



UNIVERSIDAD
PANAMERICANA

Introducción a la Ciencia de Datos

Dr. Leon Felipe Palafox Novack
lpalafox@up.edu.mx

0

Noticias del día

Qué ha pasado en el mundo de Data
Science?

Dell Medical School Launches Data Hub to Accelerate Biomedical Research, Advance Health

Aug. 7, 2018



Email



Facebook



Twitter



LinkedIn



Google+



More

AUSTIN, Texas – **Dell Medical School at The University of Texas at Austin** is accelerating innovation and research by creating a Biomedical Data Science Hub to help solve complex research and clinical problems.

Imagine having a complicated scientific question: How do we predict who will be diagnosed with Type

1

Anuncios parroquiales

Proyecto Final



Proyecto Final

El objetivo es que trabajen en el proyecto a lo largo del curso.

Conforme vayamos aprendiendo las herramientas, se recomienda las practiquen con sus propios datos.

Vayan formando sus equipos oportunamente.

- Los tutoriales de Python van a estar disponibles toda la clase
- El machote para el reporte final ya esta en la pagina web.

2

Datos

Tenemos que entender nuestra
materia prima

“ The world's most important
resource is no longer oil, is data

The Economist

Qué es un dato?



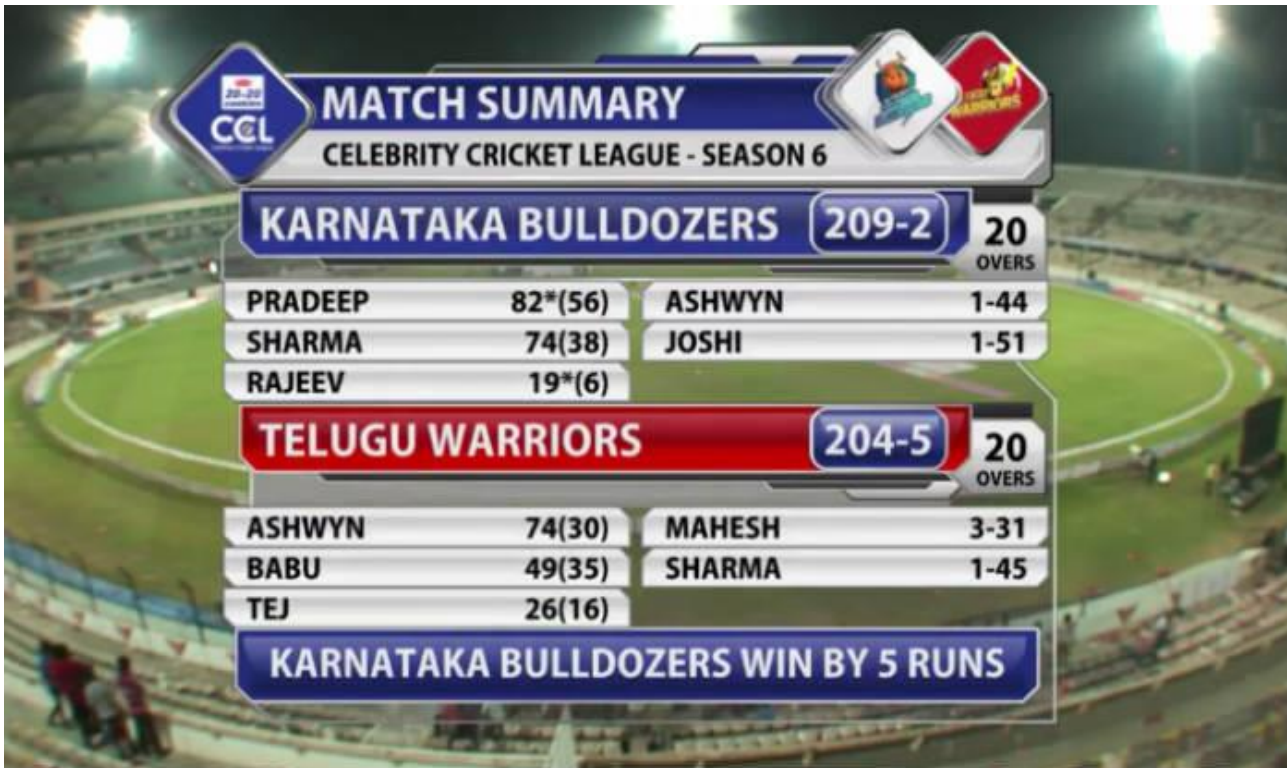
■ Es la unidad mínima de información?

