



Julio César Alonso C.

jcalonso@icesi.edu.co

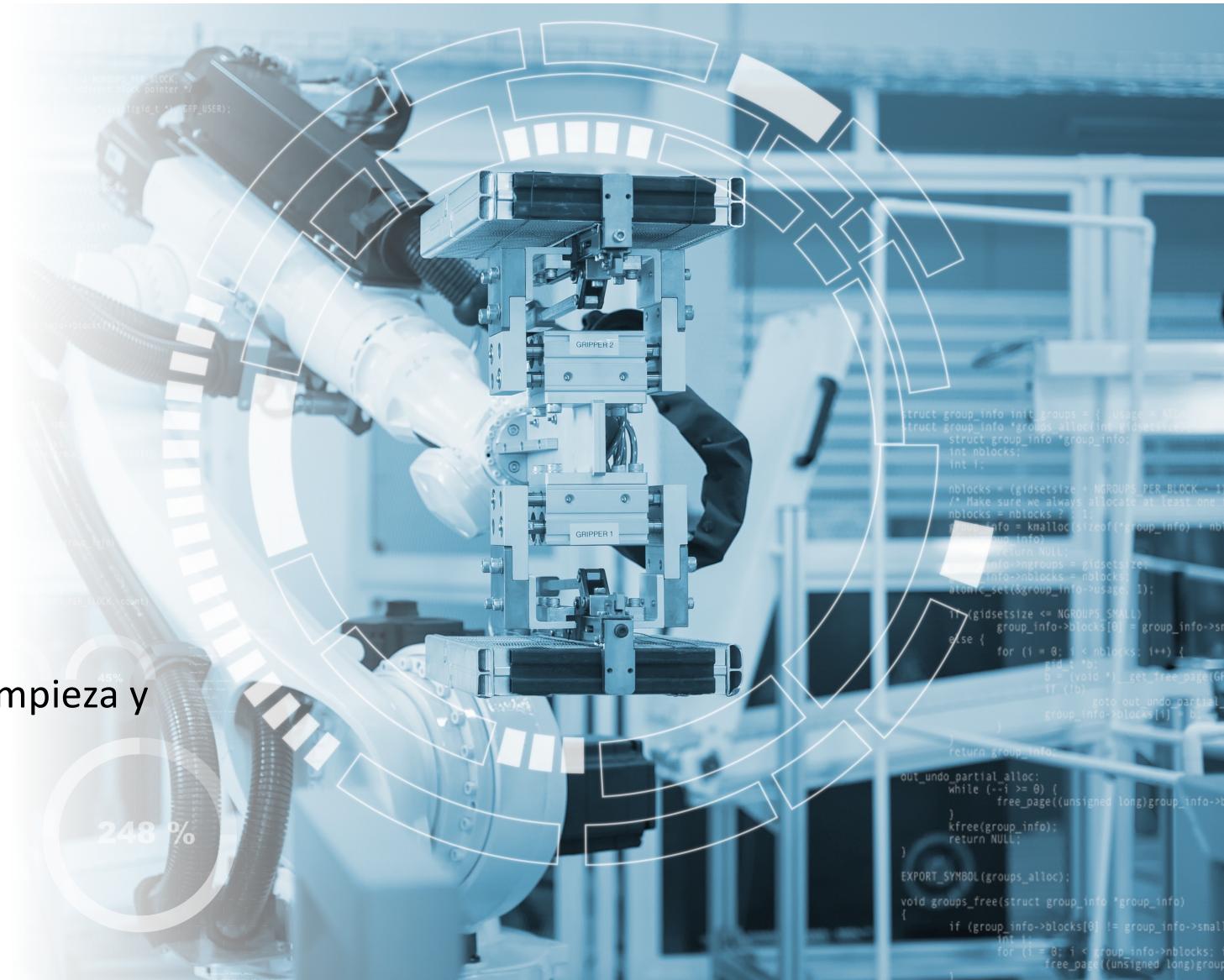
Sobre los datos, su limpieza y exploración

Introducción al Business Analytics

Unidad 4

Sobre los datos, su limpieza y exploración

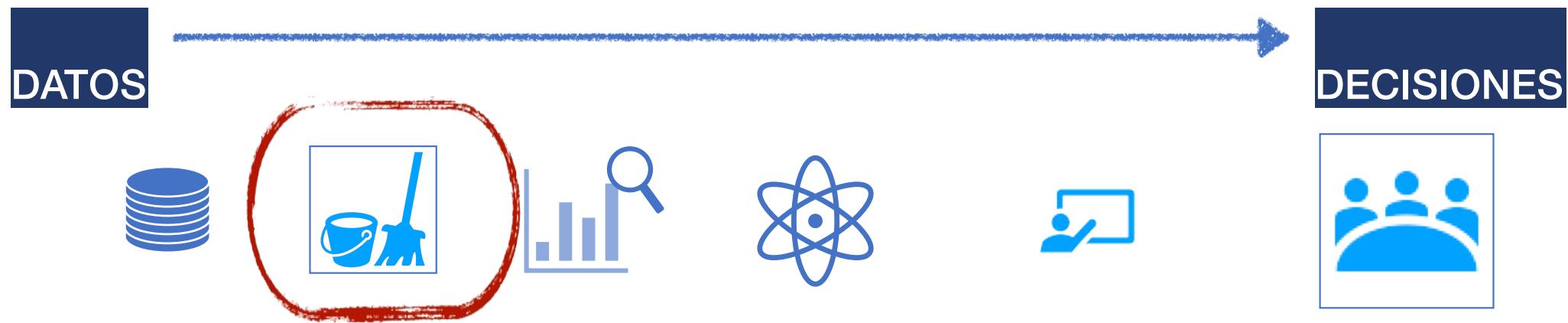
248 %



Objetivos

Al finalizar esta unidad el estudiante estará en capacidad de:

- Explicar en sus propias palabras las posibles fuentes de los datos que emplean las organizaciones para el proceso de ***business Analytics***.
- Explicar en sus propias palabras los conceptos de: base de datos, data warehouse, Data lake, ETL y Metadata.
- Realizar un proceso de limpieza inicial de una bases de datos en R.
- Explicar en sus propias palabras que tipos de problemas se pueden encontrar en una base de datos

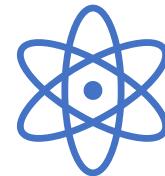
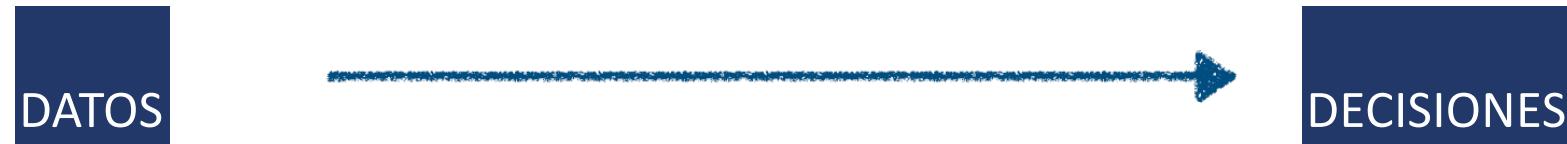


Un poco de vocabulario antes de entrar en el detalle técnico



Fuentes de Datos
en este mundo

Work flow



Fuentes de datos

De la organización

Eventos Web



Datos abiertos o públicos

Eventos Web

event_name	timestamp	user_id
homepage_visit	2019-01-01 12:01:01	1234

Grandes volúmenes de información



Datos del consumidor (solicitados)

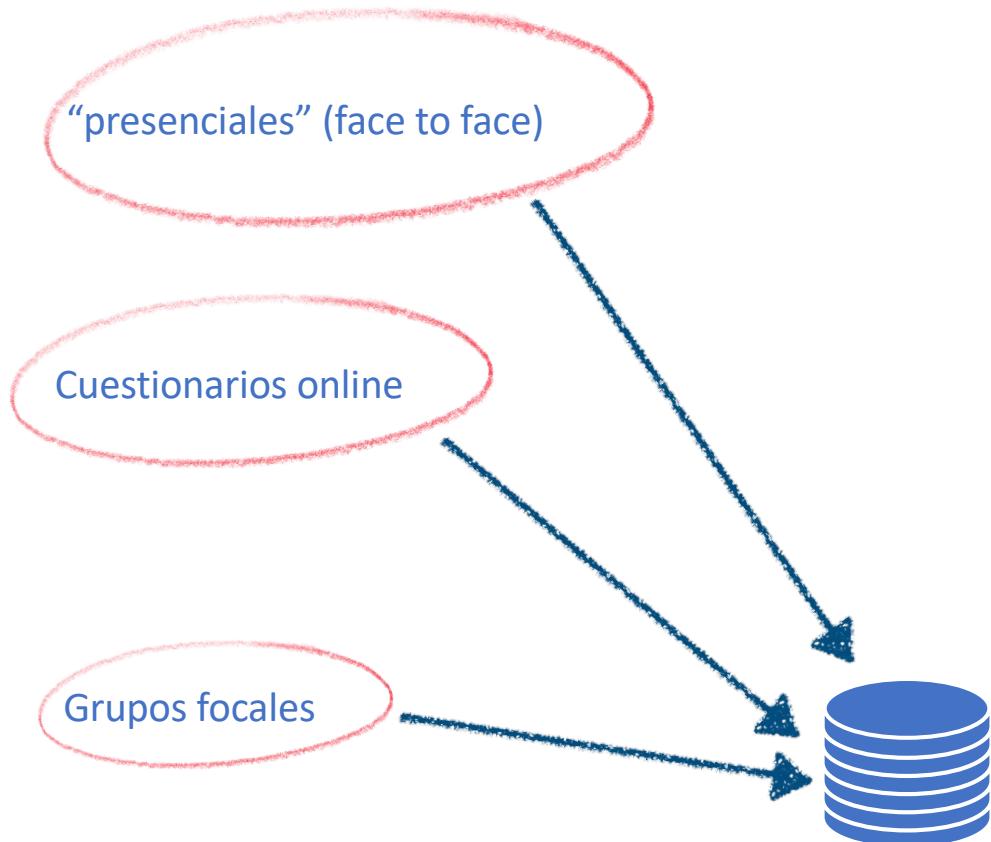
¿Para qué?

