
Procesamiento Semi-Estructurado JSON

BIG DATA ACADEMY

Registros semi-estructurados

```
{
  "PERSONA":{
    "ID_PERSONA":"26",
    "NOMBRE_PERSONA":"Brenden",
    "EDAD":33,
    "SALARIO":20549.0,
    "CONTACTO":[
      {"PREFIJO":"59","TELEFONO":"9811935"},
      {"PREFIJO":"53","TELEFONO":"9423163"}
    ]
  },
  "EMPRESA":{
    "ID_EMPRESA":"5",
    "NOMBRE_EMPRESA":"Amazon"
  },
  "TRANSACCION":{
    "MONTO":2628.0,
    "FECHA":"2021-01-23"
  }
}
```

Un registro semi-estructurado define su propio esquema de metadatos y puede tener subcampos

Dataframes semi-estructurados



Archivo JSON

En un dataframe semi-estructurado **no necesariamente todos los registros tienen el mismo esquema de metadatos**

Modelamiento de Dataframes Semi-Estructurados

```
root
|-- EMPRESA: struct (nullable = true)
|   |-- ID_EMPRESA: string (nullable = true)
|   |-- NOMBRE_EMPRESA: string (nullable = true)
|-- PERSONA: struct (nullable = true)
|   |-- CONTACTO: array (nullable = true)
|   |   |-- element: struct (containsNull = true)
|   |   |   |-- PREFIJO: string (nullable = true)
|   |   |   |-- TELEFONO: string (nullable = true)
|   |-- EDAD: long (nullable = true)
|   |-- ID_PERSONA: string (nullable = true)
|   |-- NOMBRE_PERSONA: string (nullable = true)
|   |-- SALARIO: double (nullable = true)
|-- TRANSACCION: struct (nullable = true)
|   |-- FECHA: string (nullable = true)
|   |-- MONTO: double (nullable = true)
```

dfSemiEstructurado



Modelamiento

[Crear varios
dataframes
estructurados]

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

ID_PERSONA	NOMBRE_PERSONA	EDAD	SALARIO
73	Fiona	42	9960.0
26	Brenden	33	20549.0
53	Zachery	22	23820.0
22	Kibo	22	7449.0
59	Quemby	26	12092.0
25	Pearl	52	14756.0

dfPersona

ID_EMPRESA	NOMBRE_EMPRESA
9	IBM
5	Amazon
10	Sony
2	Microsoft
7	Samsung

dfEmpresa