

Experimentación ETL con Impala: Big data 1

Máster en ciencia de datos e ingeniería de computadores - UGR

Curso 2018

M^a Cristina Heredia Gómez

Índice

Base de datos seleccionada	3
Experimentación ETL con Impala	3
Planteamiento del experimento	3
Desarrollo del experimento	4
Carga de los datos	4
Experimentación ETL	5

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 - país de nacimiento

Experimentación ETL con Impala

Planteamiento del experimento

Ante esta Base de datos, se diseña una experimentación que trata de hacer un estudio social, estudiando la situación laboral de colectivos menos favorecidos o más discriminados, como son el colectivo femenino y los inmigrantes. En concreto, a lo largo del experimento, se buscará responder a las siguientes preguntas:

1. ¿a qué sector de trabajo pertenecen mayoritariamente las personas contempladas en la BD?
2. ¿qué representación por sexo hay en cada sector de trabajo?
3. ¿influye el sexo en el sueldo anual que perciben las personas?
4. ¿influye la raza en el sueldo anual que perciben las personas?

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

Desarrollo del experimento

Carga de los datos

Comenzamos creando el directorio hadoop de nombre **input** donde posteriormente cargamos los datos:

pantalla/creacionHDFSypuestaDatos.png

```
[cloudera@quickstart ~]$ mkdir input
[cloudera@quickstart ~]$ hdfs dfs -ls input
[cloudera@quickstart ~]$ cd input/
[cloudera@quickstart input]$ ls
[cloudera@quickstart input]$ cd ..
[cloudera@quickstart ~]$ hdfs dfs -put /var/tmp/materialImpala/adult.data /input
```

Una vez copiados los datos, comprobamos que el fichero se ha cargado correctamente en HDFS:

pantalla/checkingItenExist.png

```
[cloudera@quickstart ~]$ hdfs dfs -ls /input
Found 1 items
-rw-r--r-- 1 cloudera supergroup 3974305 2018-03-24 08:05 /input/adult.data
```

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

El siguiente paso será crear la tabla en impala y cargar los datos del fichero en ella. Sin embargo, para poder hacer esto, es necesario cambiar los permisos para dotar de permiso de escritura al usuario en todo el directorio **input** creado anteriormente, pues en caso contrario mostrará un error indicando que no se tiene permiso de escritura sobre ese directorio y no podremos cargar los datos a la tabla creada en impala:

pantalla/cambioPermisos.png

```
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -chown -R impala:supergroup /input
```

Ahora que tenemos permiso de escritura en todo el directorio, creamos la tabla correspondiente en Impala para cargar esa BD y comprobamos que se ha creado correctamente. Para cargar la tabla, usamos la sentencia MySQL siguiente:

Listing 1: Sentencia SQL para creación de la tabla

```
CREATE TABLE IF NOT EXISTS Census (Age INT, Workclass STRING, Fnlwgt BIGINT,
Education STRING, EducationNum INT, MaritalStatus STRING, Occupation STRING,
Relationship STRING, Race STRING, Sex STRING, capitalGain BIGINT, CapitalLoss
INT, HoursPerWeek INT, Country STRING, YearGain STRING ) ROW FORMAT
5 DELIMITED FIELDS TERMINATED BY '\,' STORED AS TEXTFILE;
```

y comprobamos que se ha creado correctamente:

pantalla/describeCensus.png

```
[quickstart.cloudera:21000] > DESCRIBE Census;
Query: describe Census
+-----+-----+-----+
| name      | type  | comment |
+-----+-----+-----+
| age       | int   |          |
| workclass | string|          |
| fnlwgt    | bigint|          |
| education | string|          |
| educationnum | int   |          |
| maritalstatus | string|          |
| occupation | string|          |
| relationship | string|          |
| race      | string|          |
| sex       | string|          |
| capitalgain | bigint|          |
| capitalloss | int   |          |
| hoursperweek | int   |          |
| country    | string|          |
| yeargain   | string|          |
+-----+-----+-----+
Fetched 15 row(s) in 1.54s
```

Para finalizar la carga de los datos, cargamos los datos contenidos en el directorio **input** a la tabla recién creada en Impala:

pantalla/cargaDatosEnTablaImpala.png

```
[quickstart.cloudera:21000] > LOAD DATA INPATH '/input/adult.data' OVERWRITE INTO TABLE Census;
Query: load DATA INPATH '/input/adult.data' OVERWRITE INTO TABLE Census
+-----+
| summary                                     |
+-----+
| Loaded 1 file(s). Total files in destination location: 1 |
+-----+
| Fetched 1 row(s) in 0.46s                      _
```