Crear “marketing collateral”

Tomar decisiones sin riesgos

Monitorear la calidad

Fuentes de datos solicitados



Fuentes de datos

De la organización

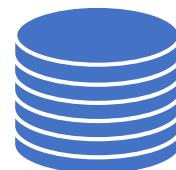
Eventos Web

Datos de clientes

Datos de logística

Transacciones financieras

Datos abiertos o públicos



Fuentes de datos

De la organización

Eventos Web

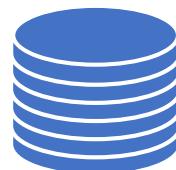
Datos de clientes

Datos de logística

Transacciones financieras

Datos abiertos o públicos

APIs (Application Programming Interface)

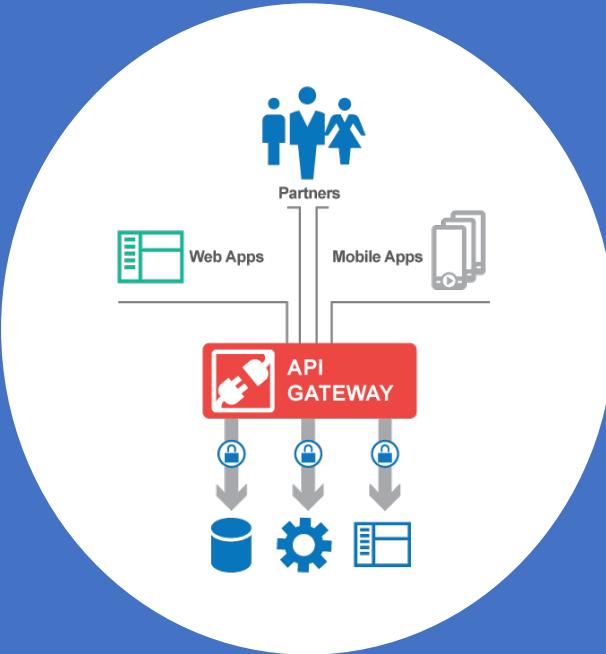


APIs

Permiten “chupar” información de internet

- Twitter
- Wikipedia
- Yahoo! Finance
- Google Maps

Las APIs permiten por ejemplo monitorear un hashtag #icesi



Una API (Application Programming Interface) es un conjunto de código que permite la transmisión de datos entre un producto de software y otro. (contiene las condiciones de este intercambio de datos).

Fuentes de datos

De la organización

Eventos Web

Datos de clientes

Datos de logística

Transacciones financieras

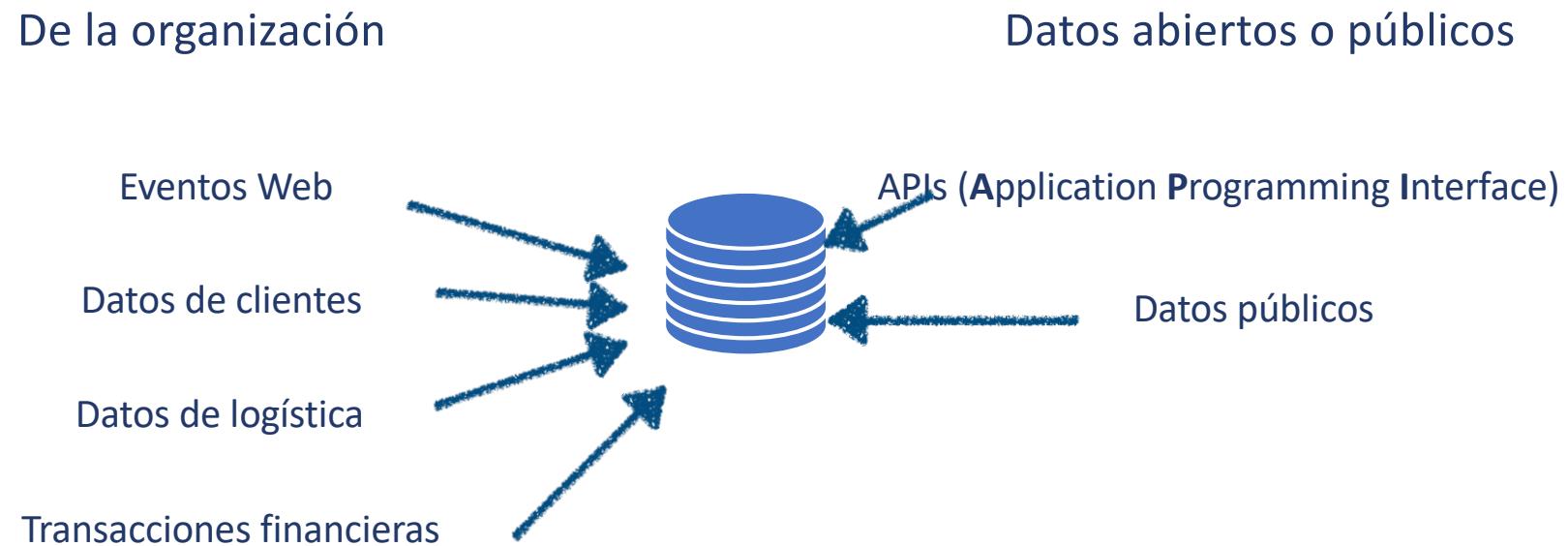


Datos abiertos o públicos

APIs (Application Programming Interface)

Datos públicos





Identificando Fuentes de datos





Tipos de Datos en
este mundo

¿Por qué nos importan los tipos de datos?

Visualización y análisis de los datos

Almacenamiento de los datos

Datos cuantitativos vs cualitativos

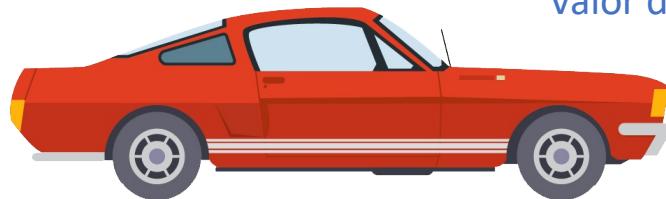
Cuantitativos

- Tienen un número naturalmente asociado
- Puede ser medido
- El número tiene un significado natural

Cualitativos

- Normalmente relacionado con descripciones
- Los datos son observables pero no medibles
- Si se asigna un número, no tiene significado natural

Datos cuantitativos



Valor de compra \$ 60 millones

2 ruedas

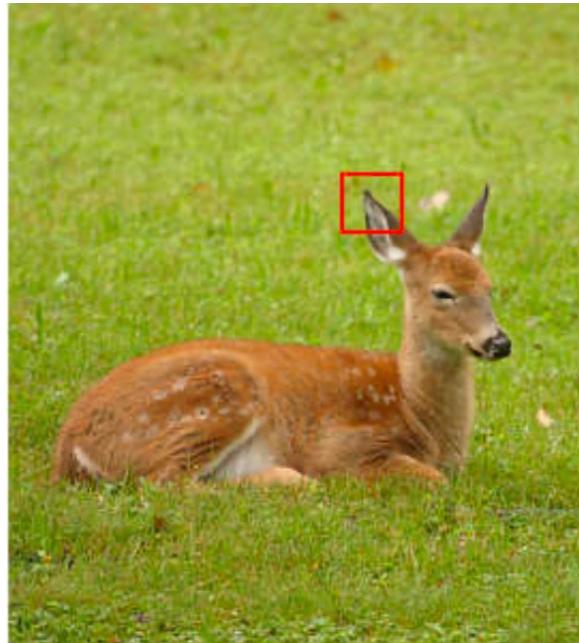
Velocidad máxima 120 km/h

Datos cualitativos



Pero hay más tipos de
datos en el mundo del
business analytics

Datos de imagen



Datos de Texto

★★★★★ 5 months ago

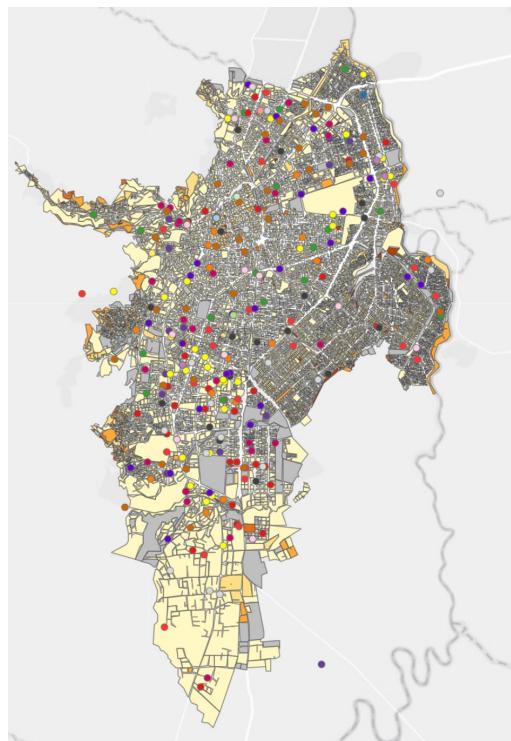
Estación de gasolina con servicio ágil, oportuno, confiable y con precios ajustados al mercado. Cuenta con un personal servicial, amable (Se esmeran en los detalles y te recuerdan aspectos importantes de revisión periódica, tales como nivel ... [More](#)