- ▷ Es aquello, que dado un contexto, nos puede proveer de información.
 - ▷ El contexto por lo general se puede inferir.
- ▷ Ejemplos:
 - ▷ Una foto
 - ▷ Un marcador



UNIVERSIDAD
PANAMERICANA





The image shows a match summary overlay for the Celebrity Cricket League (CCL) Season 6. The background is a night-time view of a cricket stadium with floodlights. The overlay is a semi-transparent box with a blue and red color scheme. It displays the match details, player statistics, and the final result.

MATCH SUMMARY			
CELEBRITY CRICKET LEAGUE - SEASON 6			
KARNATAKA BULLDOZERS		209-2	20 OVERS
PRADEEP	82*(56)	ASHWYN	1-44
SHARMA	74(38)	JOSHI	1-51
RAJEEV	19*(6)		
TELUGU WARRIORS		204-5	20 OVERS
ASHWYN	74(30)	MAHESH	3-31
BABU	49(35)	SHARMA	1-45
TEJ	26(16)		
KARNATAKA BULLDOZERS WIN BY 5 RUNS			

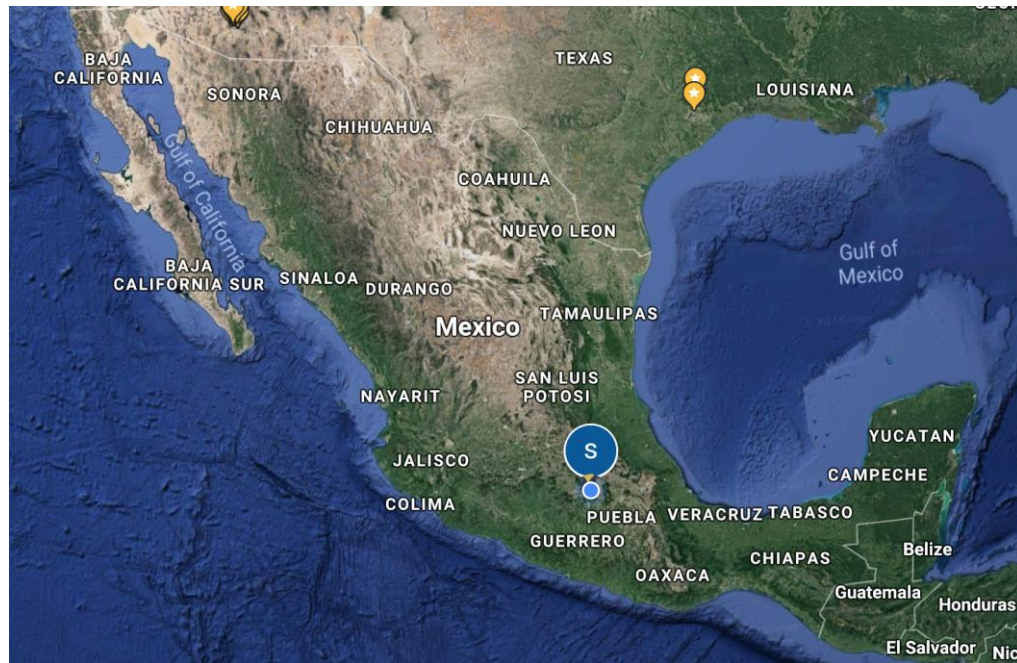
Cuantos tipos de datos hay?



