**Ejercicios PIG**

* Copiar en local file sistema de la MV el fichero datos\_pig.txt en la ruta /home/cloudera/ejercicios/pig y abrir el fichero para revisar su contenido.
* Arranca el Shell de Pig en modo local
* “pig -x local”
* Carga los datos en pig en una variable llamada “data”. Los nombres de las columnas deben ser (key, campana, fecha, tiempo, display, accion, cpc, pais, lugar). Los tipos de las columnas deben ser chararray excepto acction y cpc que son int.

data = LOAD 'datos\_pig.txt' USING PigStorage (',') as (

key: chararray, campana: chararray, fecha: chararray,

tiempo: chararray, display: chararray, accion: int,

cpc: int, pais: chararray, lugar: chararray);

* Usa el comando DESCRIBE para ver el esquema de la variable “data”

describe data;

data: {key: chararray,campana: chararray,fecha: chararray,tiempo: chararray,display: chararray,accion: int,cpc: int,pais: chararray,lugar: chararray}

* Selecciona las filas de “data” que provengan de USA;

filasUSA = FILTER data by pais == 'USA';

* Listar los datos que contengan en su key el sufijo surf:

sufijosurf = FILTER data by key MATCHES '.\*surf';

DUMP sufijosurf;

* Crear una variable llamada “ordenado” que contenga las columnas de data en el siguiente orden: (campaña, fecha, tiempo, key, display, lugar, accion, cpc)

ordenado = FOREACH data GENERATE campana, fecha, tiempo, key, display, lugar, accion, cpc;

* Guarda el contenido de la variable “ordenado” en una carpeta en el local file system de tu MV llamada resultado en la ruta /home/cloudera/ejercicios/pig

STORE ordenado INTO '/home/cloudera/ejercicios/resultado';

* Comprobar el contenido de la carpeta.

Parte 2

Usa el archivo “estudiantes”, comprueba el formato que tienen los datos

dentro de este archivo.

1. Realiza la carga en la variable estudiantes teniendo en cuenta el

formato de los datos del fichero.

data\_estudiantes = LOAD 'estudiantes' USING PigStorage (',') as (

key: chararray, nombre: chararray, apellido: chararray,

telefono: chararray, ciudad: chararray, carrera: chararray, nota: int);

2. Lista los datos con el nombre y apellido en una misma columna y

todas las columnas que consideres sin espacios en blanco.

nombreapellidos = FOREACH data\_estudiantes GENERATE CONCAT(nombre, ' ', apellido) as nombrecompleto, telefono, ciudad, REPLACE(carrera, ' ', '') as carrera, nota;

3. Guarda el resultado en un nuevo archivo con ‘:’ como delimitador.

STORE nombreapellidos INTO '[file:///home/cloudera/ejercicios/resultado\_estudiantes](file://\home\cloudera\ejercicios\resultado_estudiantes)' using PigStorage(':');

4. Lista los datos que sean de Madrid.

estudiantesmadrid = FILTER nombreapellidos by ciudad == 'Madrid';

5. Muestra la media agrupada por grado.

grupogrado= GROUP nombreapellidos BY carrera;

6. Muestra el número de estudiantes de cada ciudad.

grupociudad = GROUP nombreapellidos BY ciudad;

estudiantesporciudad = foreach grupociudad Generate(

nombreapellidos.nombrecompleto,

nombreapellidos.telefono,

nombreapellidos.ciudad,

nombreapellidos.carrera,

nombreapellidos.nota),

COUNT(nombreapellidos.ciudad) as alumnosxciudad;

solociudadesyalumnos = FOREACH grupociudad GENERATE(nombreapellidos.ciudad), COUNT(nombreapellidos.ciudad) as estudiantesporciudad;

7. El nombre del alumno o alumnos de cada ciudad con mayor nota

media.

mayornotaciudad = foreach grupociudad Generate(nombreapellidos.nombrecompleto, nombreapellidos

.telefono, nombreapellidos

.ciudad, nombreapellidos

.carrera, nombreapellidos

.nota),

MAX(nombreapellidos.nota) as notamaxima;