Response from the owner 5 months ago

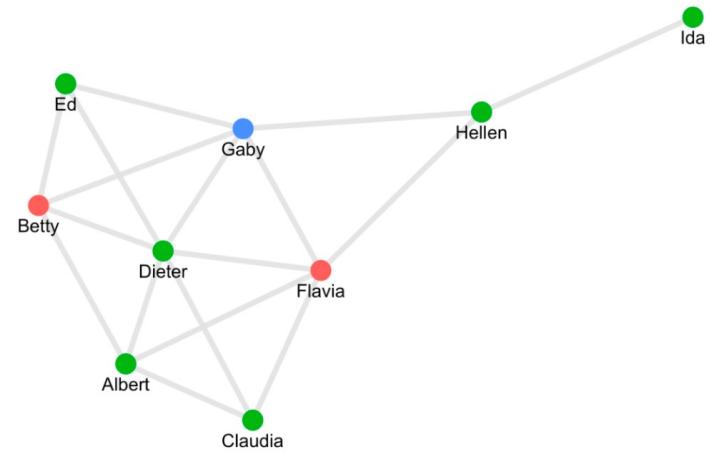
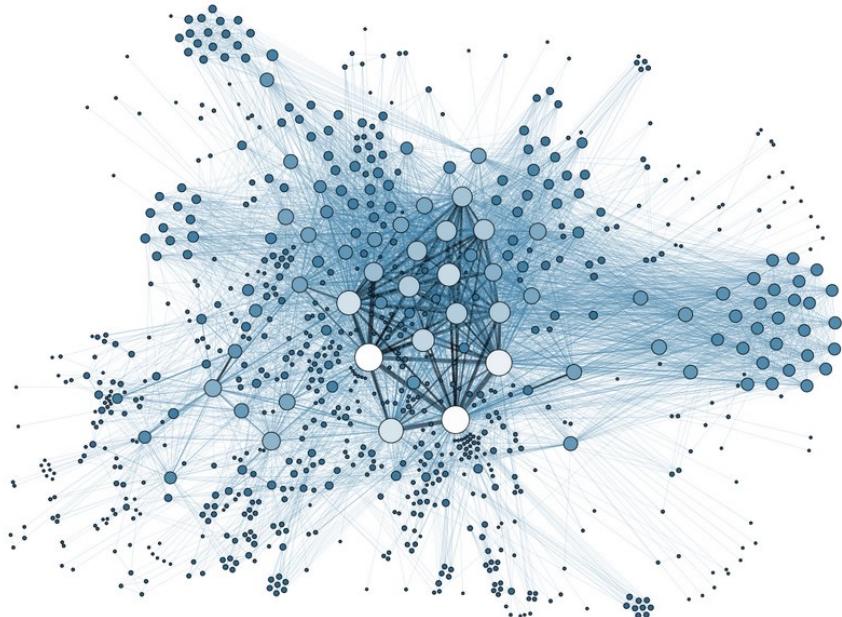
Hola, estamos muy contentos por tu visita y que haya decidido contarnos a todos la experiencia que tuviste en nuestras instalaciones. ¡Gracias por visitarnos y te esperamos pronto para que sigas eligiendo lo mejor para tu vehículo en un solo lugar!

Datos Geoespaciales



[Ver archivo anexo](#)

Datos de redes



Tipos de datos

- Quantitativos
- Qualitativos
- Imágenes
- Textos
- Geoespacial
- Redes
- Y muchos más

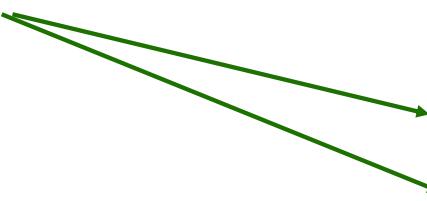
Tipos de datos

- Cuantitativos
- Cualitativos
- Imágenes
- Textos
- Geoespacial
- Redes
- Y muchos más

Y en R

Variables que se encuentran en objetos de clase *data.frame* o *tbl_df*

numeric: números reales o decimales
integer: números enteros



Tipos de datos

- Cuantitativos
- Cualitativos
- Imágenes
- Textos
- Geoespacial
- Redes
- Y muchos más

Y en R

Variables que se encuentran en objetos de clase *data.frame* o *tbl_df*

numeric: números reales o decimales

integer: números enteros

character: caracteres

logical: resultados lógicos (TRUE o FALSE)

factor: categórica (puede ser ordenada)

Tipos de datos





Los datos se pueden estar en diferentes formatos:

1. Datos estructurados,
2. Datos semiestructurados y
3. Datos no estructurados.

Datos estructurados:
se pueden almacenar en
filas y columnas (una
tabla)

Datos estructurados: “bien organizados”

Datos no estructurados:
no están organizados de
forma predefinida o no
tienen un modelo de
datos predefinido.

Ejemplo: un archivo de
Word, PDF, videos,
audios

Datos no estructurados:
son datos en bruto y no
organizados. Y a veces
dificilmente se podrán
organizar

Datos semi-
estructurados:
son datos poco
organizados.

Por ejemplo: tweets con
hashtags, información
de logs a un servidor



Almacenamiento
de los datos

Datos de sistemas internos



En una organización los datos tiene diferente origen

Aplicaciones o API's



Una base de datos es cualquier colección de datos organizada para su almacenamiento, accesibilidad y recuperación.

Otros orígenes



Datos de sistemas internos



data warehouse (DW)
(enterprise data warehouse (EDW))
(almacén de datos)



Aplicaciones o API's

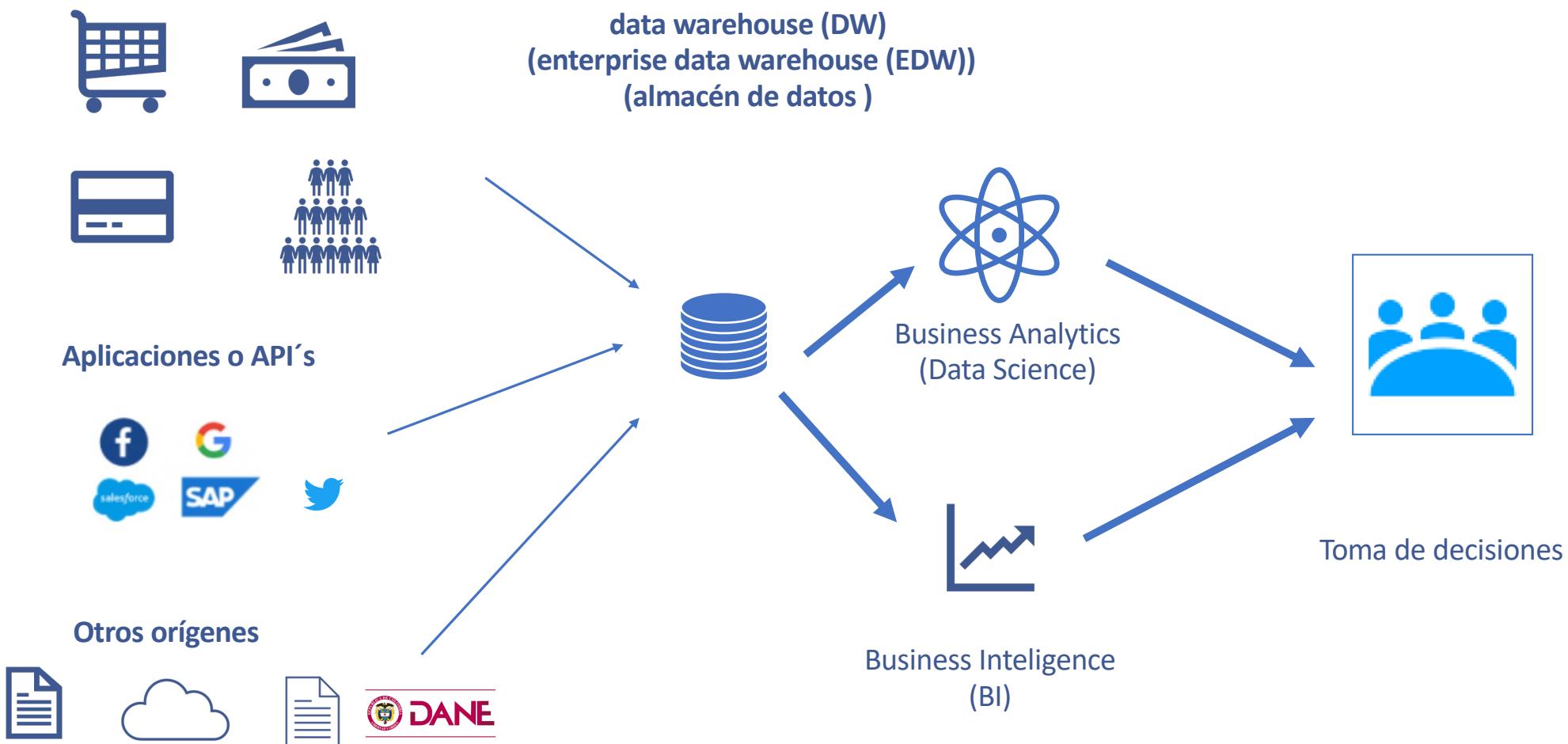


Otros orígenes



Un **Data Warehouse** es un repositorio de datos donde se almacenan después de ser extraídos de las bases de datos origen y transformados para evitar duplicidades, espacios en blanco e incoherencias. Su finalidad es crear informes y análisis de datos para la toma de decisiones

Datos de sistemas internos



Una **base de datos** es cualquier colección de datos organizada para su almacenamiento, accesibilidad y recuperación.

