ETL Process for Data Transformation

20 April 2020 12:02 pm

Step 1: Extract

I uploaded the adult-training.csv file to a Microsoft Azure Jupyter notebook, and then used some R language commands to find all distinct levels of data in each column.
For example, here is some code I used

- US_income <- read.csv("adult-training.csv")
  - reads in the data from the csv file

```
In [2]: US_income <- read.csv("adult-training.csv")
```

- summary(US_income[,4])
  - finds all levels of education data, the 4th column, and summarises into Qualifications, finding the number of rows (people) for each type of qualification

```
In [8]: summary(US_income[,4])
```

10th	933
11th	1175
12th	433
1st-4th	168
5th-6th	333
7th-8th	646
9th	514
Assoc-acdm	1067
Assoc-voc	1382
Bachelors	5354
Doctorate	413
HS-grad	10501
Masters	1723
Preschool	51
Prof-school	576
Some-college	7291

- summary(US_income[,7])
  - finds all levels of occupation data, the 7th column, and summarises into Job types, finding the number of rows (people) for each type of job

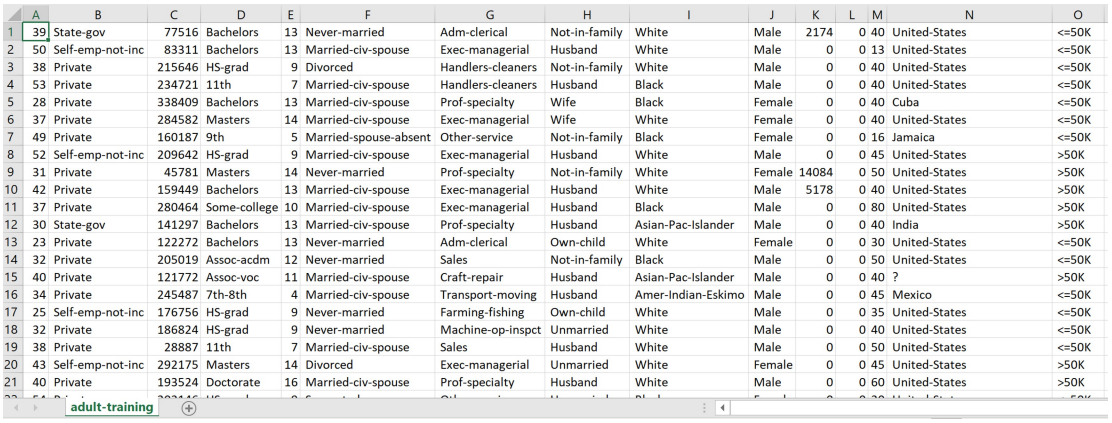
```
In [2]: summary(US_income[,7])
```

?	1843
Adm-clerical	3769
Armed-Forces	9
Craft-repair	4099
Exec-managerial	4066
Farming-fishing	994
Handlers-cleaners	1370
Machine-op-inspct	2002
Other-service	3295
Priv-house-serv	149
Prof-specialty	4140
Protective-serv	649
Sales	3650
Tech-support	928
Transport-moving	1597

Step 2: Transform

I transformed the csv file by replacing the values in each column with ID (primary/foreign key values) using the Excel VLOOKUP function
Here are the steps involved

- Original CSV



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
2	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
3	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
4	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
5	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
6	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
7	49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
8	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
9	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
10	42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
11	37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
12	30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
13	23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
14	32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
15	40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
16	34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
17	25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
18	32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
19	38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
20	43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
21	40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K

adult-training

- Inserted a blank column next to the column to transform. We are going to transform the 'Race' dimension column in this example

I	J	
White		
White		
White		
Black		
Black		
White		
Black		
White		
White		
White		
Black		
Asian-Pac-Islander		
White		
Black		
Asian-Pac-Islander		
Amer-Indian-Eskimo		
White		
White		
White		
White		
White		