Experimentación ETL

Inicialmente, hacemos una operación de proyección y otra de selección para tantear la BD (no se muestra el resultado por su larga extensión):

Listing 2: Selección de la edad - sector de trabajo y estudios

```
-- proyeccion de columnas edad, trabajo, estudios, sexo:
SELECT Age, Workclass, Education, Sex FROM Census;

-- seleccion de las mujeres mayores de edad:
5 SELECT * FROM Census WHERE Sex="Female" AND Age > 18;
```

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

1. ¿A qué sector de trabajo pertenecen mayoritariamente las personas de la BD? para responderla, seleccionamos la columna Workclass y un conteo de los valores de la misma, agrupados por área de trabajo:

pantalla/workclassCount.png

```
[quickstart.cloudera:21000] > SELECT Workclass, COUNT(*) FROM Census GROUP BY Workclass;
Query: select Workclass, COUNT(*) FROM Census GROUP BY Workclass
Query submitted at: 2018-03-29 09:43:43 (Coordinator: http://quickstart.cloudera:25000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?query_id=a5486c179d7df0a4:4ed273d100000000
+-----+
| workclass | count(*) |
+-----+
| NULL      | 1         |
| Self-emp-not-inc | 2541      |
| Federal-gov | 960       |
| Private   | 22696     |
| Self-emp-inc | 1116      |
| State-gov | 1298      |
| Without-pay | 14        |
| Local-gov | 2093      |
| Never-worked | 7         |
| ?         | 1836      |
+-----+
| Fetched 10 row(s) in 0.47s
```

Donde podemos ver que el área de trabajo más numerosa es el ámbito privado, es decir, la mayoría de la gente de esta BD trabaja en empresas privadas.

2. ¿Qué representación por sexo hay en cada sector de trabajo? obtenemos las columnas de área de trabajo y sexo y el conteo de área de trabajo por sexo agrupadas por área de trabajo y sexo, y lo ordenamos por área de trabajo, para que la salida sea más interpretable:

pantalla/workclassBysex.png

Query: select Workclass,Sex,count(*) FROM Census GROUP BY Workclass,Sex order by workclass
 Query submitted at: 2018-03-24 10:46:40 (Coordinator: http://quickstart.cloudera:25000)
 Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?query_id=124e8874fac2724e:ac547c6400000000

workclass	sex	count(*)
?	Female	839
?	Male	997
Federal-gov	Male	645
Federal-gov	Female	315
Local-gov	Male	1258
Local-gov	Female	835
Never-worked	Female	2
Never-worked	Male	5
Private	Female	7752
Private	Male	14944
Self-emp-inc	Male	981
Self-emp-inc	Female	135
Self-emp-not-inc	Male	2142
Self-emp-not-inc	Female	399
State-gov	Female	489
State-gov	Male	809
Without-pay	Female	5
Without-pay	Male	9
NULL		1

Donde podemos ver, para cada sector, la cantidad de mujeres y hombres que hay trabajando en él. Por ejemplo, trabajando en el gobierno federal se encuentran más del doble de hombres(645) que mujeres (315). Para el gobierno local, gobierno estatal autónomos y área privada se da una situación similar, aunque esto puede deberse a que hay más datos recogidos de hombres que de mujeres. Hacemos un conteo de las muestras contenidas en la BD por sexo:

pantalla/sexcount.png

[quickstart.cloudera:21000] > SELECT Sex, Count(*) FROM Census GROUP BY Sex;
 Query: select Sex, Count(*) FROM Census GROUP BY Sex
 Query submitted at: 2018-03-29 10:32:34 (Coordinator: http://quickstart.cloudera:25000)
 Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?query_id=84e1bb40aead1eb:ccf9181700000000

sex	count(*)
	1
Male	21790
Female	10771

Fetches 3 row(s) in 0.43s

y efectivamente, nos encontramos con que los hombres encuestados duplican a las mujeres encuestadas.