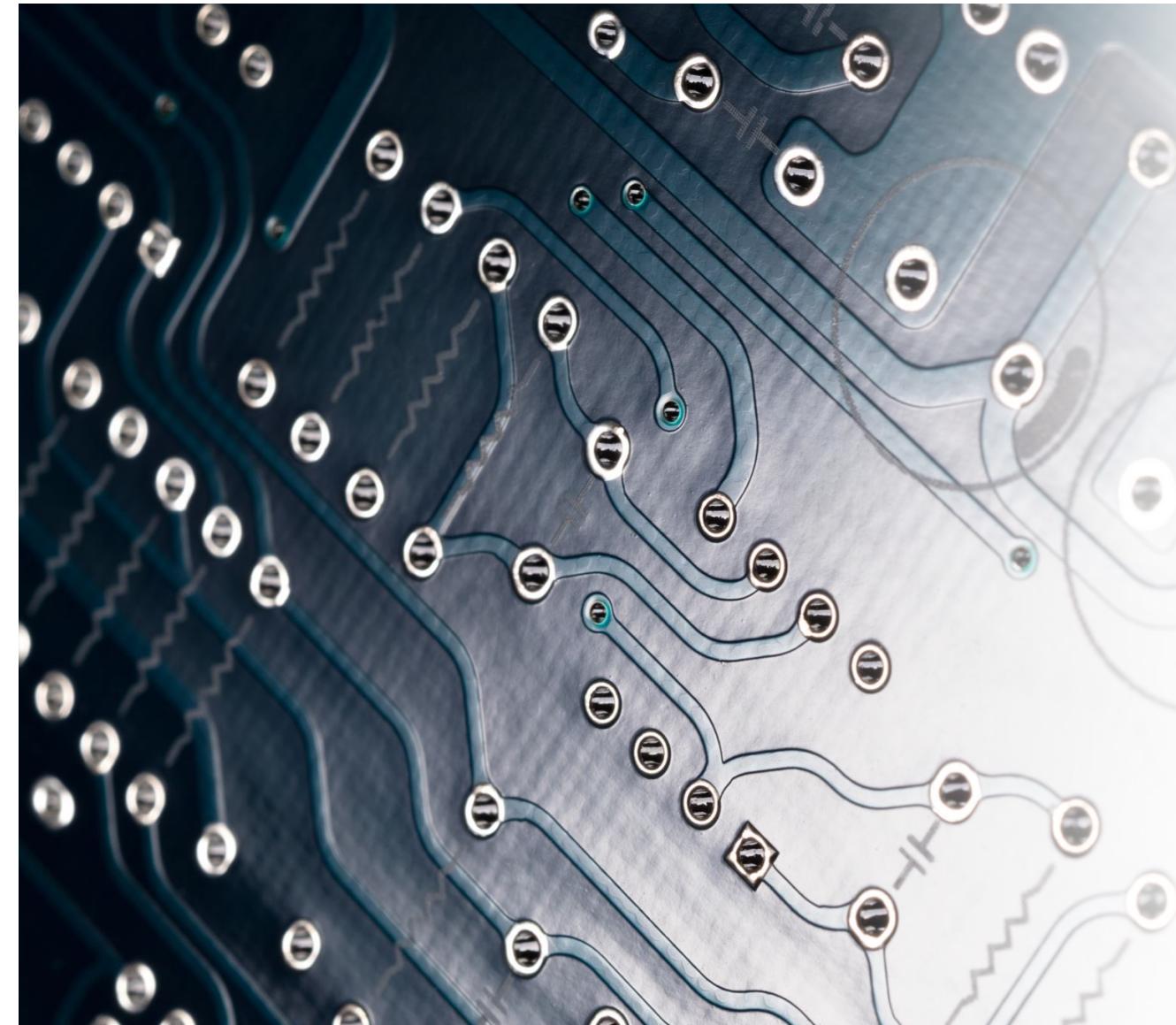
Un **Data Warehouse** es un tipo de base de datos que integra copias de datos procedentes de sistemas de origen dispares y los guarda para hacer BI o B Analytics

Un **Data Warehouse** es un repositorio de datos donde se almacenan después de ser extraídos de las bases de datos originales y transformados para evitar duplicidades, espacios en blanco e incoherencias. Su finalidad es crear informes y análisis de datos para la toma de decisiones

Tradicionalmente data estructurada



Un ***data lake*** es un repositorio de almacenamiento central que contiene ***big data*** de muchas fuentes en un formato crudo y granular.



Puede almacenar datos estructurados, semiestructurados o no estructurados, lo que significa que los datos pueden conservarse en un formato más flexible para su uso futuro.



ETL (extract,
transform, load):

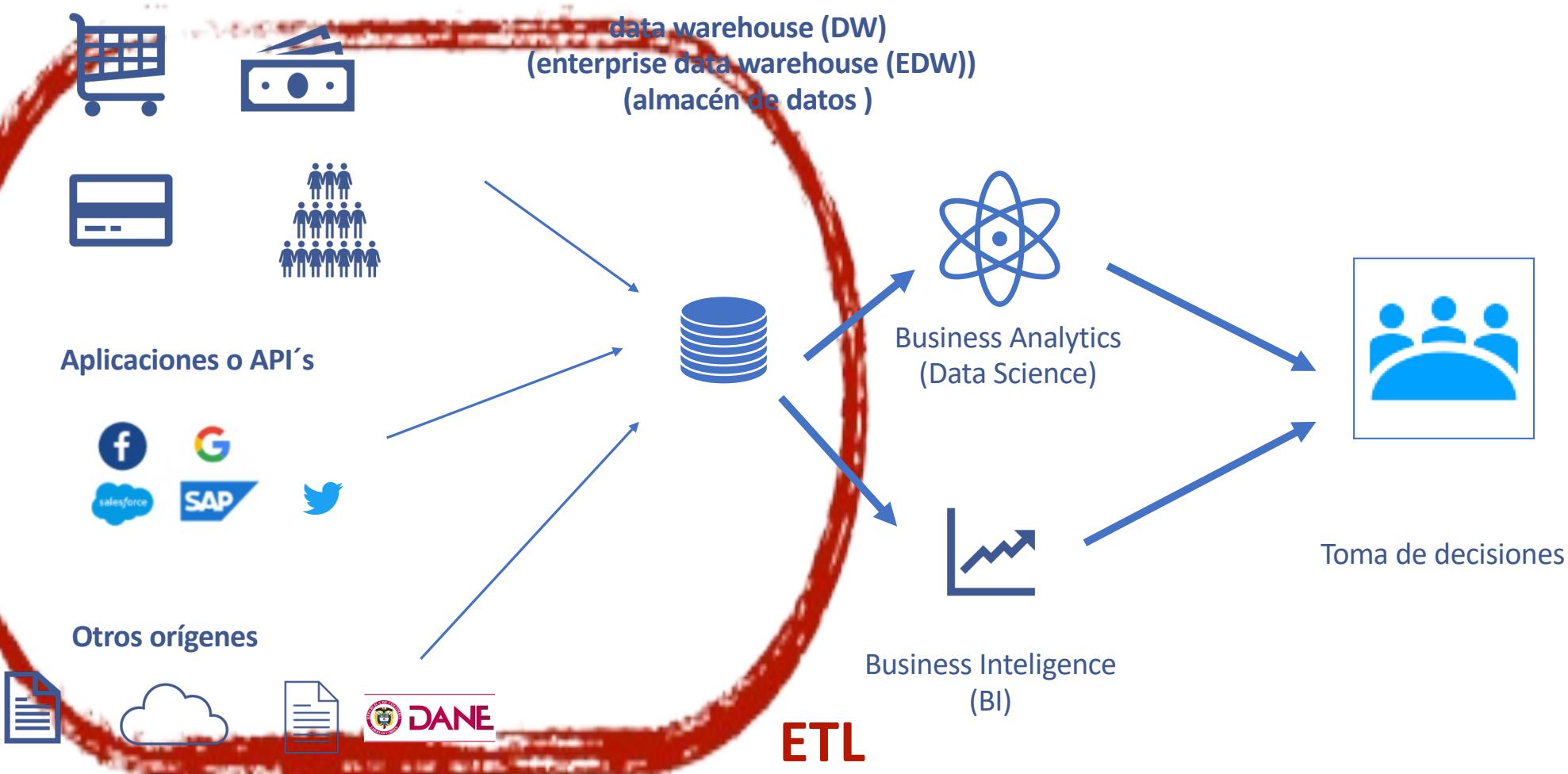
ETL (extract, transform, load):

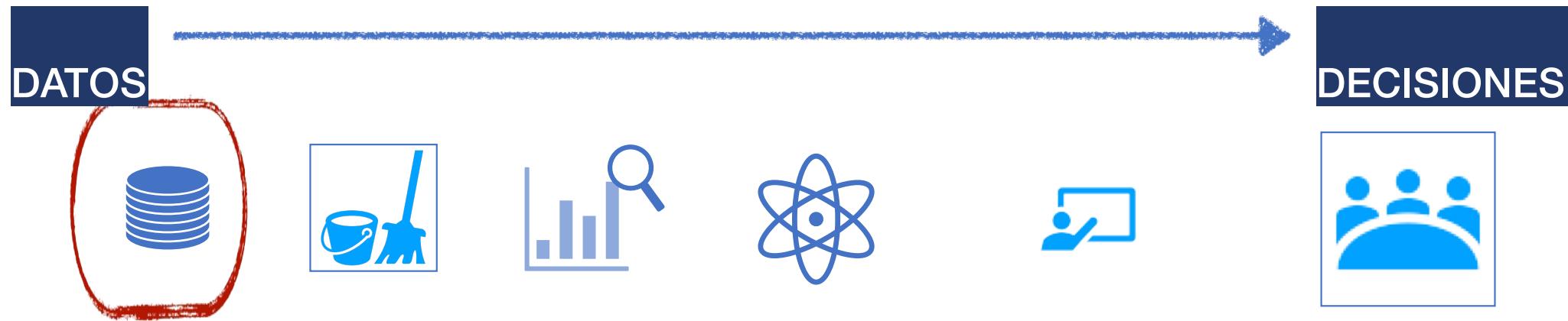
Extracción es el proceso de lectura de datos de una base de datos

Transformación es el proceso de convertir los datos extraídos de su forma anterior a la forma que deben tener para poder colocarlos en otra base de datos

La carga es el proceso de escribir los datos en la base de datos de destino.

