Experimentación ETL con Pig: Big data 2

Máster en ciencia de datos e ingeniería de computadores - UGR

Curso 2018

Mª Cristina Heredia Gómez

Índice

Base de datos seleccionada

La base de datos seleccionada es la de Census Income. La BD está compuesta por 48842 instancias y 14 atributos numéricos y categóricos. Contiene valores perdidos y la siguiente información sobre distintos ciudadanos:

- age - edad
- workclass - sector de trabajo al que pertenecen
- fnlwgt - salario anual
- education - grado de estudios
- education-num - nivel de estudios (en número)
- marital-status - estado civil
- occupation] - profesión
- relationship - relación social (hijo/a, cónyuge, etc)
- race - raza (blanca,negra,Asiática, etc)
- sex - mujer u hombre
- capital-gain - ganancia capital
- capital-loss - pérdida capital
- hours-per-week - horas de trabajo semanales
- native-country - paìs de nacimiento

Experimentación ETL con Pig

Planteamiento del experimento

Ante esta Base de datos, se diseña una experimentación con Pig que trate de responder a las siguientes preguntas:

1. ¿Cuántas mujeres mayores de 18 años hay en cada sector de trabajo?
2. Para cada sector de trabajo, ¿cual es la edad media de las mujeres por sector?

A través de diversas consultas a la BD con Pig que nos permitan obtener la información de nuestro interés.

Desarrollo del experimento

Carga de los datos

Esta experimentación se ha realizado mediante acceso al servidor hadoop de la UGR, donde se dispone de Pig por consola. Comenzamos creando el directorio hadoop de nombre **input2** donde posteriormente cargamos los datos:

```
[CD_76668203@hadoop-master ~]$ hdfs dfs -mkdir input2
```

Una vez creado el directorio, descargamos los datos de Census y los copiamos al directorio que acabamos de crear, comprobando que el fichero se ha cargado correctamente en HDFS:

```
[CD_76668203@hadoop-master ~]$ wget https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data
[CD_76668203@hadoop-master ~]$ hdfs dfs -put adult.data input2
[CD_76668203@hadoop-master ~]$ hdfs dfs -ls input2
Found 1 items
-rw-r--r--   2 CD_76668203 supergroup    3974305 2018-04-12 12:17 input2/adult.data
[CD_76668203@hadoop-master ~]$
```

La salida nos muestra que en el directorio se encuentra el item que acabamos de copiar.

El siguiente paso será especificar el esquema de los datos a cargar en Pig y cargar los datos del fichero en ella, para ello usamos la sentencia Pig siguiente:

Listing 1: Sentencia Pig para creación de la tabla

```
measure = load '/user/CD_76668203/input2/adult.data' using PigStorage(',') as
(Age: int, Workclass: chararray, Fnlwgt: long, Education: chararray,
EducationNum: int, MaritalStatus: chararray, Occupation: chararray,
Relationship: chararray, Race: chararray, Sex: chararray,
capitalGain: long, CapitalLoss: int, HoursPerWeek: int,
Country: chararray, YearGain: chararray);
```

Tras esto, volcamos los resultados en el directorio **pigResults/CensusProcessed** y comprobamos que se ha creado correctamente:

```
grunt> store measure into 'pigResults/CensusProcessed' using PigStorage (',');
[CD_76668203@hadoop-master ~]$ hdfs dfs -cat pigResults/CensusProcessed/part-m-00000 | less
cat: Unable to write to output stream.
```

El resultado nos muestra un resumen de la salida:

```
55, Local-gov,98545, 10th,6, Married-civ-spouse, Adm-clerical, Husband, White, Male,0,0,40, United-States, <=50K
53, Private,242606, HS-grad,9, Married-civ-spouse, Transport-moving, Husband, White, Male,0,0,40, United-States, <=50K
17, Private,270942, 5th-6th,3, Never-married, Other-service, Other-relative, White, Male,0,0,48, Mexico, <=50K
30, Private,94235, HS-grad,9, Never-married, Craft-repair, Other-relative, White, Male,0,0,40, United-States, <=50K
49, Private,71195, Masters,14, Never-married, Prof-specialty, Not-in-family, White, Male,0,0,60, United-States, <=50K
19, Private,104112, HS-grad,9, Never-married, Sales, Unmarried, Black, Male,0,0,30, Haiti, <=50K
45, Private,261192, HS-grad,9, Married-civ-spouse, Other-service, Husband, Black, Male,0,0,40, United-States, <=50K
26, Private,94936, Assoc-acdm,12, Never-married, Sales, Not-in-family, White, Male,0,0,40, United-States, <=50K
```

Experimentación ETL

Comenzamos haciendo una operación de proyección, que selecciona las columnas edad, área de trabajo, educación y sexo, y otra de selección para contemplar solo aquellas instancias que se correspondan con mujeres de más de 18 años:

Operación de proyección:

```
grunt> proyeccion = foreach measure generate Age, Workclass, Education, Sex;
```

Operación de selección:

```
grunt> seleccion = filter measure by Sex == ' Female' AND Age>18;
```

En las siguientes imágenes se muestran las últimas líneas de ambas operaciones, que podemos consultar con **dump proyeccion;** o **dump seleccion;** respectivamente:

