Tema 2.1

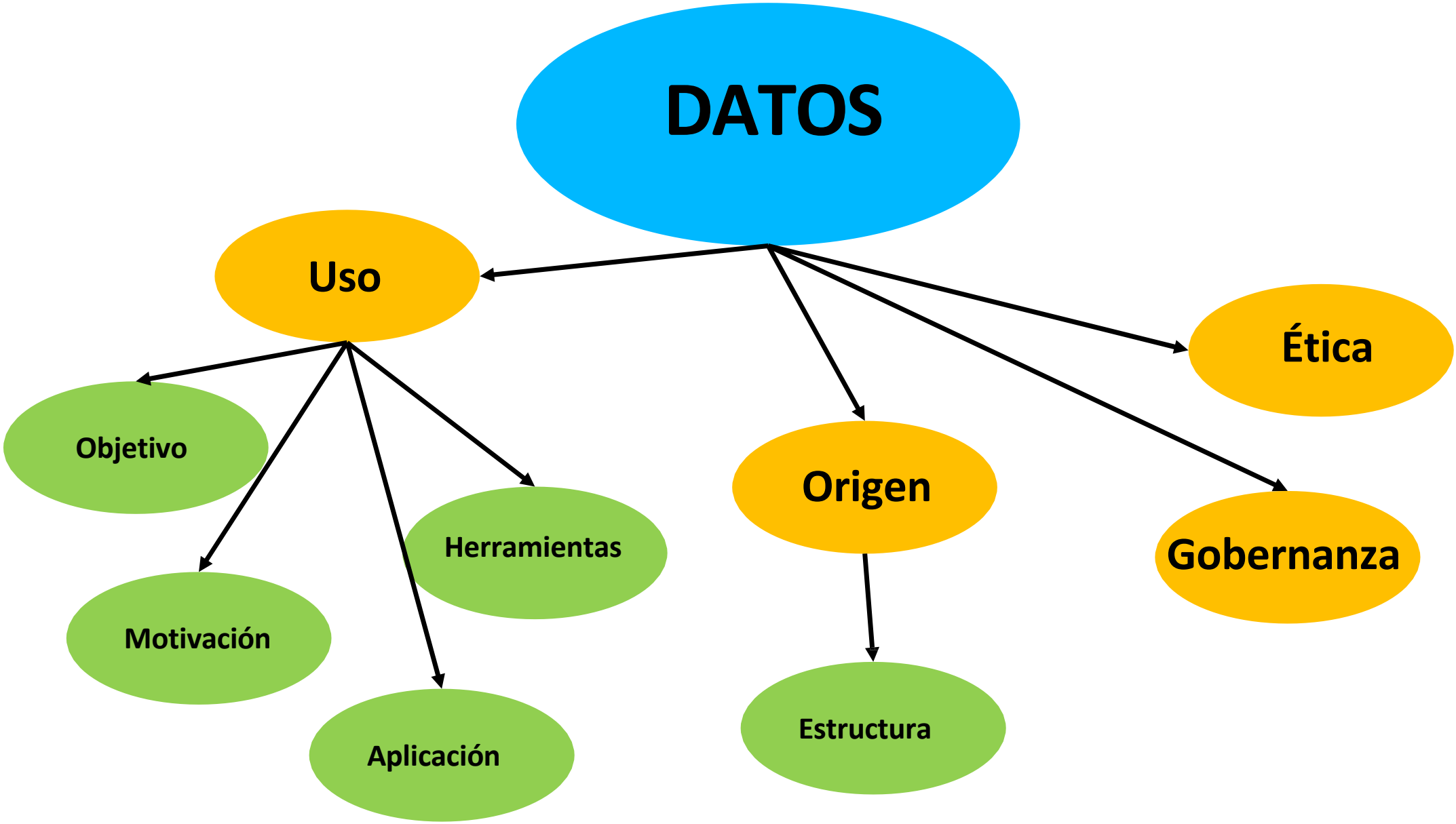
Data Scientist

UNIDAD DIDACTICA 2

2.1 Datos web: estructura y origen

Ing. Gary David Guzmán Muñoz

Curso 2025


DATOS
Uso
Ética
Objetivo
Herramientas
Origen
Gobernanza
Motivación
Aplicación
Estructura

Datos web

La principal fuente generadora de datos es la WEB.