Datos de sistemas internos





Nosotros nos
concentraremos en la
limpieza y exploración

Limpieza de datos - fases

Limpieza de datos - fases

Análisis de los datos

Definición y registro de los pasos de limpieza

Verificación

Transformación

Reemplazo de datos sucios con datos limpios

Veamos unos ejemplos
y cómo deberían
limpiarse los datos



No Correspondencia entre el formato de la variable y su tipo

Edad	Edad en R	Estado Civil	Estado Civil
4	"4"	1	1
5	"5"	4	4
6	"6"	5	5
8	"8"	2	2
12	"12"	3	3

1=Casado
 2=Unión Libre
 3=Soltero
 4=Viudo
 5=Divorciado

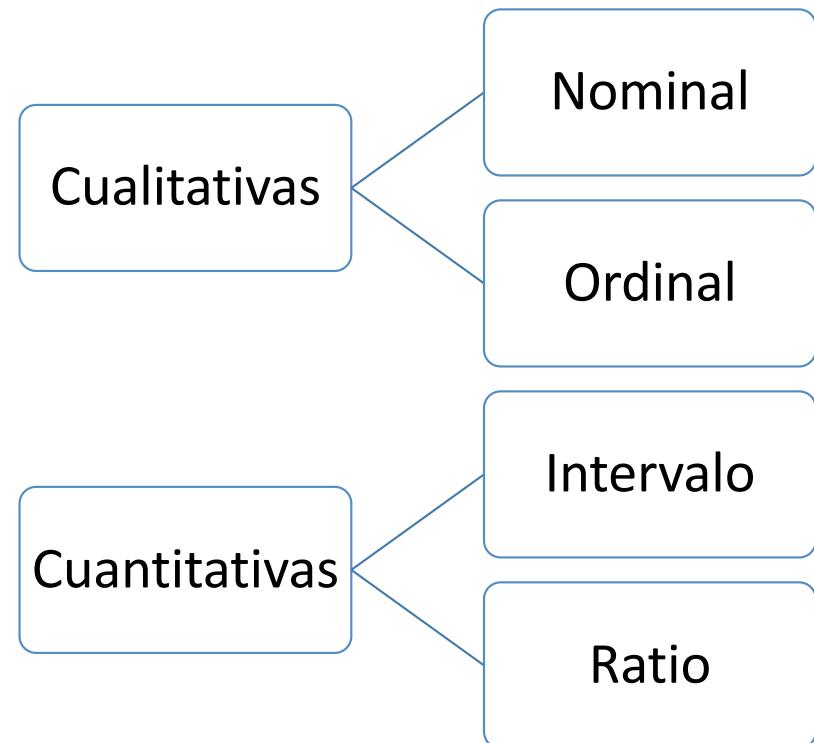
No Correspondencia entre el formato de la variable y su tipo

Solución

Convertir los datos a su respectivo formato.

Funciones

- as.character
- as.numeric
- as.integer
- as.factor
- as.Date



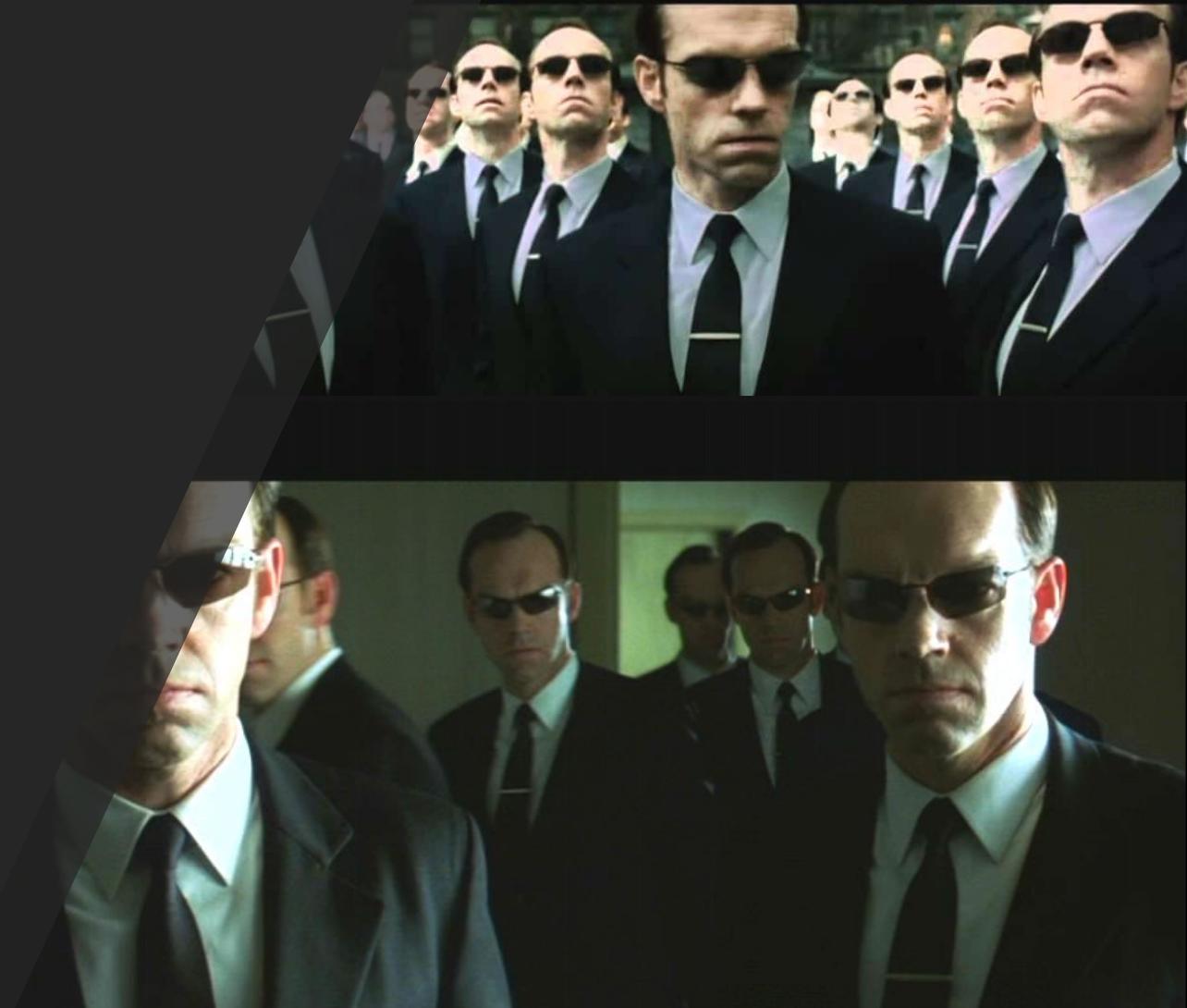
Observaciones duplicadas

¿Por qué este paso es importante?

- Para entender mejor la base de datos
- Evitar la doble contabilización

Solución

- Eliminar las observaciones duplicadas*
- Revisar si en el proceso de carga de datos o de extracción de datos se cometió algún error.
- Al final del proceso



Valores perdidos



“ ”

99



NA

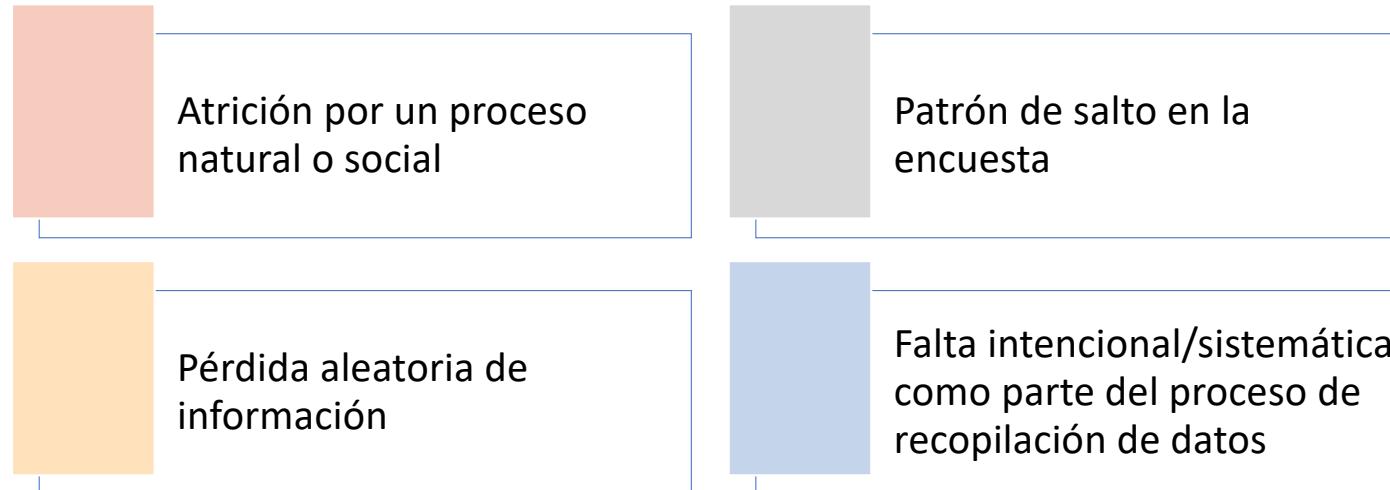
1. Identifique donde se encuentran los valores perdidos y si es necesario recodifiquelos correctamente.

Tip: Maneje el mismo código para indicar que falta un valor en un campo de datos.

2. Intente entender la naturaleza de los datos perdidos.

3. Determine cuál es el mejor método para tratarlos.

Entender el porqué hay valores perdidos



Tratamiento de los valores perdidos

1. Reconocer su existencia y no efectuar tratamiento alguno. Pero tenerlo en cuenta en la aplicación de los modelos y el análisis.

Tratamiento de los valores perdidos

2. Eliminar observaciones con valores perdidos

Gender	8 th grade math test score	12 th grade math score
F	45	.
M	.	99
F	55	86
F	85	88
F	80	75
.	81	82
F	75	80
M	95	.
M	86	90
F	70	75
F	85	.