- ¿influye el sexo en el sueldo anual que perciben las personas? a pesar de que ahora sabemos que los distintos géneros no están bien representados en esta BD, continuamos el experimento. Para resolver esta cuestión comenzamos mirando cuántas personas hay que ganan menos de 50.000 al año, por sexo y por grupo de trabajo. Limpiamos los datos filtrando los valores nulos:

pantalla/gananMenosde50000PorSexo.png

Query: select Workclass,Sex, count(*) FROM Census WHERE workclass!="NULL" AND Fnlwgt < 50000 GROUP BY Workclass,Sex order by workclass
 Query submitted at: 2018-03-27 04:07:38 (Coordinator: http://quickstart.cloudera:25000)
 Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?query_id=434a436352908c74:5f80070500000000

workclass	sex	count(*)
?	Female	70
?	Male	85
Federal-gov	Male	61
Federal-gov	Female	68
Local-gov	Female	61
Local-gov	Male	96
Private	Male	877
Private	Female	471
Self-emp-inc	Male	64
Self-emp-inc	Female	8
Self-emp-not-inc	Male	248
Self-emp-not-inc	Female	38
State-gov	Female	47
State-gov	Male	76
Without-pay	Female	1
Without-pay	Male	1

Aunque no podemos afirmar nada de los casos en los que hay más hombres que mujeres que cobran

menos de 50.000, dado que la representación por género en la BD está sesgada, sí que podemos decir que resulta llamativo el caso del gobierno federal, ya que hay más mujeres que hombres cobrando menos de 50.000 dólares, a pesar de que hay la mitad de mujeres que de hombres trabajando para el gobierno deferal. Esto puede deberse a que las pocas mujeres que trabajan para el gobierno federal ocupan cargos inferiores a los hombres que trabajan para esta misma área.

4. ¿influye la raza en el sueldo anual que perciben las personas? Para resolver esta pregunta, comprobamos cuantas personas ganan menos de 50000 dólares por sexo y por raza, para todos los tipos de trabajo:

pantalla/cuantossexoyrazaporWorkclassgananMenosDe50000.png

```
[quickstart.cloudera:21000] > SELECT Sex,Race, count(workclass) FROM Census WHERE workclass!="NULL" AND Fnlwgt < 50000 GROUP BY Sex, Race order by Race ;
Query: select Sex,Race, count(workclass) FROM Census WHERE workclass!="NULL" AND Fnlwgt < 50000 GROUP BY Sex, Race order by Race
Query submitted at: 2018-03-27 04:15:43 (Coordinator: http://quickstart.cloudera:25000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?query_id=24e86905e3bb7d8:5bfc5d4000000000
```

sex	race	count(workclass)
Female	Amer-Indian-Eskimo	40
Male	Amer-Indian-Eskimo	48
Male	Asian-Pac-Islander	21
Female	Asian-Pac-Islander	12
Male	Black	24
Female	Black	48
Female	Other	9
Male	Other	2
Male	White	1413
Female	White	655

Fetch 10 row(s) in 0.65s

Aunque para el resto de razas no se puede afirmar una diferencia significativa, sí se aprecia que para el caso de la raza negra, existen el doble de mujeres que cobran menos que los hombres de raza negra, a pesar de que en los datos las mujeres están la mitad de representadas. Por esta razón, sería un buen caso de estudio la discriminación salarial hacia las mujeres de raza negra.

Para concluir el estudio, comprobamos si hay diferencias significativas por raza y por sexo entre las personas que perciben más de 50.000 dólares, para todas las áreas de trabajo:

pantalla/cuantosSexoyRazaPorWorkclassGananMasDE50000.png

```
[quickstart.cloudera:21000] > SELECT Sex,Race, count(workclass) FROM Census WHERE workclass!="NULL" AND Fnlwgt >= 50000 GROUP BY Sex, Race order by Race ;
Query: select Sex,Race, count(workclass) FROM Census WHERE workclass!="NULL" AND Fnlwgt >= 50000 GROUP BY Sex, Race order by Race
Query submitted at: 2018-03-27 04:17:23 (Coordinator: http://quickstart.cloudera:25000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?query_id=4946c0226da3ce24:61f7128500000000
```

sex	race	count(workclass)
Female	Amer-Indian-Eskimo	79
Male	Amer-Indian-Eskimo	144
Male	Asian-Pac-Islander	672
Female	Asian-Pac-Islander	334
Male	Black	1545
Female	Black	1507
Female	Other	100
Male	Other	160
Male	White	17761
Female	White	7987

Fetch 10 row(s) in 0.66s

En este caso, comprobamos que para todas las razas se cumple que hay más hombres que mujeres con un salario superior o igual a 50.000 dólares. También parece existir una gran diferencia no sólo en el salario percibido por sexo, sino también por raza, pues por ejemplo, hay 17761 hombres de raza blanca ganando más de 50.000 dólares anuales frente a 144 indios americanos percibiendo ese salario.