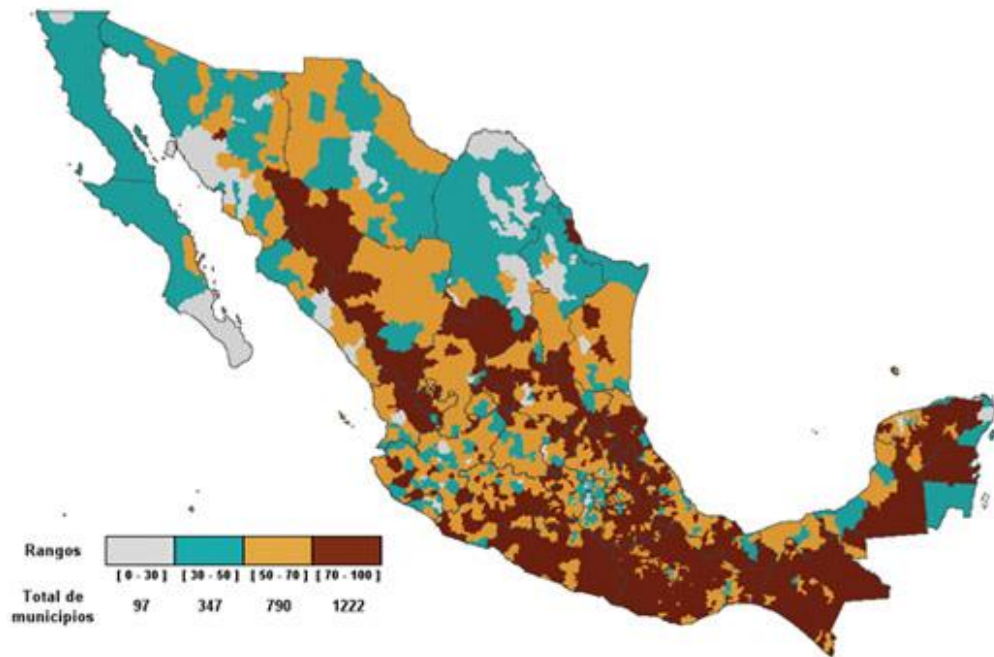
■ Hay muchas definiciones acerca de los tipos de datos que hay:

- ▷ Continuos y discretos
- ▷ Big and Small
- ▷ Aqui vamos a hablar de estructurados, no estructurados y semi estructurados.

Datos Estructurados y No estructurados



Datos Estructurados y No estructurados



**Pobreza a nivel
municipio (CONEVAL)**

Datos No Estructurados



- Se estima que el grueso de los datos que existen son no estructurados
- Mucho valor del negocio esta escondido en datos no estructurados
- Imagenes, Audio, Video

Datos no Estructurados



- Son datos que cuentan con una estructura pre-definida.
 - ▷ Estructuras de Datos (También llamados modelos de datos)
 - ▷ Colas, Listas, Arreglos, Pilas (stack)
 - ▷ Tablas como hojas de cálculo, un archivo de excel, etc
 - ▷ Mapas cartográficos tipo GIS (Google Earth)

Datos Estructurados



- Son aquellos que organizan elementos de los datos y tienen patrones de como se relacionan uno con otro, y como se relacionan con el resto del mundo.
- También se dice que un dato estructurado es aquel que se asocia a un modelo de dato.
 - ▶ Un modelo de dato da estructura y capacidad de relación al dato

