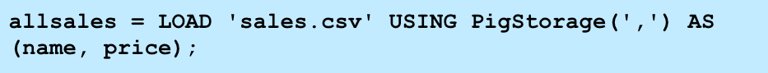
**Apuntes**

Pig Latin es un lenguaje de flujo de datos. El flujo de datos es representado por una secuencia de instrucciones.



Se puede utilizar PigStorage para poner un delimitador en la tabla. El delimitador por defecto es un espacio.

El comando usado para obtener la salida nos indicará el formato de la misma

* DUMP identificador: Muestra la salida por pantalla
* STORE identificador into “ruta”: Envía los resultados al disco (HDFS)

**Ejercicios**

1. 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.
   1. Cp /media/sf\_sharedFolder/datos\_pig.txt /home/cloudera/ejercicios/pig
2. Arranca el Shell de Pig en modo local
   1. “pig -x local”
3. 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.
   1. data= LOAD ‘/home/cloudera/ejercicios/pig/datos\_pig.txt’ USING PigStorage(‘,’) as (key: chararray, campana: chararray, fecha: chararray, tiempo: chararray, display: chararray, acction:int, cpc:int, pais: chararray, lugar: chararray);
4. Usa el comando DESCRIBE para ver el esquema de la variable “data”
   1. describe data
5. Selecciona las filas de “data” que provengan de USA;
   1. filasUSA = FILTER data by pais == ‘USA’;
6. Listar los datos que contengan en su key el sufijo surf:

datosSurf= FILTER data BY key MATCHES ‘surf.\*’;

1. Crear una variable llamada “ordenado” que contenga las columnas de data en el siguiente orden: (campaña, fecha, tiempo, key, display, lugar, accion, cpc)
   1. ordenado = FOREACH data GENERATE campana, fecha, tiempo, key, display, lugar, acction, cpc;
2. 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
   1. STORE ordenado INTO ‘/home/Cloudera/ejercicios/pig/resultado’;
3. Comprobar el contenido de la carpeta.
   1. Gedit resultado (dentro de la carpeta pig)