UNIVERSIDAD DE GRANADA

MÁSTER DE CIENCIA DE DATOS E INGENIERÍA DE COMPUTADORES

CURSO ACADÉMICO 2019-2020

BIG DATA 2

ALMACENAMIENTO DE DATOS MASIVOS PARA PROCESAMIENTO Y ANÁLISIS: ETL

Procesamiento de datos con *Pig Latín*

Documentación del proceso ejecutado con Pig Latin para el procesamiento y obtención de información de un dataset masivo.

Nicolás Cubero

22 de Mayo de 2020

1. Descripción del *dataset*

En este proyecto se trabaja con el *dataset* **Adult**, disponible en el repositorio web de la *UCI Machine Learning* [1]

Este *dataset* recoge datos tomados del censo de EEUU de 1994 con diversos datos demográficos sobre la población estadounidense en edad adulta, presentando un problema de clasificación en el que se trata de predecir el nivel de ingresos de cada ciudadano/a (superior a 50.000 dólares anuales o inferior o igual a 50.000 dólares anuales) en función del resto de datos demográficos recogidos para cada ciudadano/a.

Para este proyecto se ha hecho uso del *dataset* reservado para entrenamiento de modelos predictivos de clasificación que consta de 32561 instancias y, para cada ciudadano/a se recogen un total de 15 características demográficas incluídas la etiqueta a predecir: el nivel de ingresos de cada individuo.

2. Experimento realizado

Con los datos recogidos en este *dataset*, se pretende realizar un experimento con *Pig Latín* para analizar el número de ciudadanos/as de cada raza que desempeña algún cargo público para el gobierno estadounidense según su nivel de estudios, con la finalidad de comprobar, si a igualdad de nivel de estudios, existen diferencias significativas en el número de personas de cada raza que desempeñan cargos públicos dentro del gobierno estadounidense.

Para obtener esta información, se requiere realizar una operación de filtrado mediante la cual, se seleccionen los ciudadanos/as del *dataset* que desempeñen un cargo público dentro del gobierno (todos los ciudadanos/as cuyo valor del atributo `workclass` presente alguno de los siguientes valores: `Federal-gov, Local-gov` y `State-gov`).

[1] Enlace al repositorio web de la UCI con el dataset Adult: http://archive.ics.uci.edu/ml/datasets/Adult