Cuando se habla de WEB hablamos en su sentido amplio, no solo consideramos la pagina web como tal, sino servicios web y aplicaciones.

Los datos web son el punto de partida para cualquier análisis en la era digital.

Generadores de datos web

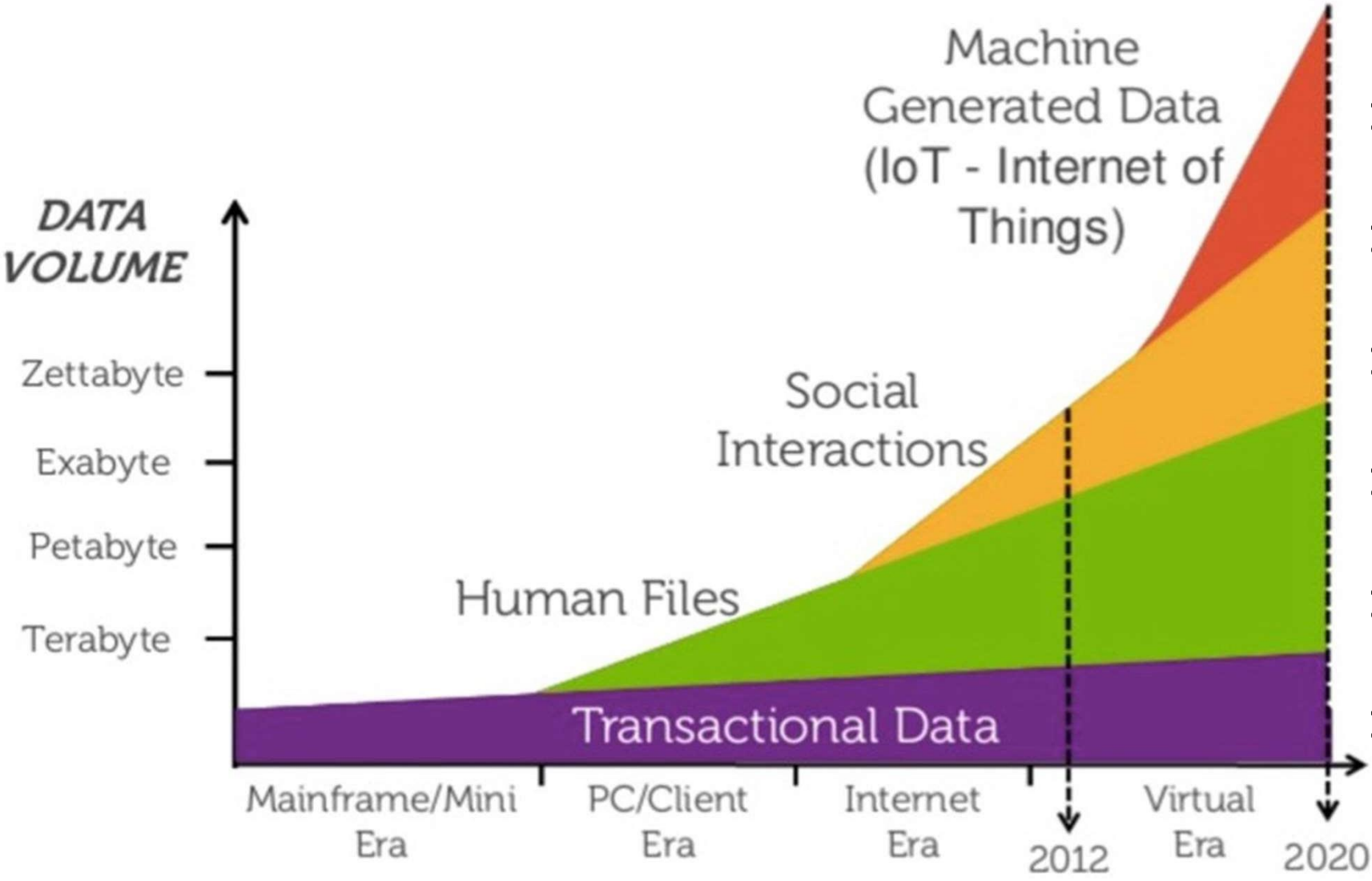
Aplicaciones

Sitios web

Buscadores

Analizadores de datos web

Volumen y Origen de los Datos



1 GB = 1024 MB

1 TB = 1024 GB

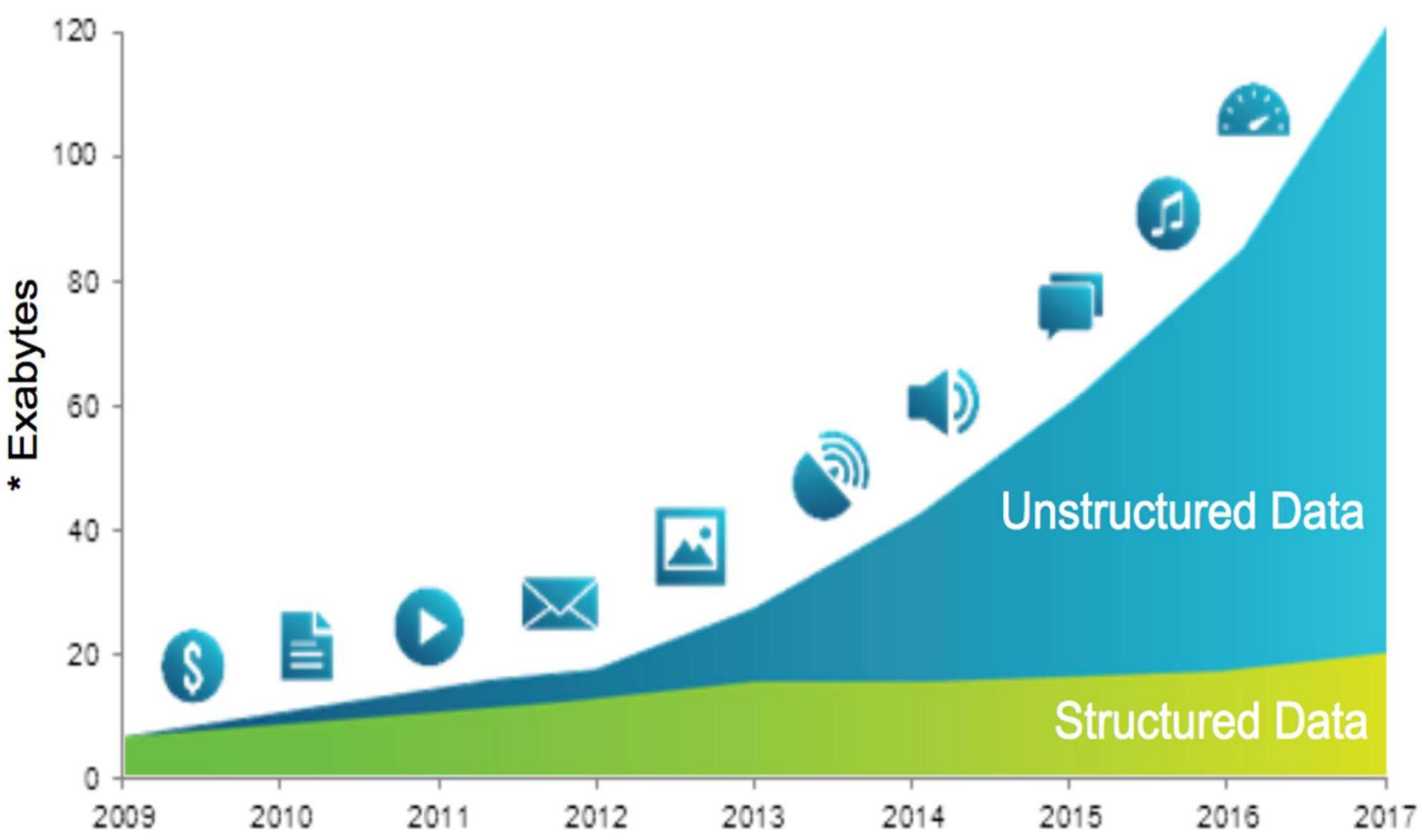
1 PB = 1024 TB

1 EB = 1024 PB

1 ZB = 1024 EB

1 YB = 1024 ZB

Volumen y Tipos de Datos



*1 exabyte = 1,000 petabytes =1 million terabytes = 1 billion gigabytes

- Unstructured data growth of
- 60–80%
  per year
- creates Web-scale storage needs

Datos ESTRUCTURADO

Los datos **tienen perfectamente definido la longitud, el formato y el tamaño de sus datos**. Se almacenan en formato tabla, hojas de cálculo o en bases de datos relacionales.

	nombre	color	edad	altura	peso	puntuacion
1:	Paco	Rojo	24	182	74.8	83
2:	Juan	Green	30	170	70.1	500