Resultado de operación de proyección:

```
(41, ?, HS-grad, Female)
(72, ?, HS-grad, Male)
(45, Local-gov, Assoc-acdm, Female)
(31, Private, Masters, Female)
(39, Local-gov, Assoc-acdm, Female)
(37, Private, Assoc-acdm, Female)
(43, Private, HS-grad, Male)
(65, Self-emp-not-inc, Prof-school, Male)
(43, State-gov, Some-college, Female)
(43, Self-emp-not-inc, Some-college, Male)
(32, Private, 10th, Male)
(43, Private, Assoc-voc, Male)
(32, Private, Masters, Male)
(53, Private, Masters, Male)
(22, Private, Some-college, Male)
(27, Private, Assoc-acdm, Female)
(40, Private, HS-grad, Male)
(58, Private, HS-grad, Female)
(22, Private, HS-grad, Male)
(52, Self-emp-inc, HS-grad, Female)
(,,,)
```

Resultado de operación de selección:

```
(81, ?,120478, Assoc-voc,11, Divorced, ?, Unmarried, White, Female,0,0,1, ?, <=50K)
(31, Private,292592, HS-grad,9, Married-civ-spouse, Machine-op-inspct, Wife, White, Female,0,0,40, United-States, <=50K)
(29, Private,125976, HS-grad,9, Separated, Sales, Unmarried, White, Female,0,0,35, United-States, <=50K)
(35, ?,320084, Bachelors,13, Married-civ-spouse, ?, Wife, White, Female,0,0,55, United-States, >50K)
(30, ?,33811, Bachelors,13, Never-married, ?, Not-in-family, Asian-Pac-Islander, Female,0,0,99, United-States, <=50K)
(37, Private,179137, Some-college,10, Divorced, Adm-clerical, Unmarried, White, Female,0,0,39, United-States, <=50K)
(34, Private,160216, Bachelors,13, Never-married, Exec-managerial, Not-in-family, White, Female,0,0,55, United-States, >50K)
(38, Private,139180, Bachelors,13, Divorced, Prof-specialty, Unmarried, Black, Female,15020,0,45, United-States, >50K)
(45, State-gov,252208, HS-grad,9, Separated, Adm-clerical, Own-child, White, Female,0,0,40, United-States, <=50K)
(41, ?,202822, HS-grad,9, Separated, ?, Not-in-family, Black, Female,0,0,32, United-States, <=50K)
(45, Local-gov,119199, Assoc-acdm,12, Divorced, Prof-specialty, Unmarried, White, Female,0,0,48, United-States, <=50K)
(31, Private,199655, Masters,14, Divorced, Other-service, Not-in-family, Other, Female,0,0,30, United-States, <=50K)
(39, Local-gov,111499, Assoc-acdm,12, Married-civ-spouse, Adm-clerical, Wife, White, Female,0,0,20, United-States, >50K)
(37, Private,198216, Assoc-acdm,12, Divorced, Tech-support, Not-in-family, White, Female,0,0,40, United-States, <=50K)
(43, State-gov,255835, Some-college,10, Divorced, Adm-clerical, Other-relative, White, Female,0,0,40, United-States, <=50K)
(27, Private,257302, Assoc-acdm,12, Married-civ-spouse, Tech-support, Wife, White, Female,0,0,38, United-States, <=50K)
(58, Private,151910, HS-grad,9, Widowed, Adm-clerical, Unmarried, White, Female,0,0,40, United-States, <=50K)
(52, Self-emp-inc,287927, HS-grad,9, Married-civ-spouse, Exec-managerial, Wife, White, Female,15024,0,40, United-States, >50K)
```

Tras esto, pasamos a resolver las preguntas que se plantearon inicialmente en la sección **Planteamiento del experimento**.

1. ¿Cuántas mujeres mayores de edad hay en cada sector de trabajo? para responderla, primero agrupamos por área de trabajo la selección hecha anteriormente que contiene mujeres mayores de 18 años:

```
grunt> measure_by_workclass = group seleccion by Workclass;
```

   y luego sobre la agrupación creada en el paso anterior, generamos una nueva agrupación sobre la que realizamos un conteo por área de trabajo:

```
grunt> women_by_workclass = foreach measure_by_workclass generate group, COUNT(seleccion.Workclass) as wk parallel 12;
```

   Obteniendo el resultado:

```
( ?,758)
( Private,7406)
( Never-worked,1)
( State-gov,486)
( Federal-gov,313)
( Without-pay,5)
( Local-gov,824)
( Self-emp-not-inc,392)
( Self-emp-inc,132)
```

Donde podemos ver que hay un gran número de mujeres que trabajan en el ámbito privado, y que las demás encuestadas ocupan puestos en el gobierno estatal, federal o local, que hay muchas cuyo trabajo se desconoce y que hay 5 que no reciben remuneración.

2. Para cada sector de trabajo, ¿cual es la edad media de las mujeres por sector? Para resolver esta pregunta de nuevo usamos la agrupación por área de trabajo de la selección del grupo de interés realizada al principio, y sobre ella generamos una nueva agrupación sobre la que realizamos la media del atributo edad por área de trabajo:

```
grunt> women_age_by_workclass = foreach measure_by_workclass generate group, AVG(seleccion.Age) as measure parallel 12;
```

Obteniendo como resultado:

```
( ?,38.60290237467019)
( Private,36.411153119092624)
( Never-worked,30.0)
( State-gov,38.53086419753087)
( Federal-gov,42.0926517571885)
( Without-pay,58.6)
( Local-gov,42.05582524271845)
( Self-emp-not-inc,44.31377551020408)
( Self-emp-inc,44.265151515151516)
grunt>
```

Donde podemos ver, para cada sector, la edad media de las mujeres que trabajan en él. Por ejemplo, podemos ver que las mujeres no remuneradas son las que tienen en media la edad más alta (58 años), que la edad media de las mujeres autónomas es de 44 años, o que la edad media de las mujeres que no han trabajado nunca es de 30 años.