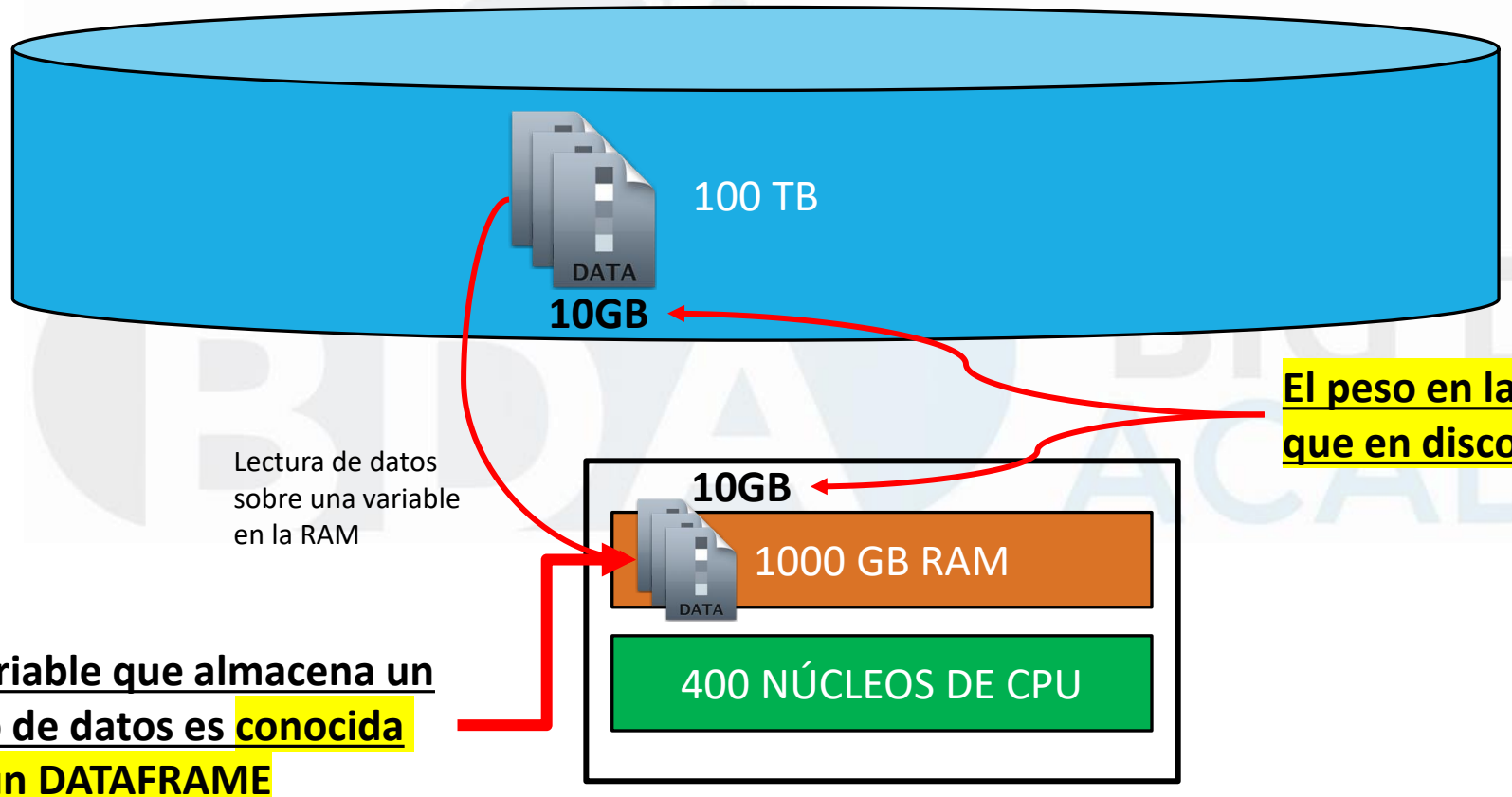
Lectura de Datos con Spark

BIG DATA ACADEMY

Alonso Melgarejo [alonsoraulmgs@gmail.com]