3:	Andres	Amarillo	41	169	60.0	20
4:	Natalia	Green	22	183	75.0	865
5:	Vanesa	Verde	31	178	83.9	221
6:	Miriam	Rojo	35	172	76.2	413
7:	Juan	Amarillo	22	164	68.0	902

Referencia	Fecha Alta	Tipo	Operación	Provincia	Superfici	Precio Vent	Fecha Venta	Vendedor
1	01/01/17	Parking	Alquiler	Lleida	291 m2	2.133.903,00 €	19/06/17	Carmen
2	01/01/17	Local	Venta	Girona	199 m2	1.945.424,00 €	19/04/17	Pedro
3	01/01/17	Oficina	Alquiler	Girona	82 m2	712.416,00 €	08/11/17	Joaquín
4	02/01/17	Parking	Alquiler	Girona	285 m2	1.815.450,00 €	27/04/17	Jesús
5	02/01/17	Suelo	Venta	Tarragona	152 m2	1.138.024,00 €	10/07/17	María
6	03/01/17	Industrial	Alquiler	Girona	131 m2	953.156,00 €	05/09/17	Pedro
7	03/01/17	Parking	Alquiler	Tarragona	69 m2	406.686,00 €	07/06/17	Pedro
8	03/01/17	Oficina	Venta	Girona	235 m2	2.158.475,00 €	31/10/17	Jesús
9	04/01/17	Piso	Alquiler	Lleida	108 m2	1.024.380,00 €	28/12/17	Jesús
10	04/01/17	Parking	Venta	Lleida	299 m2	2.042.768,00 €	06/10/17	Joaquín

	EMPNO	ENAME	JOB	MGR	HIREDATE	SAL	COMM	DEPTNO
