

Estructura estándar de Archivos XML

BIG DATA ACADEMY



Cabecera de archivos XML

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

- El campo "version" indica la versión del estándar XML usado, sólo exista al día de hoy una versión (1.0)
- El campo "encoding" indica el juego de caracteres usados en el archivo (UTF-8)
- El campo "standalone" indica si el archivo se auto-describe (yes)



Registros en un archivo XML

```
<root>
    <element>
        INFORMACIÓN DE REGISTRO 1
    </element>
    <element>
        INFORMACIÓN DE REGISTRO 1
    </element>
    <element>
        INFORMACIÓN DE REGISTRO N
    </element>
  root>
```

Dentro de la etiqueta "root" se encuentran los registros, y cada registro es definido dentro de la etiqueta "element"



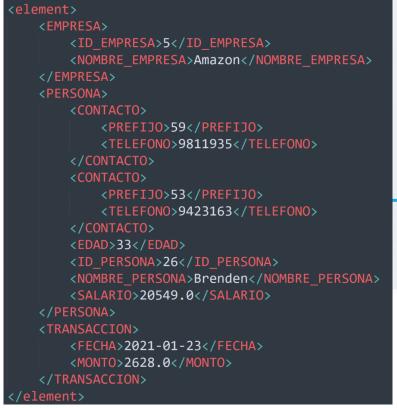
Registro semi-estructurado XML

```
<element>
   <EMPRESA>
       <ID EMPRESA>5</ID EMPRESA>
       <NOMBRE EMPRESA>Amazon/NOMBRE EMPRESA>
   </EMPRESA>
   <PERSONA>
       <CONTACTO>
            <PREFIJO>59</PREFIJO>
            <TELEFONO>9811935</TELEFONO>
       </CONTACTO>
       <CONTACTO>
            <PREFIJO>53</PREFIJO>
            <TELEFONO>9423163</TELEFONO>
       </CONTACTO>
       <EDAD>33</EDAD>
       <ID PERSONA>26</ID PERSONA>
       <NOMBRE PERSONA>Brenden/NOMBRE PERSONA>
       <SALARIO>20549.0</SALARIO>
   </PERSONA>
   <TRANSACCION>
       <FECHA>2021-01-23</FECHA>
       <MONTO>2628.0</MONTO>
   </TRANSACCION>
:/element>
```

Un registro semiestructurado define su propio esquema de metadatos y puede tener subcampos



Modelamiento de Dataframes Semi-Estructurados

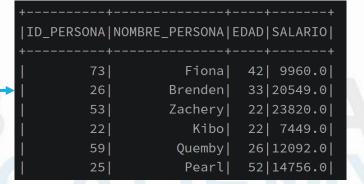


dfSemiEstructurado



+-----+
|ID_PERSONA|ID_EMPRESA| MONTO| FECHA|
+-----+
26	5	2628.0	2021-01-23
21	9	4261.0	2021-01-23
73	7	1429.0	2021-01-23
14	5	3385.0	2021-01-23
80	4	3514.0	2021-01-23

dfTransaccion



dfPersona



dfEmpresa