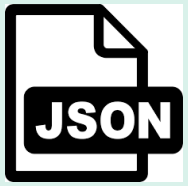
Laboratorio 07

PYSPARK SQL

Alonso Melgarejo [alonsoraulmgs@gmail.com]

Formatos de archivos

DATOS ESTRUCTURADOS	DATOS SEMI-ESTRUCTURADOS	DATOS NO ESTRUCTURADOS
 Tablas	 JSON	 IMÁGENES
 Archivos con separadores	 XML	 VIDEOS
		 AUDIOS
		 DOCUMENTOS ENRIQUECIDOS

Dependiendo del formato de archivo, existen diferentes herramientas de procesamiento

Datos estructurados

Tablas

Archivos con separadores

En los datos estructurados todos los registros tienen el mismo esquema de metadatos

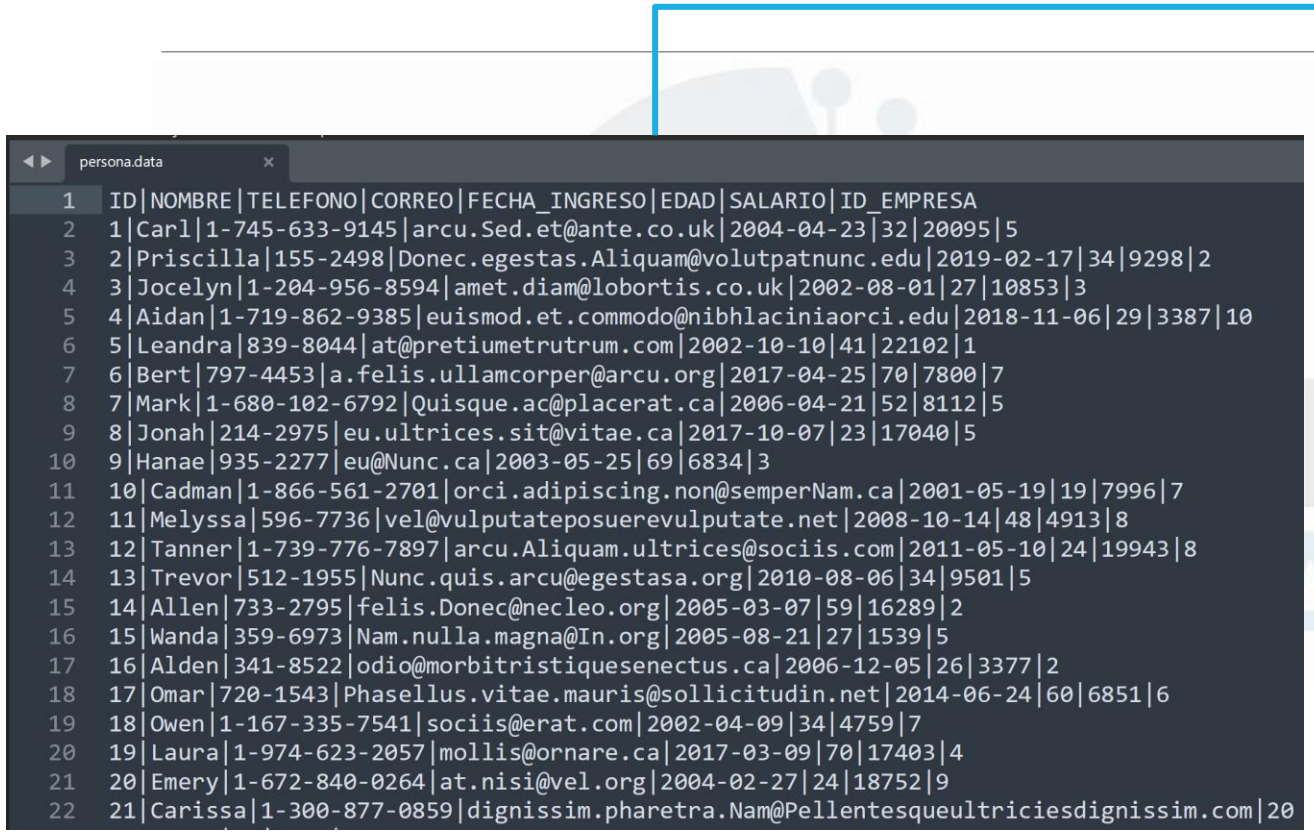
Ejemplo

persona.data

```
ID|NOMBRE|TELEFONO|CORREO|FECHA_INGRESO|EDAD|SALARIO|ID_EMPRESA
1|Carl|1-745-633-9145|arcu.Sed.et@ante.co.uk|2004-04-23|32|20095|5
2|Priscilla|155-2498|Donec.egestas.Aliquam@volutpatnunc.edu|2019-02-17|34|9298|2
3|Jocelyn|1-204-956-8594|amet.diam@lobortis.co.uk|2002-08-01|27|10853|3
4|Aidan|1-719-862-9385|euismod.et.commodo@nibhlaciniaorci.edu|2018-11-06|29|3387|10
5|Leandra|839-8044|at@pretiumetrutrum.com|2002-10-10|41|22102|1
6|Bert|797-4453|a.felis.ullamcorper@arcu.org|2017-04-25|70|7800|7
7|Mark|1-680-102-6792|Quisque.ac@placerat.ca|2006-04-21|52|8112|5
8|Jonah|214-2975|eu.ultrices.sit@vitae.ca|2017-10-07|23|17040|5
9|Hanae|935-2277|eu@Nunc.ca|2003-05-25|69|6834|3
10|Cadman|1-866-561-2701|orci.adipiscing.non@semperNam.ca|2001-05-19|19|7996|7
11|Melyssa|596-7736|vel@vulputateposuerevulputate.net|2008-10-14|48|4913|8
12|Tanner|1-739-776-7897|arcu.Aliquam.ultrices@sociis.com|2011-05-10|24|19943|8