1	7369	SMITH	CLERK	7902	17/12/80	800	(null)	20
2	7499	ALLEN	SALESMAN	7698	20/02/81	1600	300	30
3	7521	WARD	SALESMAN	7698	22/02/81	1250	500	30
4	7566	JONES	MANAGER	7839	02/04/81	2975	(null)	20
5	7654	MARTIN	SALESMAN	7698	28/09/81	1250	1400	30
6	7698	BLAKE	MANAGER	7839	01/05/81	2850	(null)	30
7	7782	CLARK	MANAGER	7839	09/06/81	2450	(null)	10
8	7788	SCOTT	ANALYST	7566	19/04/87	3000	(null)	20
9	7839	KING	PRESIDENT	(null)	17/11/81	5000	(null)	10
10	7844	TURNER	SALESMAN	7698	08/09/81	1500	0	30
11	7876	ADAMS	CLERK	7788	23/05/87	1100	(null)	20
12	7900	JAMES	CLERK	7698	03/12/81	950	(null)	30
13	7902	FORD	ANALYST	7566	03/12/81	3000	(null)	20
14	7934	MILLER	CLERK	7782	23/01/82	1300	(null)	10

Datos SEMIESTRUCTURADO

Los datos **no presentan una estructura perfectamente definida** como los datos estructurados, pero si presentan una **organización definida** en sus metadatos donde describen los objetos y sus relaciones, y que en algunos casos están aceptados por convención, como por ejemplo los formatos **HTML**, **XML** o **JSON**.