Lectura y procesamiento de datos



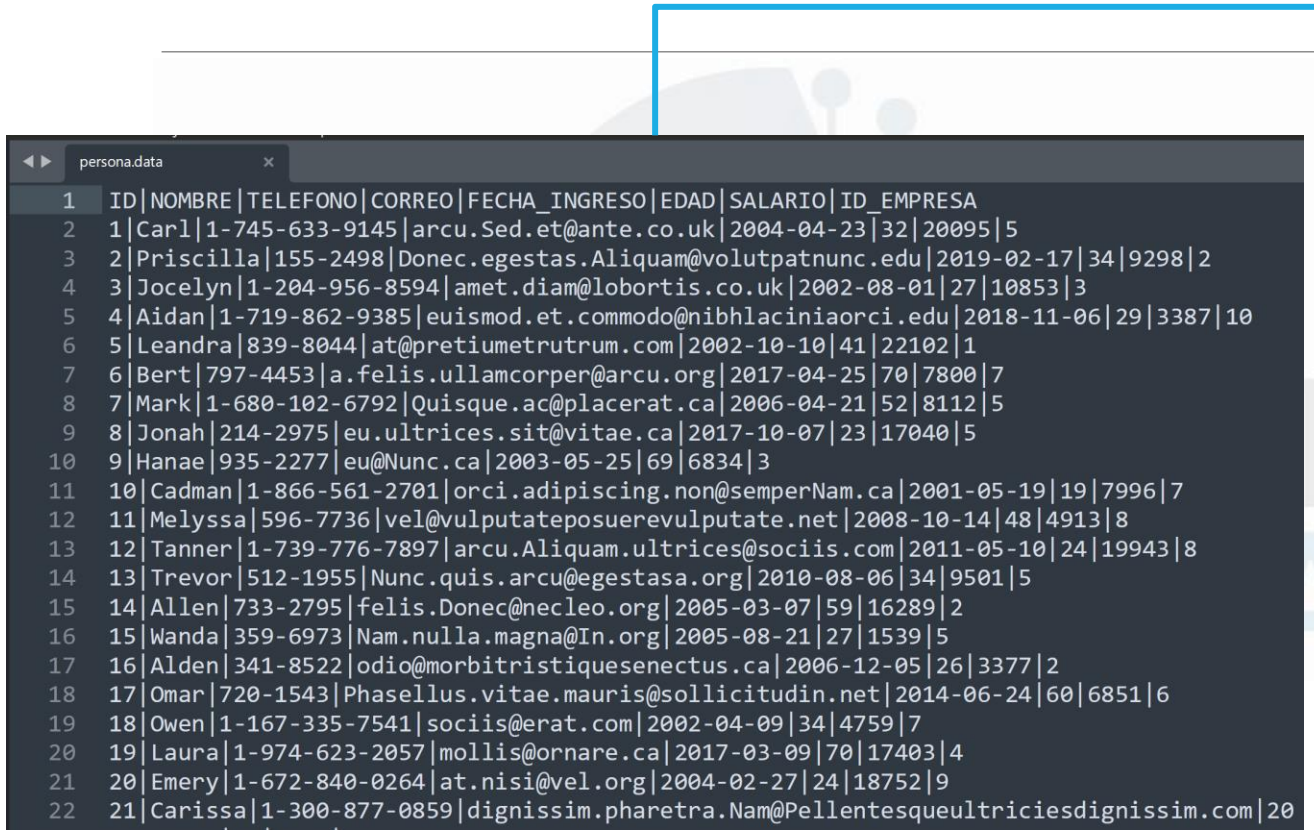
INFRAESTRUCTURA DE ALMACENAMIENTO

El peso en la RAM es el mismo que en disco duro

INFRAESTRUCTURA DE PROCESAMIENTO

Una variable que almacena un archivo de datos es conocida como un DATAFRAME

Dataframes estructurados desde archivos



Archivo con separador

dfEstructurado

```
root
 |-- ID: string (nullable = true)
 |-- NOMBRE: string (nullable = true)
 |-- TELEFONO: string (nullable = true)
 |-- CORREO: string (nullable = true)
 |-- FECHA_INGRESO: string (nullable = true)
 |-- EDAD: string (nullable = true)
 |-- SALARIO: string (nullable = true)
 |-- ID_EMPRESA: string (nullable = true)
```

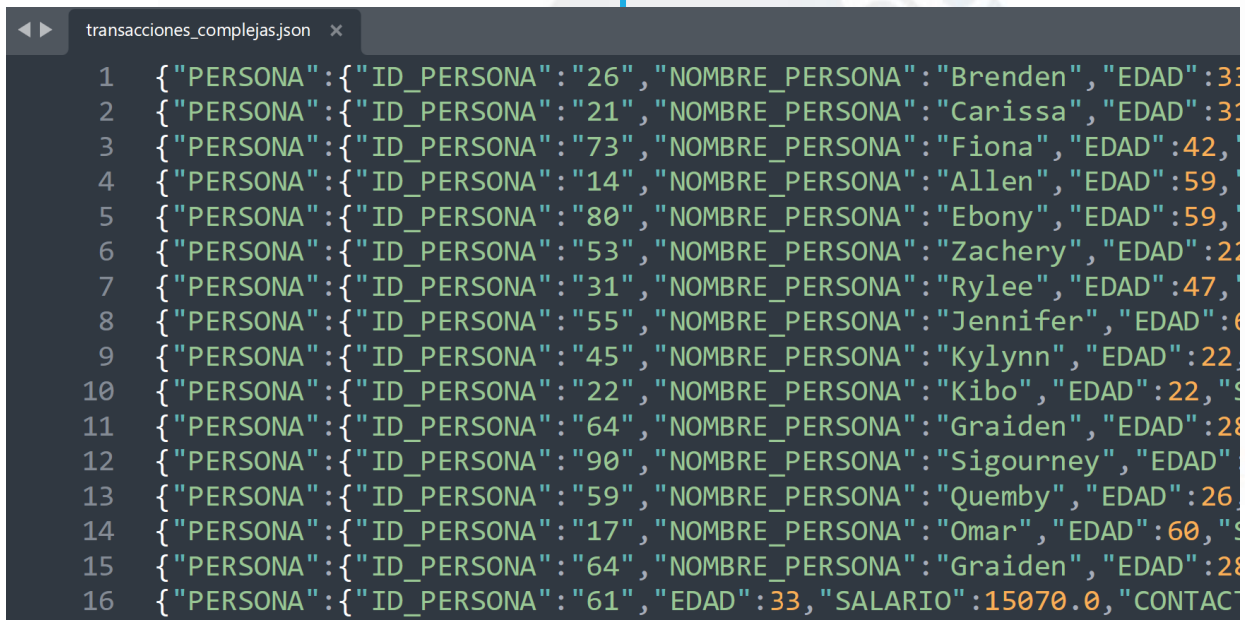
Todos los registros tienen el **mismo esquema de metadatos**

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 sub-campos**

Dataframes semi-estructurados



Archivo JSON

dfSemiEstructurado

```
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)
```

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