Laboratorio No 1

Profesor: Juan David Velasquez

Fecha: Mayo 4 / 2019

**LAB: Transformación de archivos con sed**

Una tarea común en Analytics es el procesamiento de archivos en bruto para que puedan ser usados en otros procesos, donde gran parte de dicho procesamiento corresponde a la transforamción del formato. Estos archivos suelen tener millones de registros por lo que la aplicación de procedimientos manuales no es práctica

El archivo data.csv contiene una muestra de los archivos generados automáticamente por un software. El archivo contiene cuatro (4) columnas, donde los valores válidos para los campos son los siguientes:

* Columna 1: DD/MM/AA.
* Columna 2: A, B, C. \N indica NULL.
* Columna 3: Número entero. \N indica NULL.
* Columna 4: Número decimal. \N indica NULL.

Usted debe escribir un script en bash (el archivo se llama script.sh), que realice las siguientes acciones:

* Convierta el formato de las fechas de DD/MM/YY a YYYY-MM-DD.
  + Si el día es menor que 9 tendremos que adicionar cero (0).
  + Si el mes es de una posición adicionarle cero (0).
  + Si el año:
    - Tiene una menos de 4 posiciones habra que adicionarle 3 posiciones con cero (000+numero).
    - Tiene dos posiciones puede ser del siglo pasado y habra que adicionarle el número 19 eoc adicionarle el número 20.
  + Elimine cualquier registro que no contenga [\*/\*/\*]
  + Elimine cualquier registro que tenga dia cero/mes 0.
  + Transponer la fecha
* Transforme el archivo para que todos los campos nulos aparezcan como \N.
  + Si existe una letra diferente de ABC reemplacela por \N.
  + Diferencie entre mayúscula y minúscula.
  + Si aparecen 2 comas seguidas debe ser anulado el registro.
* Reemplace los (punto y coma ;) por (coma ,)
  + Tener cuidado porque si se reemplaza sin validar podemos dañar el archivo porque la siguiente instrucción consiste en usar el punto para indicar decimales.
  + Se debe cambiar primero la coma (,) por el punto (.) y después reemplazar el punto y coma (;) por coma (,).
  + Elimine la cadena siguiente al punto (incluyéndo el punto pues se trata de un campo entero.

* Use el (punto .) para indicar decimales.
  + Buscar la coma (,) en el número y reemplazarla por el punto (.).
* Extraiga los registros que no tienen campos nulos.
  + Esta es la primera instrucción del programa porque si se encuentra un null debemos borrar el registro.
* Otras condiciones que se deben cumplir
  + Elimine cualquier registro que contenga \N
  + Elimine cualquier registro con campos en blanco [,,]
  + Reemplace punto y coma (;) por coma (,)

**Estrategia Resumida**

|  |  |
| --- | --- |
| Estrategia | |
| Registro Inválido | Registro Válido |
| Elimine cualquier registro que contenga \N |  |
| Elimine cualquier registro con campos en blanco [,,] |  |
| Campo 2:  Elimine cualquier registro que no contenga [A,B,C] |  |
| Campo 1 :  Elimine cualquier registro que no contenga [\*/\*/\*] |  |
| Campo 1: Elimine cualquier registro que tenga dia cero/mes 0 |  |
|  | Reemplace la coma (,) por el punto (.) |
|  | Elimine la cadena siguiente al punto para el campo 3 |
|  | Reemplace punto y coma (;) por coma (,) |
|  | Completar el dd |
|  | Completar el mm |
|  | Completar el aa |
|  | Transponer la fecha |
|  | Escribir el archivo resultante |
|  | Borrar el temporal |

**Documentos consultados**

<https://thales.cica.es/rd/glinex/practicas-glinex05/manuales/bash/practica.pdf>