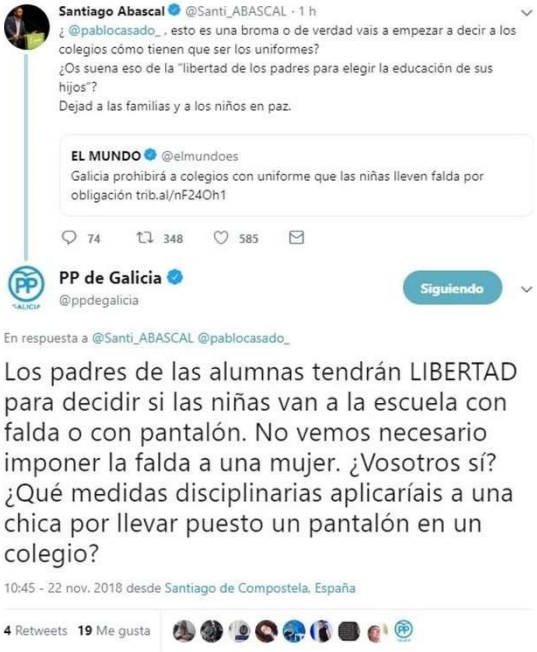
```
<HTML>
  <HEAD><TITLE>Listado de Personas</TITLE>
  <BODY>
  <%DIM doc, raiz, i
      set doc =
  Server.CreateObject("microsoft.xmldom")
      doc.load "c:\personas.xml"
      set raiz = doc.documentElement
      for i = 0  to raiz.childNodes.length -1 %>
  <%=raiz.childNodes.item(i).text%><BR>
  <%next%>
  </BODY>
</HTML>
```

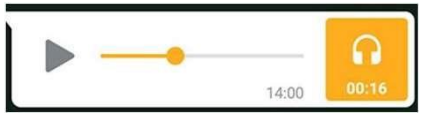
```
{
    "marcadores": [
      {
        "latitude": 40.416875,
        "longitude": -3.703308,
        "city": "Madrid",
        "description": "Puerta del Sol"
      },
      {
        "latitude": 40.417438,
        "longitude": -3.693363,
        "city": "Madrid",
        "description": "Paseo del Prado"
      },
      {
        "latitude": 40.407015,
        "longitude": -3.691163,
        "city": "Madrid",
        "description": "Estación de
          Atocha"
      }
    ]
}
```

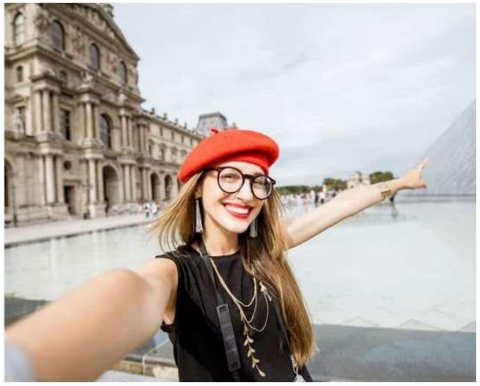
```
<?xml version="1.0" encoding="UTF-8"?>

    <sensor>

        <id> Identificador del sensor </id>
        <nombre> Nombre del Sensor </nombre>
        <descripcion> Sensor de temperatura </descripcion>
        <valor> Valor actual </valor>
        <idPedido> Pedido 1 </idPedido>

    </sensor>
```

