecturas Distribuidas

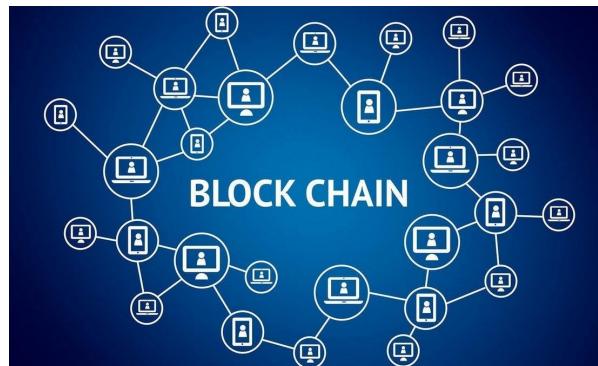
*Blockchain
(cadena de bloques)*



https://www.youtube.com/watch?v=Yn8WGao_aka

Sistema para:
Producir
Almacenar
Gestionar
datos

Incorruptible



<https://www.xataka.com/especiales/que-es-blockchain-la-explicacion-definitiva-para-la-tecnologia-mas-de-moda>

Base de datos distribuida y segura (por el cifrado) que se puede aplicar a cualquier transacción (no solo económica), que elimina a los intermediarios para el control de los datos (el control es global)

Tecnología para Análisis de “Big Data”

Analítica

“Big Data Analytics”: Identificar patrones “Conocimiento” en Big Data

Inteligencia de Negocios (BI)



Visualizar estadísticas de los datos. Los usuarios infieren el conocimiento a partir de las gráficas.

Dashboard
OLAP
Datalake



PYTHON



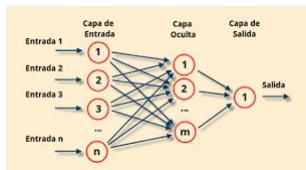
Visualización



Tomado de
https://www.researchgate.net/figure/Analytics-and-big-data-related-word-cloud_fig1_327202967

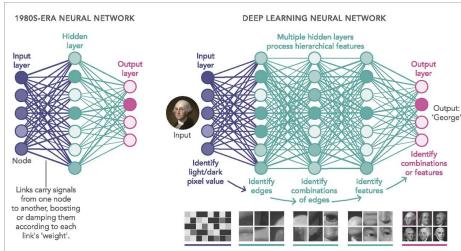
Inteligencia Artificial Aprendizaje Maquinal

Automatiza y optimiza la tarea de encontrar “patrones” (valor) en datos, por medio de algoritmos de computación avanzados



Watson Machine Learning

Aprendizaje Profundo



<https://www.pnas.org/content/116/4/1074>

Aprendizaje maquinal usando una red neuronal que se compone de varios niveles jerárquicos

Videos Venecia



“El fin de la memoria?” de Vincent Amouroux
Minutos del 45:45 al 50:26

Hechos
QUE CONECTAN

UNIVERSIDAD NACIONAL
DE COLOMBIA

MIDAS

Misión
TIC 2022

Aplicaciones de “Big Data Analytics”

- Compañías de servicios/ventas
 - Conocer los usuarios/clientes
 - Mejorar Servicios
 - Identificar oportunidades
 - Definir estrategias de marketing
 - Establecer nuevos modelos de recomendación

Aplicaciones de “Big Data Analytics”

- Tráfico
 - Visualización del tráfico
 - Análisis de los patrones de desplazamiento
 - Rutas de congestión
 - Planificación urbana
 - Uso de las carreteras
- Clima

Aplicaciones de “Big Data Analytics”

Deportes

- seguimientos al rendimiento de cada jugador (análisis de videos), tecnología de sensores (cestas, mallas, etc.) usando teléfonos inteligentes y servicios de la nube.
- NBA -> preparación de partidos
- NFL (National Football League)
<https://youtu.be/aztUUcZfXb8> (optimización de las agendas de los partidos usando IBM tool)

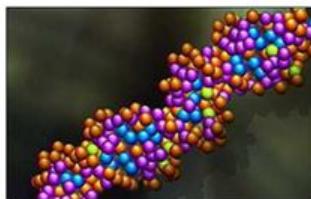
Aplicaciones de “Big Data Analytics”

Astronomía



- Astronomical sky surveys
- 120 Gigabytes/week
- 6.5 Terabytes/year

Genómica



- 25,000 genes in human genome
- 3 billion bases
- 3 Gigabytes of genetic data

Telefonía



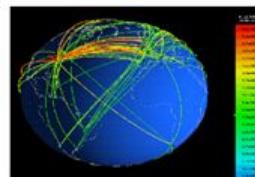
- 250M calls/day
- 60G calls/year
- 40 bytes/call
- 2.5 Terabytes/year

Transacciones de tarjetas de crédito



- 47.5 billion transactions in 2005 worldwide
- 115 Terabytes of data transmitted to VisaNet data processing center in 2004

Tráfico en Internet



- Traffic in a typical router:
- 42 kB/second
 - 3.5 Gigabytes/day
 - 1.3 Terabytes/year

Procesamiento de información WEB



- 25 billion pages indexed
- 10kB/Page
- 250 Terabytes of indexed text data
- "Deep web" is supposedly 100 times as large



Problemas de Almacenamiento

Almacenamiento (dispositivos)

- Cuarzo
- ADN
- Programación Cuántica

Videos



["¿El fin de la memoria?"](#) de Vincent Amouroux
Minutos del 14:18 al 16:41 y del 22:40 al 26:42

Hechos
QUE CONECTAN ✓



MIDAS

Misión
TIC 2022

Resumen...

- Historia de los datos
- Datos como recurso corporativo
- Bases de datos
- Evolución de los dispositivos de almacenamiento
- Arquitecturas que surgen dado el reto del *BigData*
 - Bases de datos
 - Análisis de datos



Internet



Continua...

Referencias

- [1] Guillenson. *Administración de Bases de Datos*. LIMUSA WILEY
- [2] Vincent Amouroux. Documental ¿el fin de la memoria?. 2016. Francia. Señal Colombia

<https://www.youtube.com/watch?v=BBrBOMTzWDo>

Gracias

?

eleonguz@unal.edu.co