Gender	8 th grade math test score	12 th grade math score
F	45	.
M	.	99
F	55	86
F	85	88
F	80	75
.	81	82
F	75	80
M	95	.
M	86	90
F	70	75
F	85	.

Fuente: https://liberalarts.utexas.edu/prc/_files/cs/Missing-Data.pdf

Tratamiento de los valores perdidos

3. Imputación de datos: sustituir valores perdidos en un campo por otros empleando:

- Media
- Mediana
- Moda
- Predicción de esos valores faltantes empleado modelos estadísticos

Errores de digitación

Convenciones de denominación

- NYC vs New York

Representaciones diferentes

- Si, si, Sí

Espacios vacíos

- “Mujer” vs “ Mujer ”

¿Soluciones?

Valores inconsistentes

Estos valores se hallan analizando la consistencia de las respuestas entre variables que se encuentran relacionadas.

Edad	Fecha de nacimiento	Edad calculada
27	1990/01/17	27
21	1995/05/14	21
60	11/11/1956	60
47	7/04/1940	77
37	12/08/1980	37

- La detección de estos valores inconsistentes dependerán del conocimiento y la lógica del analista.

Valores sin referencia en el diccionario de variables

Nombre	Estado Civil
Paola	1
Esteban	4
Cesar	5
Carlos	8
María	3
Mayra	2
Roberto	1

Código Estado Civil	
1	Soltero
2	Casado
3	Separado o divorciado
4	Unión libre
5	Viudo

¿Soluciones?

Importancia del diccionario de variables

Metadata

Datos que describen las características de otros datos

employee_id	first_name	last_name	nin	department_id
44	Simon	Martinez	HH 45 09 73 D	1
45	Thomas	Goldstein	SA 75 35 42 B	2
46	Eugene	Comelsen	NE 22 63 82	2
47	Andrew	Petculescu	XY 29 87 61 A	1
48	Ruth	Stadick	MA 12 89 36 A	15
49	Barry	Scardelis	AT 20 73 18	2
50	Sidney	Hunter	HW 12 94 21 C	6
51	Jeffrey	Evans	LX 13 26 39 B	6
52	Doris	Berndt	YA 49 88 11 A	3
53	Diane	Eaton	BE 08 74 68 A	1
54	Bonnie	Hall	WW 53 77 68 A	15
55	Taylor	Li	ZE 55 22 80 B	1

Metadata

Column	Data Type	Description
employee_id	int	Primary key of a table
first_name	nvarchar(50)	Employee first name
last_name	nvarchar(50)	Employee last name
nin	nvarchar(15)	National Identification Number
position	nvarchar(50)	Current position title, e.g. Secretary
department_id	int	Employee department. Ref: Departments
gender	char(1)	M = Male, F = Female, Null = unknown
employment_start_date	date	Start date of employment in organization.
employment_end_date	date	Employment end date. Null if employee st

Diccionario de variable (*Codebook*)

Un Diccionario de variable es una descripción técnica de los datos. Describe cómo se organizan los datos en el archivo, qué significan los diferentes números y letras, y cualquier instrucción especial sobre cómo usar los datos correctamente.

Características de un buen diccionario de variables:

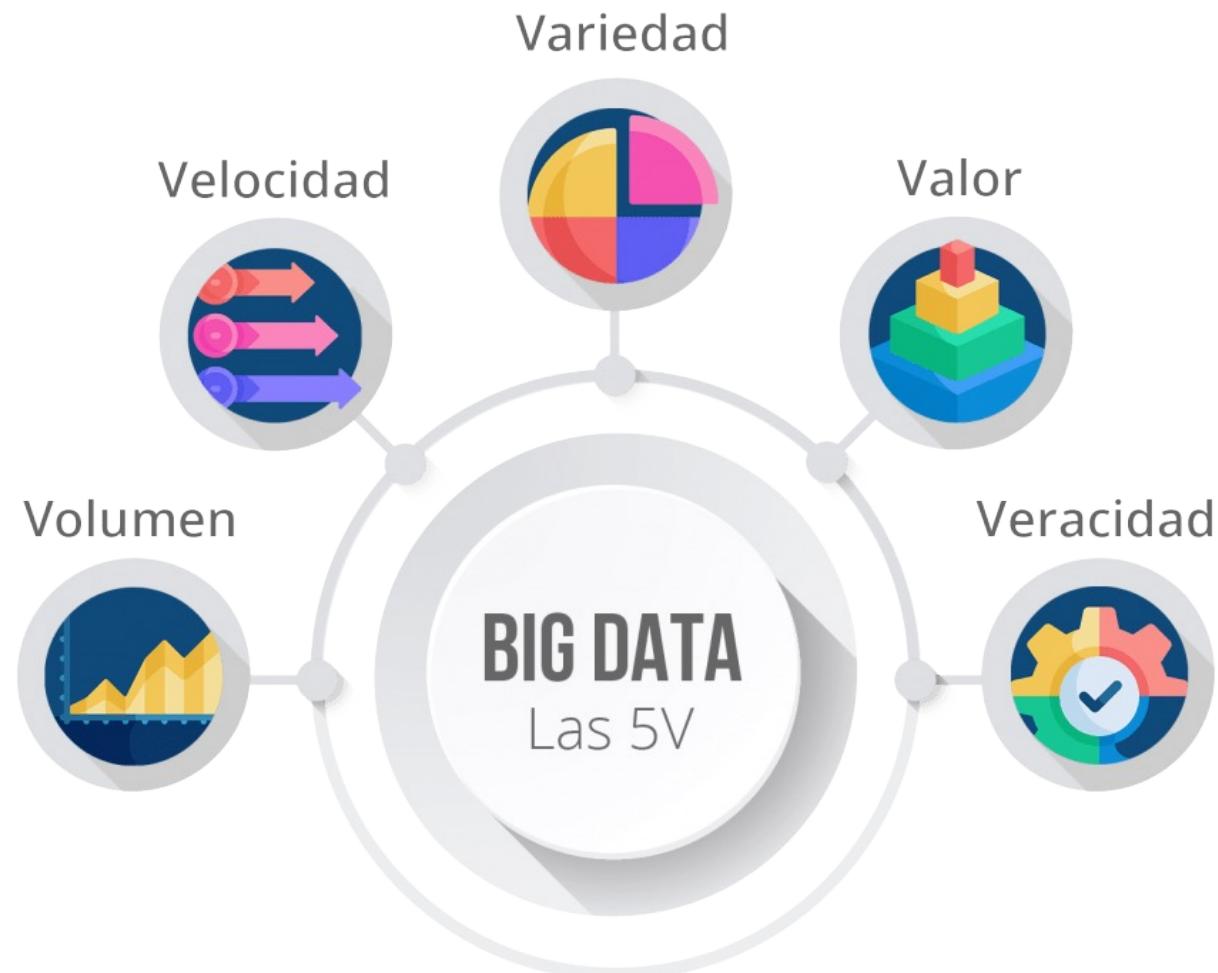
- Descripción del estudio: quién lo hizo, por qué lo hicieron, cómo lo hicieron.
- Información muestral: cuál fue la población estudiada, cómo se extrajo la muestra, cuál fue la tasa de respuesta.
- Información técnica sobre los archivos en sí: número de observaciones, longitud de registro, número de registros por observación, etc.
- Estructura de los datos dentro del archivo: jerárquico, etc.
- Detalles sobre los datos: columnas en las que se pueden encontrar variables específicas, ya sean de carácter o numéricas, y si son numéricas, qué formato.
- Texto de las preguntas y respuestas: algunos incluso tienen cuántas personas respondieron de una manera particular.

Big Data

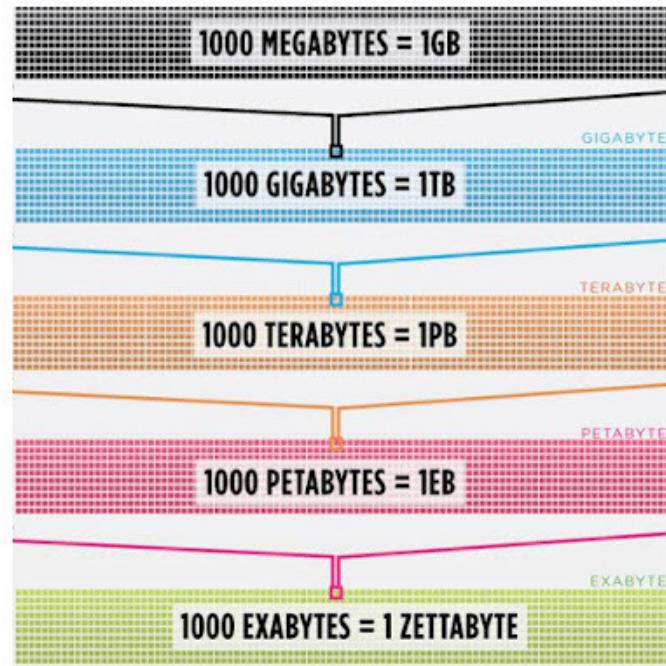
grandes volúmenes de
datos que son muy
variados y veloces

Es muy complicado
capturarlos y
procesarlos con
métodos tradicionales

los datos deben cumplir
con las 5V (8V)



El *volumen* de datos
tiende a ser inmenso



Fuente: <http://tecnoenlaces.blogspot.com/2013/05/que-es-un-kilo-mega-giga-tera-peta-exa.html>

How big is a **Yottabyte?**

TERABYTE

Will fit 200,000 photos or mp3 songs on a single 1 terabyte hard drive.



PETABYTE

Will fit on 16 Backblaze storage pods racked in two datacenter cabinets.



EXABYTE

Will fit in 2,000 cabinets and fill a 4 story datacenter that takes up a city block.



ZETTABYTE

Will fill 1,000 datacenters or about 20% of Manhattan, New York.