13|Trevor|512-1955|Nunc.quis.arcu@egestasa.org|2010-08-06|34|9501|5
14|Allen|733-2795|felis.Donec@necleo.org|2005-03-07|59|16289|2
15|Wanda|359-6973|Nam.nulla.magna@In.org|2005-08-21|27|1539|5
16|Alden|341-8522|odio@morbitristiquesenectus.ca|2006-12-05|26|3377|2
17|Omar|720-1543|Phasellus.vitae.mauris@sollicitudin.net|2014-06-24|60|6851|6
18|Owen|1-167-335-7541|sociis@erat.com|2002-04-09|34|4759|7
19|Laura|1-974-623-2057|mollis@ornare.ca|2017-03-09|70|17403|4
20|Emery|1-672-840-0264|at.nisi@vel.org|2004-02-27|24|18752|9
```

Si tomamos cualquier registro del archivo de datos, **todos tienen el mismo esquema de metadatos** (juego de campos)

Dataframes estructurados desde archivos



Archivo con separador

dfEstructurado

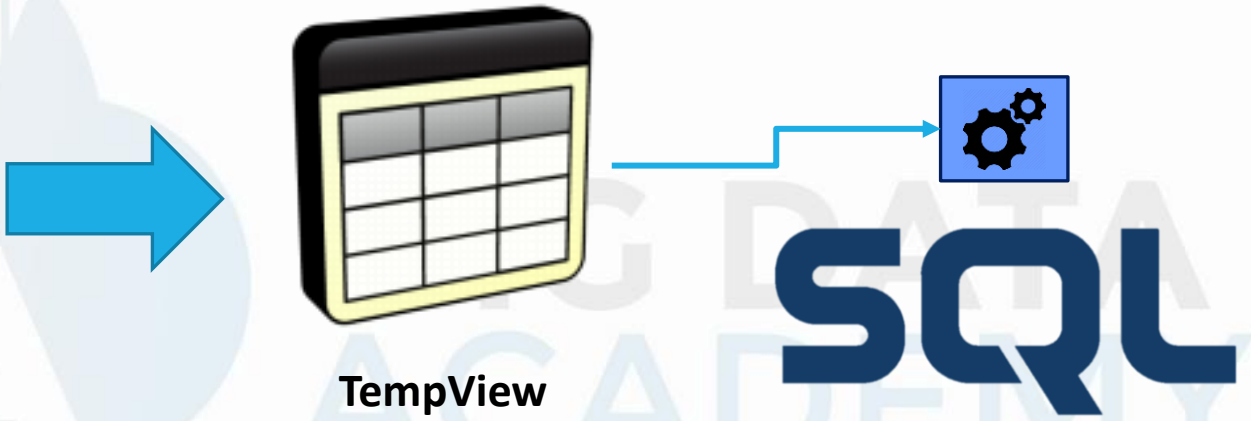
```
root
 |-- ID: string (nullable = true)
 |-- NOMBRE: string (nullable = true)
 |-- TELEFONO: string (nullable = true)
 |-- CORREO: string (nullable = true)
 |-- FECHA_INGRESO: string (nullable = true)
 |-- EDAD: string (nullable = true)
 |-- SALARIO: string (nullable = true)
 |-- ID_EMPRESA: string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

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

Dataframe como TempView

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)
```



Si conocemos de lenguajes de programación (Scala, Python, otros) podemos procesar dataframes estructurados usando la sintaxis SQL de procesamiento de tablas. Deberemos convertir al dataframe en un TempView.