Datos Estructurados

Food Inventory Sheet - Microsoft Excel

File

Home

Insert

Page Layout

Formulas

Data

Review

View

Team

Normal

Page Layout

Page Break Preview

Custom Views

Full Screen

Ruler

Formula Bar

Gridlines

Headings

Zoom 100%

Zoom to Selection

New Window

Arrange All

Freeze Panes

Split

Hide

Unhide

View Side by Side

Synchronous Scrolling

Reset Window Position

Save Workspace

Switch Windows

Macros

Workbook Views

Show

Zoom

Window

Macros

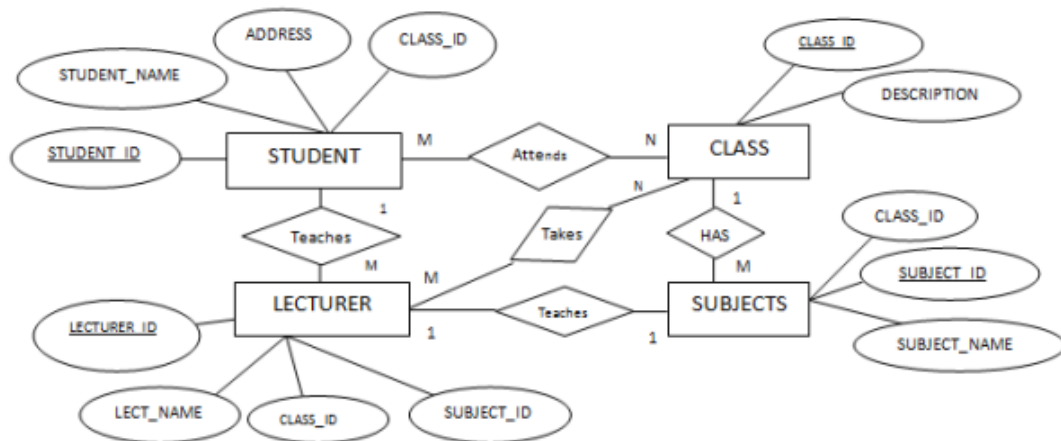
H8

</

Datos Estructurados

Bases de Datos

► Diagrama Entidad Relación

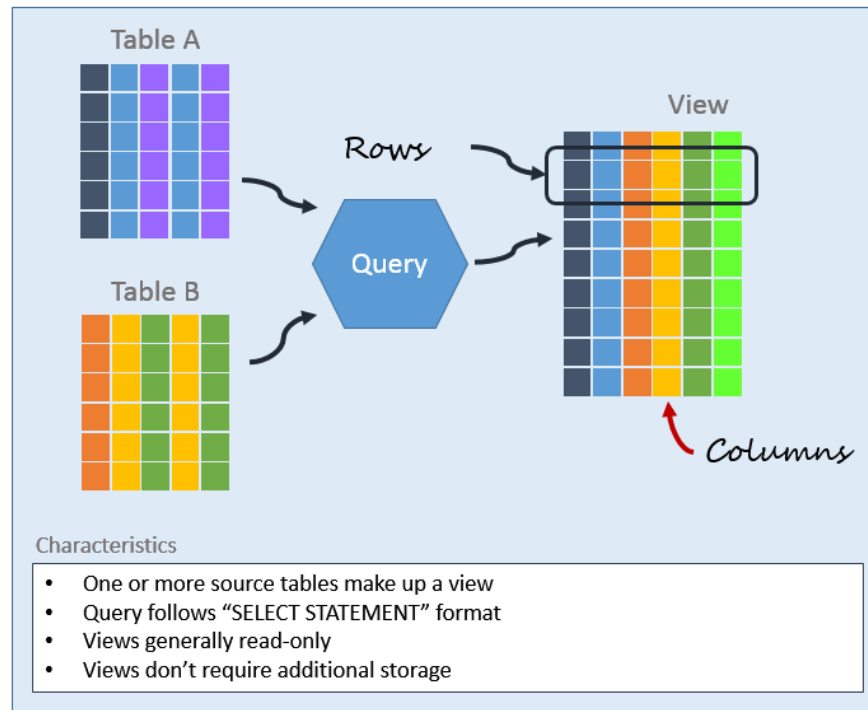


Base de Datos

UNIVERSIDAD
PANAMERICANA[illegible][illegible][illegible]