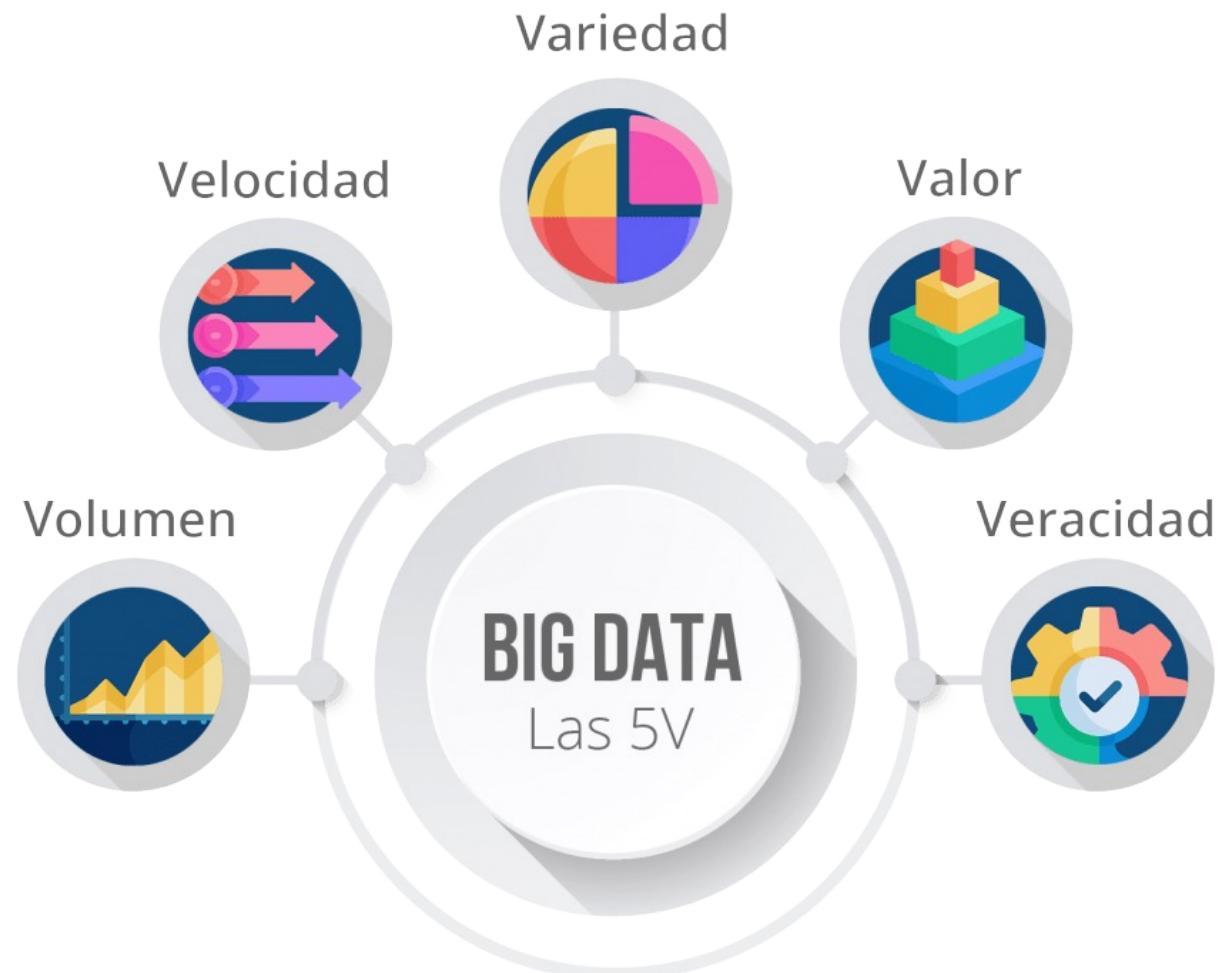


YOTTABYTE

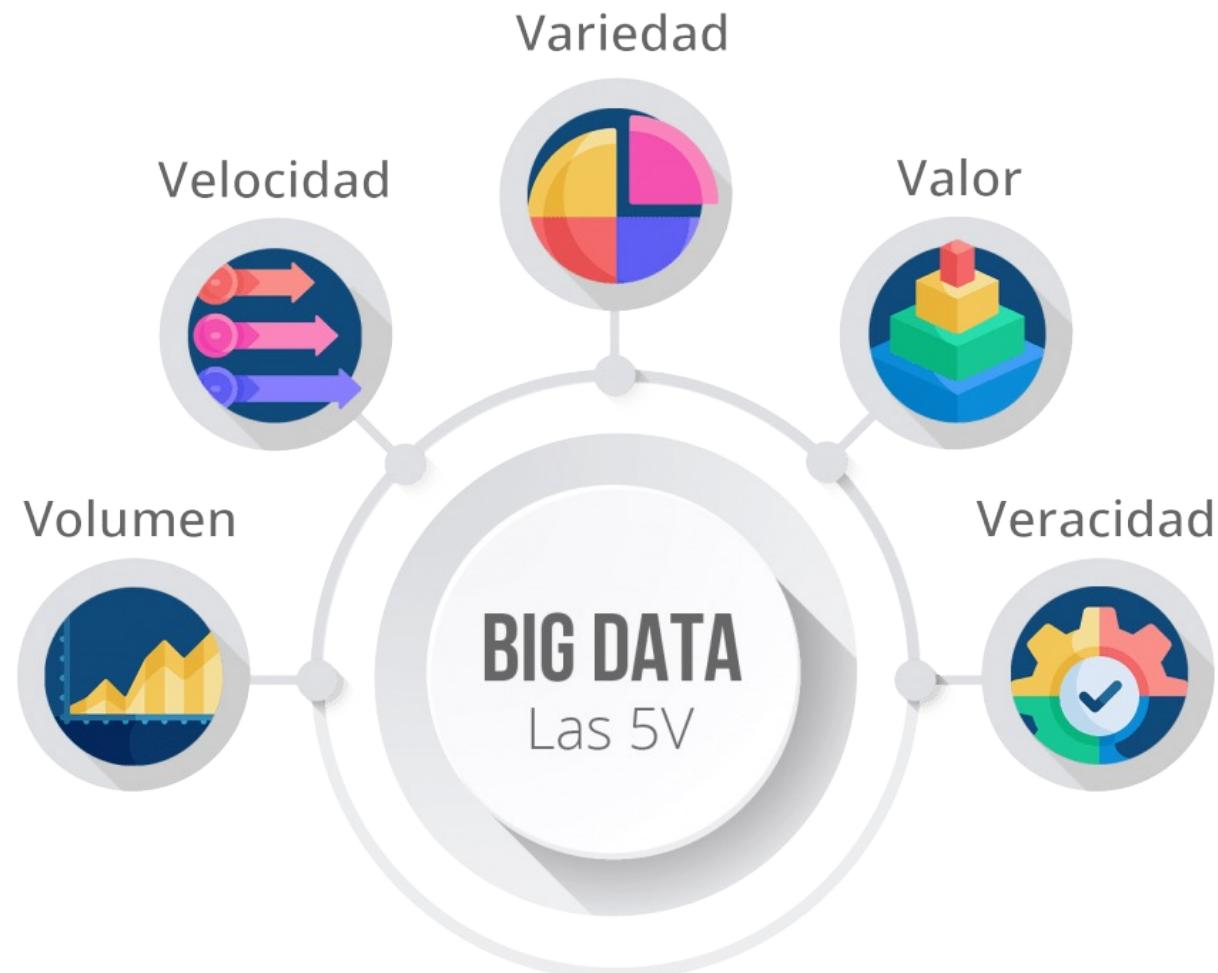
Will fill the states of Delaware and Rhode Island with a million datacenters.



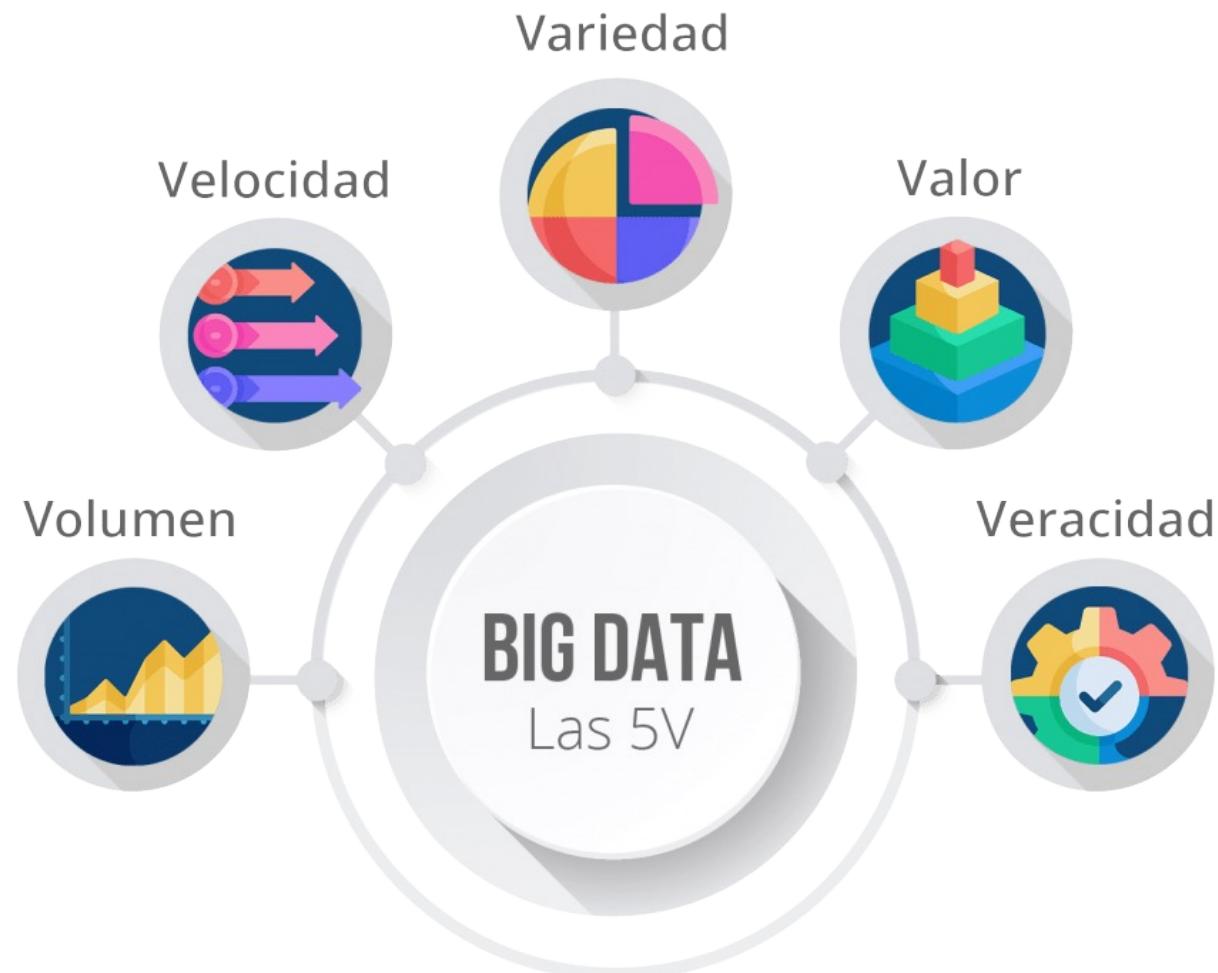
Fuente: <https://byteme404.wordpress.com/2010/05/31/kilo-mega-tera-peta-exa-zetta-ÿotta/>