Datos DESESTRUCTURADO

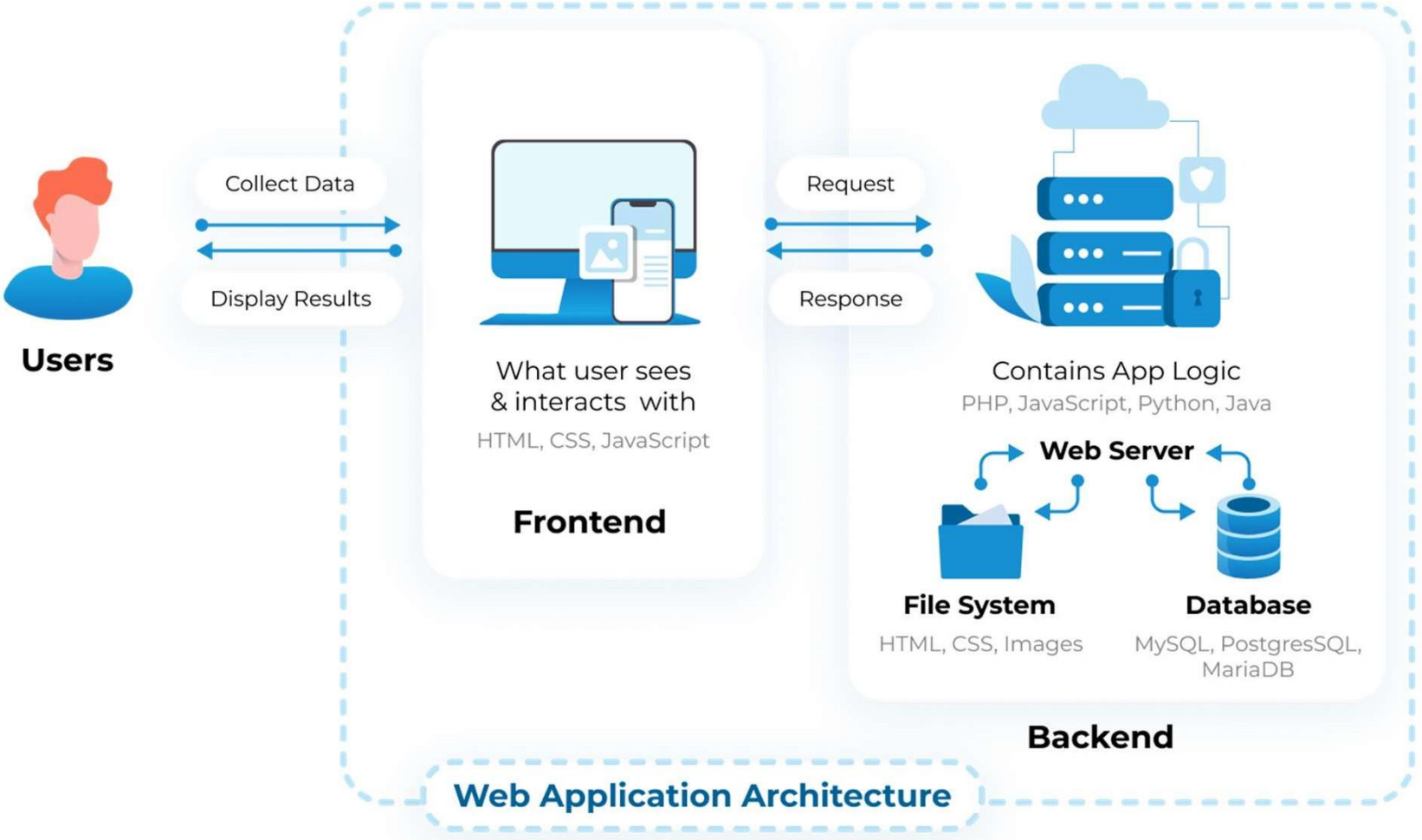
Los datos no estructurados se caracterizan por **no tener un formato específico**. Se almacenan en múltiples formatos como documentos PDF, JPG, MPG4, emails, ….







Estructura o Arquitectura básica que genera dato



Recursos y Bibliografía

- **Iofullstack repository.**

Disponible en línea:

https://github.com/iofullstack/data-scientist/tree/main/Web-Mining

- **REF 1 - what_is_datwhata_science**