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	DESCRIPTION	CATEGORY	UNIT	PRICE	MARKING	STATUS	PRICE	PRICE	PRICE	PRICE	PRICE	PRICE	PRICE	PRICE	PRICE	PRICE	PRICE	PRICE	PRICE	PRICE	PRICE	PRICE	PRICE	PRICE	PRICE
2	82222Y TURKEY BREAST 1/2	Meat & Poultry	100g	0.00			0.00																		
3	82222Y TURKEY BREAST 1/2	Meat & Poultry	500g	0.00			0.00																		
4	82222Y TURKEY BREAST 1/2	Meat & Poultry	1kg	0.00			0.00																		
5	754545 MARGARITA LUNCH CREAMY	Meat & Poultry	100g	0.00			0.00																		
6	754545 MARGARITA LUNCH CREAMY	Meat & Poultry	500g	0.00			0.00																		
7	754545 MARGARITA LUNCH CREAMY	Meat & Poultry	1kg	0.00			0.00																		
8	860006 LUSH BAKED SLOTTED	Meat & Poultry	100g	0.00			0.00																		
9	860006 LUSH BAKED SLOTTED	Meat & Poultry	500g	0.00			0.00																		
10	860006 LUSH BAKED SLOTTED	Meat & Poultry	1kg	0.00			0.00																		
11	747676 PANCAKE MIX	Meat & Poultry	100g	0.00			0.00																		
12	747676 PANCAKE MIX	Meat & Poultry	500g	0.00			0.00																		
13	747676 PANCAKE MIX	Meat & Poultry	1kg	0.00			0.00																		
14	800000 MILD MEXICAN	Meat & Poultry	100g	0.00			0.00																		
15	800000 MILD MEXICAN	Meat & Poultry	500g	0.00			0.00																		
16	800000 MILD MEXICAN	Meat & Poultry	1kg	0.00			0.00																		
17	700000 CHICKEN MARINADE	Meat & Poultry	100g	0.00			0.00																		
18	700000 CHICKEN MARINADE	Meat & Poultry	500g	0.00			0.00																		
19	700000 CHICKEN MARINADE	Meat & Poultry	1kg	0.00			0.00																		
20	700000 CHICKEN MARINADE	Meat & Poultry	5kg	0.00			0.00																		
21	700000 CHICKEN MARINADE	Meat & Poultry	10kg	0.00			0.00																		
22	700000 CHICKEN MARINADE	Meat & Poultry	20kg	0.00			0.00																		
23	700000 CHICKEN MARINADE	Meat & Poultry	50kg	0.00			0.00																		
24	700000 CHICKEN MARINADE	Meat & Poultry	100kg	0.00			0.00																		
25	700000 CHICKEN MARINADE	Meat & Poultry	200kg	0.00			0.00																		
26	700000 CHICKEN MARINADE	Meat & Poultry	500kg	0.00			0.00																		
27	700000 CHICKEN MARINADE	Meat & Poultry	1000kg	0.00			0.00																		
28	700000 CHICKEN MARINADE	Meat & Poultry	2000kg	0.00			0.00																		
29	700000 CHICKEN MARINADE	Meat & Poultry	5000kg	0.00			0.00																		
30	700000 CHICKEN MARINADE	Meat & Poultry	10000kg	0.00			0.00																		
31	700000 CHICKEN MARINADE	Meat & Poultry	20000kg	0.00			0.00																		
32	700000 CHICKEN MARINADE	Meat & Poultry	50000kg	0.00			0.00																		
33	700000 CHICKEN MARINADE	Meat & Poultry	100000kg	0.00			0.00																		
34	700000 CHICKEN MARINADE	Meat & Poultry	200000kg	0.00			0.00																		
35	700000 CHICKEN MARINADE	Meat & Poultry	500000kg	0.00			0.00																		
36	700000 CHICKEN MARINADE	Meat & Poultry	1000000kg	0.00			0.00																		
37	700000 CHICKEN MARINADE	Meat & Poultry	2000000kg	0.00			0.00																		
38	700000 CHICKEN MARINADE	Meat & Poultry	5000000kg	0.00			0.00																		
39	700000 CHICKEN MARINADE	Meat & Poultry	10000000kg	0.00			0.00																		
40	700000 CHICKEN MARINADE	Meat & Poultry	20000000kg	0.00			0.00																		
41	700000 CHICKEN MARINADE	Meat & Poultry	50000000kg	0.00			0.00																		
42	700000 CHICKEN MARINADE	Meat & Poultry	100000000kg	0.00			0.00																		