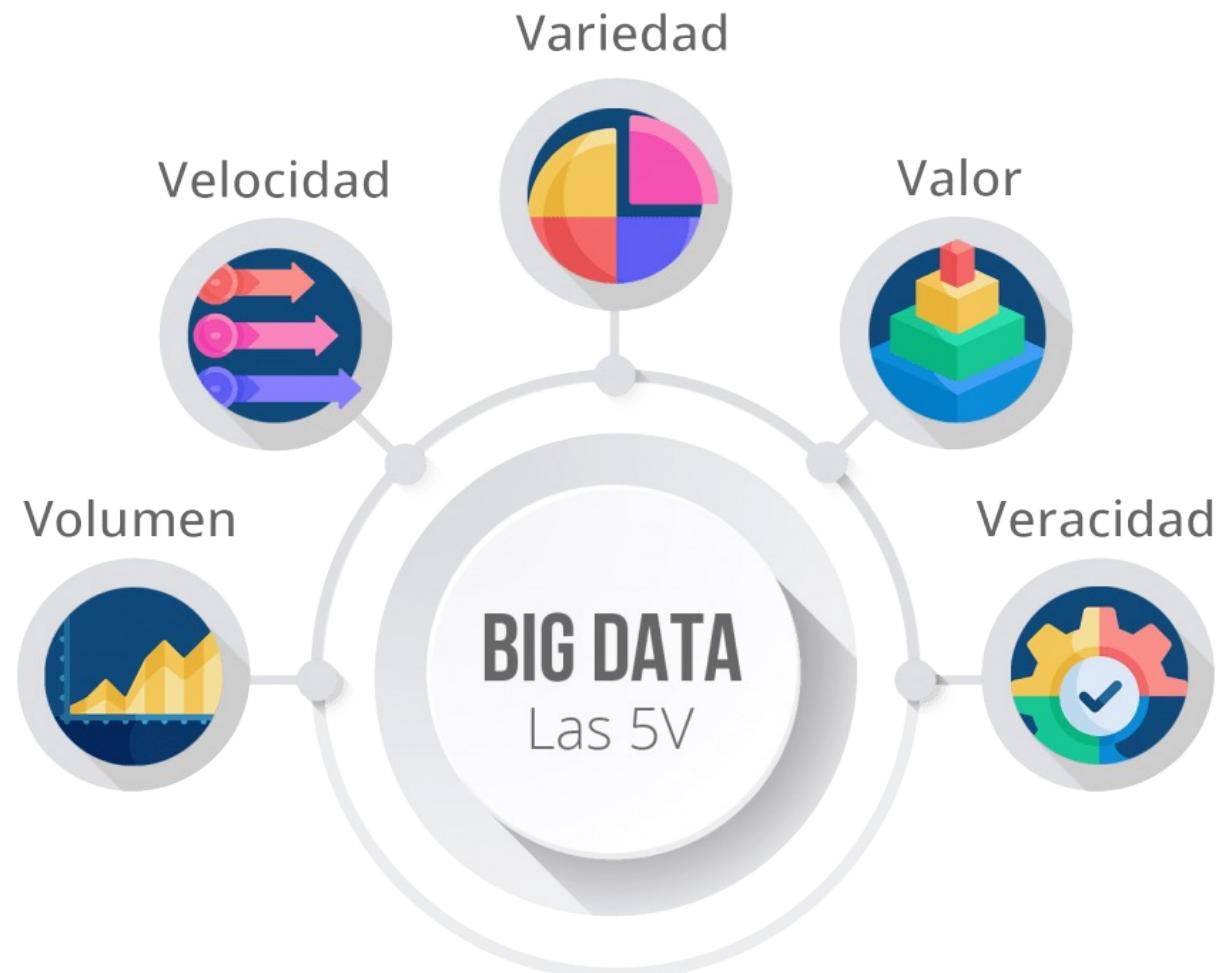
La *velocidad* a la que se generan es muy alta



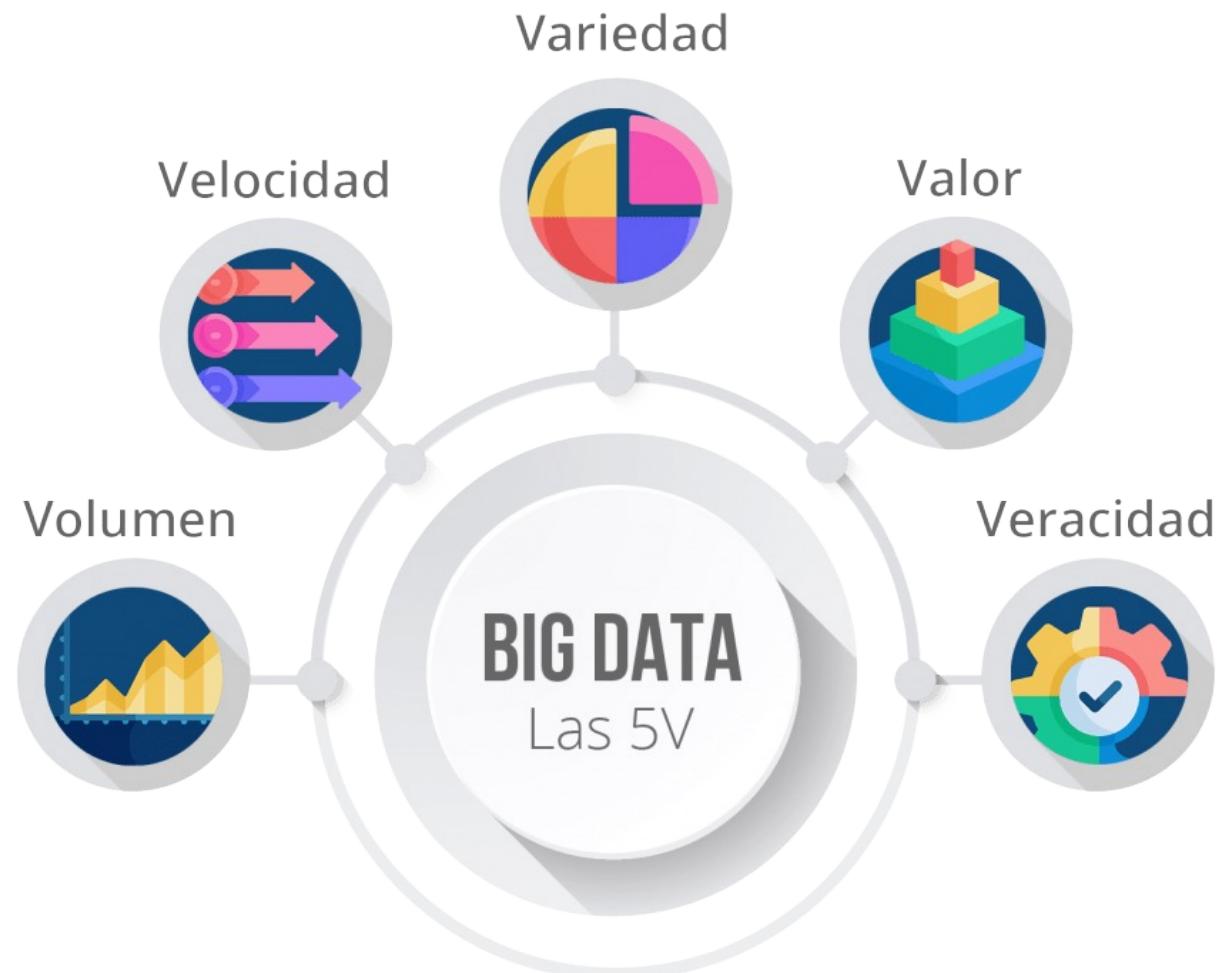
Todo tipo de datos, ya sea estructurados o no estructurados. (tablas, texto, imágenes, videos, audio) (*Variedad*)



Los datos deben poder proporcionar un *valor* beneficio a la organización.



Se debe garantizar la calidad y confiabilidad de los datos. El Big Data debe alimentarse con datos relevantes y verdaderos. (veracidad)



Julio César Alonso C.
jcalonso@icesi.edu.co

Sobre los datos, su limpieza y exploración

Introducción al Business Analytics