Práctica 5: Introducción a los formatos Avro y Parquet

Para realizar esta práctica son necesarios los archivos paises.avro y demo.parquet

Uso de archivos Avro

Avro se basa en esquemas. Cuando los datos .avro son leídos siempre está presente el esquema con el que han sido escritos. Esto
permite aumentar el rendimiento al escribir los datos, haciendo la serialización rápida y viable en espacio.
Para poder analizar, visualizar sus datos y transformar los ficheros avro vamos a utilizar las herramientas avro-tools

Esquema de un fichero avro

Conseguimos el formato de los datos de un fichero avro a través de su esquema. Se ejecuta la siguiente instrucción:

```
[alumno@pasarela datos_pr1]$ avro-tools getschema paises.avro > paises.avsc
[main] WARN org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your plat
form... using builtin-java classes where applicable
log4j:WARN No appenders could be found for logger (org.apache.htrace.core.Tracer).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
[alumno@pasarela datos_pr1]$
```

cat paises.avsc

```
[alumno@pasarela datos_pr1]$ cat paises.avsc
{
  "type" : "record",
  "name" : "Root",
  "fields" : [ {
    "name" : "country",
    "type" : [ "null", "string" ]
  }, {
    "name" : "year",
    "type" : [ "null", "long" ]
  }, {
    "name" : "population",
    "type" : [ "null", "double" ]
  }, {
    "name" : "continent",
    "type" : [ "null", "string" ]
  }, {
    "name" : "lifeExp",
    "type" : [ "null", "double" ]
  }, {
    "name" : "gdpPercap",
    "type" : [ "null", "double" ]
  } ]
}
```

Extracción de datos desde un fichero avro

Con las avro-tools se pueden obtener los datos de un fichero avro en distintos formatos. En este caso, el esquema y datos del
fichero vienen en formato json.

Exportar datos a formato JSON
Utilizamos la opción tojson (redireccionando la salida a un fichero .json)

```
[alumno@pasarela datos_pr1]$ avro-tools tojson paises.avro > paises.json
[main] WARN org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your plat
form... using builtin-java classes where applicable
log4j:WARN No appenders could be found for logger (org.apache.htrace.core.Tracer).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
[alumno@pasarela datos_pr1]$ head paises.json | jq .
{
  "country": {
    "string": "Afghanistan"
  },
  "year": {
    "long": 1952
  },
  "population": {
    "double": 8425333
  },
  "continent": {
    "string": "Asia"
  },
  "lifeExp": {
    "double": "NaN"
  },
  "gdpPercap": {
    "double": 779.4453145
  }
}
```

Uso de ficheros Parquet

Es un formato de almacenamiento columnar disponible para cualquier proyecto en el ecosistema de Hadoop, independiente del
framework utilizado para procesar los datos, o el lenguaje de programación.

En este caso utilizamos la aplicación parquet-tools, y el fichero de muestra es demo.parquet

Visualización de los datos del archivo parquet

Se pueden obtener las líneas del archivo en texto plano, usando la opción cat, y a partir de la salida procesar el fichero obtenido.

```
[alumno@pasarela datos_pr1]$ parquet-tools cat demo.parquet > demo.txt
[alumno@pasarela datos_pr1]$ head -n 10 demo.txt
id = 1
name = Toy Story
year = 1995

id = 2
name = Jumanji
year = 1995

id = 3
name = Grumpier Old Men
```

Obtención del esquema de los datos

parquet-tools permite obtener el esquema de la estructura de los datos de forma simple, o bien de forma detallada usando la opción -d

```
[alumno@pasarela datos_pr1]$ parquet-tools schema demo.parquet
message movie {
  optional int32 id;
  optional binary name (UTF8);
  optional int32 year;
}
```

Obtención de los datos de una columna

Con la opción dump se pueden extraer los datos y metadatos de una o varias columnas.

```
[alumno@pasarela datos_pr1]$ FRANCISCO GARRIDO LARA^C
[alumno@pasarela datos_pr1]$ parquet-tools dump -c name demo.parquet > demo.dump.out
[alumno@pasarela datos_pr1]$ head -10 demo.dump.out
row group 0
--------------------------------------------------------------------------------------------------
name:  BINARY SNAPPY DO:0 FPO:15589 SZ:53871/76846/1,43 VC:3881 ENC:RLE,BIT_PACKED,PLAIN ST:[min: $1,000
,000 Duck, max: eXistenZ, num_nulls: 0]

    name TV=3881 RL=0 DL=1
    ----------------------------------------------------------------------------------------------
    page 0:  DLE:RLE RLE:BIT_PACKED VLE:PLAIN ST:[min: $1,000,000 Duck, max: eXistenZ, num_nulls: 0] SZ:
76793 VC:3881

BINARY name
--------------------------------------------------------------------------------------------------
[alumno@pasarela datos_pr1]$
```