43	700000 CHICKEN MARINADE	Meat & Poultry	200000000kg	0.00			0.00																		
44	700000 CHICKEN MARINADE	Meat & Poultry	500000000kg	0.00			0.00																		
45	700000 CHICKEN MARINADE	Meat & Poultry	1000000000kg	0.00			0.00																		
46	700000 CHICKEN MARINADE	Meat & Poultry	2000000000kg	0.00			0.00																		
47	700000 CHICKEN MARINADE	Meat & Poultry	5000000000kg	0.00			0.00																		
48	700000 CHICKEN MARINADE	Meat & Poultry	10000000000kg	0.00			0.00																		
49	700000 CHICKEN MARINADE	Meat & Poultry	20000000000kg	0.00			0.00																		
50	700000 CHICKEN MARINADE	Meat & Poultry	50000000000kg	0.00			0.00																		
51	700000 CHICKEN MARINADE	Meat & Poultry	100000000000kg	0.00			0.00																		
52	700000 CHICKEN MARINADE	Meat & Poultry	200000000000kg	0.00			0.00																		
53	700000 CHICKEN MARINADE	Meat & Poultry	500000000000kg	0.00			0.00																		
54	700000 CHICKEN MARINADE	Meat & Poultry	1000000000000kg	0.00			0.00																		
55	700000 CHICKEN MARINADE	Meat & Poultry	2000000000000kg	0.00			0.00																		
56	700000 CHICKEN MARINADE	Meat & Poultry	5000000000000kg	0.00			0.00																		
57	700000 CHICKEN MARINADE	Meat & Poultry	10000000000000kg	0.00			0.00																		
58	700000 CHICKEN MARINADE	Meat & Poultry	20000000000000kg	0.00			0.00																		
59	700000 CHICKEN MARINADE	Meat & Poultry	50000000000000kg	0.00			0.00																		
60	700000 CHICKEN MARINADE	Meat & Poultry	100000000000000kg	0.00			0.00																		
61	700000 CHICKEN MARINADE	Meat & Poultry	200000000000000kg	0.00			0.00																		
62	700000 CHICKEN MARINADE	Meat & Poultry	500000000000000kg	0.00			0.00																		
63	700000 CHICKEN MARINADE	Meat & Poultry	1000000000000000kg	0.00			0.00																		
64	700000 CHICKEN MARINADE	Meat & Poultry	2000000000000000kg	0.00			0.00																		
65	700000 CHICKEN MARINADE	Meat & Poultry	5000000000000000kg	0.00			0.00																		
66	700000 CHICKEN MARINADE	Meat & Poultry	10000000000000000kg	0.00			0.00																		
67	700000 CHICKEN MARINADE	Meat & Poultry	20000000000000000kg	0.00			0.00																		
68	700000 CHICKEN MARINADE	Meat & Poultry	50000000000000000kg	0.00			0.00																		
69	700000 CHICKEN MARINADE	Meat & Poultry	100000000000000000kg	0.00			0.00																		
70	700000 CHICKEN MARINADE	Meat & Poultry	200000000000000000kg	0.00			0.00																		
71	700000 CHICKEN MARINADE	Meat & Poultry	500000000000000000kg	0.00			0.00																		
72	700000 CHICKEN MARINADE	Meat & Poultry	1000000000000000000kg	0.00			0.00																		
73	700000 CHICKEN MARINADE	Meat & Poultry	2000000000000000000kg	0.00			0.00																		
74	700000 CHICKEN MARINADE	Meat & Poultry	5000000000000000000kg	0.00			0.00																		
75	700000 CHICKEN MARINADE	Meat & Poultry	10000000000000000000kg	0.00			0.00																		
76	700000 CHICKEN MARINADE	Meat & Poultry	20000000000000000000kg	0.00			0.00																		
77	700000 CHICKEN MARINADE	Meat & Poultry	50000000000000000000kg	0.00			0.00																		
78	700000 CHICKEN MARINADE	Meat & Poultry	100000000000000000000kg	0.00			0.00																		
79	700000 CHICKEN MARINADE	Meat & Poultry	200000000000000000000kg	0.00			0.00																		
80	700000 CHICKEN MARINADE	Meat & Poultry	500000000000000000000kg	0.00			0.00										</								