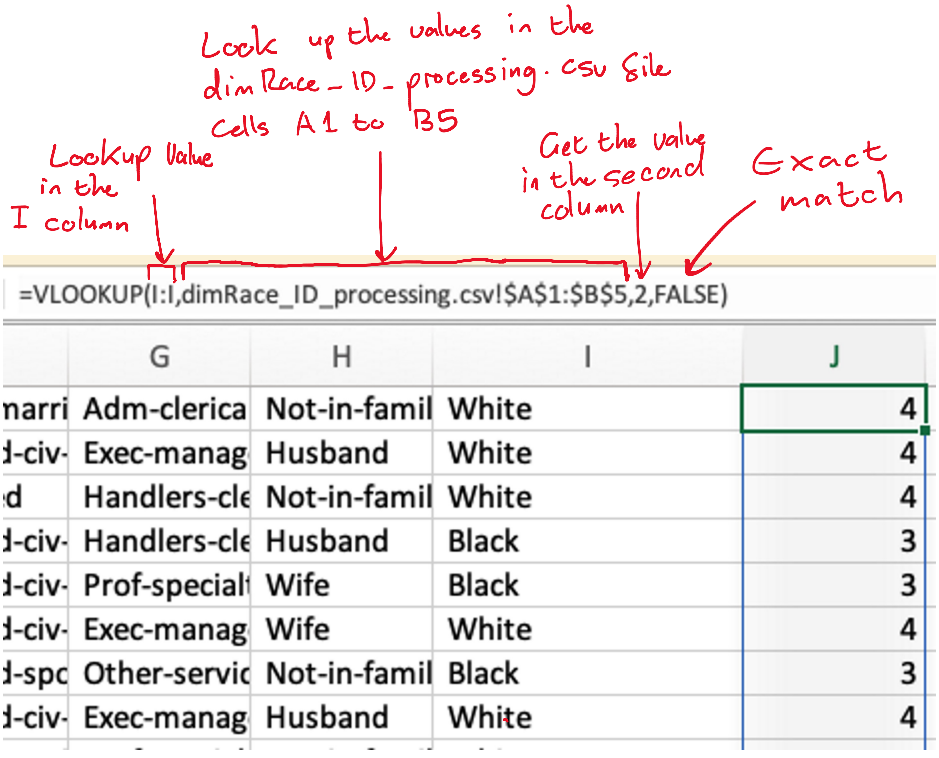
- Created a new CSV file, specifically for processing, for each dimension with the data value (race in this example) and a unique integer value to be used as the ID/Primary Key
  Filename: dimRace_ID_processing.csv

	A	B
1	Amer-Indian-Eskimo	1
2	Asian-Pac-Islander	2
3	Black	3
4	White	4
5	Other	5
6		

So, the Amer-Indian-Eskimo race will have an ID of 1, Asian-Pac-Islander race will have an ID of 2 etc.

- Used the VLOOKUP Excel function in the adult-training.csv file to find the ID/Key value in the processing CSV file for that dimension (in this example, Race)

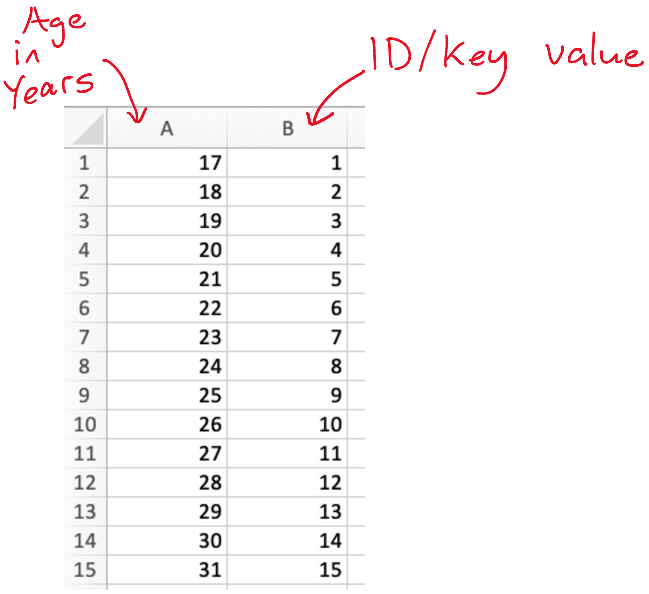
  Code: =VLOOKUP(I:I,[dimRace_ID_processing.csv]dimRace_ID_processing!$A$1:$B$5,2,FALSE)