Anatomy of a View



- Para que querriamos una vista?
- Por que es importante conocer el diagrama entidad-relación?
- Que otras cosas necesita una base de datos?

■ Componentes de una base de datos:

- ▷ Diagrama Entidad – Relación
- ▷ Catálogo de Datos
- ▷ Bitácora de Datos

Datos semi-estructurados

- Hay varios formatos, pero en la industria estan los dos mas usados:

▶ JSON

```
{  
  "Rail Booking": {  
    "reservation": {  
      "ref_no": 1234567,  
      "time_stamp": "2016-06-24T14:26:59.125",  
      "confirmed": true  
    },  
    "train": {  
      "date": "07/04/2016",  
      "time": "09:30",  
      "from": "New York",  
      "to": "Chicago",  
      "seat": "57B"  
    },  
    "passenger": {  
      "name": "John Smith"  
    },  
    "price": 1234.25,  
    "comments": ["Lunch & dinner incl.", "\"Have a nice day!\""]  
  }  
}
```

Datos semi-estructurados

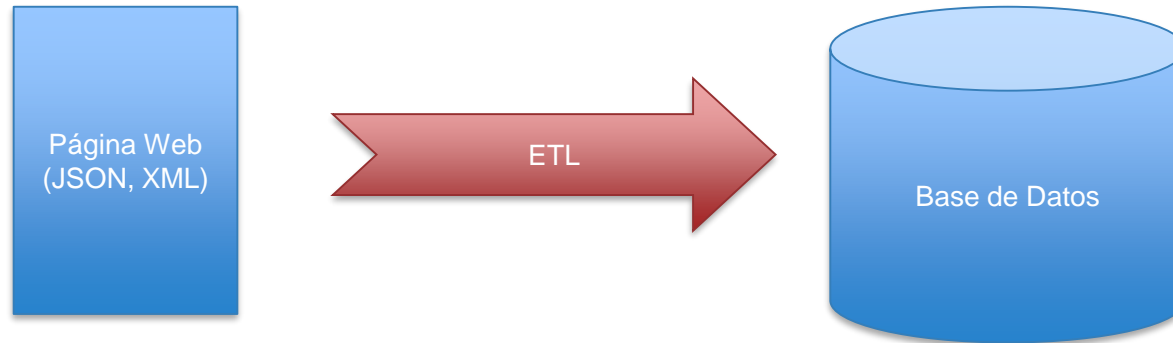
XML

```
<?xml version="1.0"?>
<quiz>
  <qanda seq="1">
    <question>
      Who was the forty-second
      president of the U.S.A.?
    </question>
    <answer>
      William Jefferson Clinton
    </answer>
  </qanda>
  <!-- Note: We need to add
  more questions later.-->
</quiz>
```

XML

Usos de JSON y XML

- Se utilizan primordialmente para transportar datos de un lado al otro



■ Extract, Transform and Load:

- ▶ Que creen que hace un ETL?
- ▶ Por que necesitamos ETLs?
- ▶ Creen que hay negocio haciendo ETLs?

- Un ETL (Extract, Transform, Load) se encarga de extraer los datos de una fuente, y transformarlos para que se coloquen en otra.
- Dichas transformaciones son por lo general simples y no involucran algoritmos complejos.
- Un uso muy común de un ETL es cuando se migran bases de datos o arquitecturas.

Data Lake vs Data Warehouse

- En la industria de Big Data hay dos conceptos:
 - ▷ Data Lake: Almacena TODOS los datos que se pueda, estructurados y no estructurados.
 - ▷ Data Warehouse: Almacena sólo datos estructurados. Tiene un esquema mucho mas estricto acerca de que tipo de datos se van a almacenar.
- En un Data Lake es muy dificil encontrar información, mientras que en un Data Warehouse es fácil. (No simple, sólo facil)

Como transformar datos?



- Que se necesita para hacer que un dato no estructurado tenga estructura?
 - ▷ Que se necesita para que su album de fotos tenga estructura?
 - ▷ Que se necesita para que un libro tenga estructura?
 - ▷ Que se necesita para que una canción tenga estructura?

Introducción a Python

La herramienta del Data Scientist