	G	H	I	J
narri	Adm-clerica	Not-in-famil	White	4
d-civ-	Exec-manag	Husband	White	4
d	Handlers-cle	Not-in-famil	White	4
d-civ-	Handlers-cle	Husband	Black	3
d-civ-	Prof-specialt	Wife	Black	3
d-civ-	Exec-manag	Wife	White	4
d-spc	Other-servic	Not-in-famil	Black	3
d-civ-	Exec-manag	Husband	White	4

A second example, the ID processing for the Age dimension.

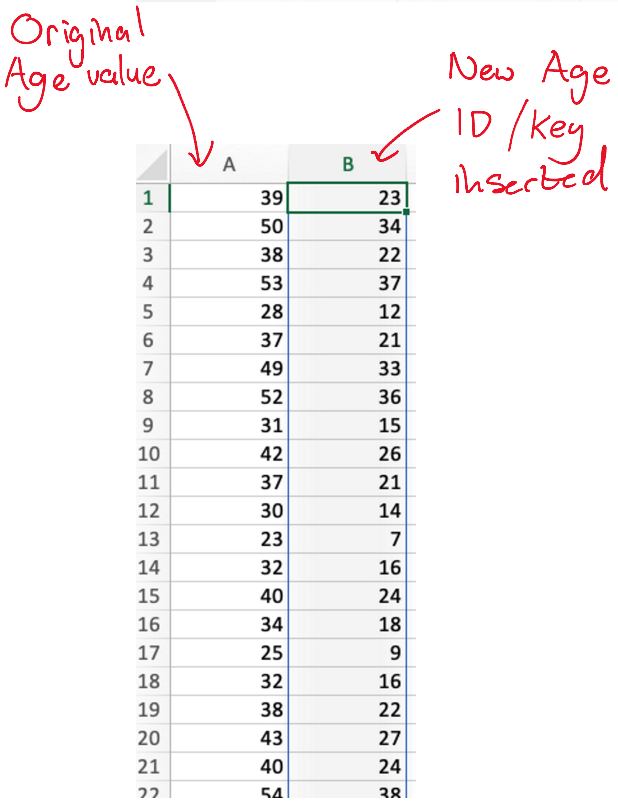
- A csv file was created with the Age in Years and ID/Key integer values
  Filename: dimAge_ID_processing.xlsx



	A	B
1	17	1
2	18	2
3	19	3
4	20	4
5	21	5
6	22	6
7	23	7
8	24	8
9	25	9
10	26	10
11	27	11
12	28	12
13	29	13
14	30	14
15	31	15

- New column is inserted next to the Age column in the adult-training.csv file
- VLOOKUP function is executed in cells of new column
- New column now contains ID/Key values

```
=VLOOKUP(A:A,[dimAge_ID_processing.xlsx]Sheet1!$A$1:$B$74,2,FALSE)
```



	A	B
1	39	23
2	50	34
3	38	22
4	53	37
5	28	12
6	37	21
7	49	33
8	52	36
9	31	15
10	42	26
11	37	21
12	30	14
13	23	7
14	32	16
15	40	24
16	34	18
17	25	9
18	32	16
19	38	22
20	43	27
21	40	24
22	54	38

- The same process is repeated for the dimension columns:
    - Age
    - Gender
    - Race
    - Native Country
    - Education
    - Occupation
    - Marital Status

- The original columns are deleted after processing is complete as they are no longer needed
- First column is left empty for auto incrementation of the 'Person_ID' column in the Database which contains primary key values for each row in the Fact Table
- Column headers are removed
- Fact Table is ready to be loaded into the Database

Finished Product: Processed Fact Table

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Age_ID	Gender_ID	Race_ID	NativeCountry_ID	Education_ID	Occupation_ID	MaritalStatus_ID	Income_Bracket	Capital Gain	Capital Loss	Hours_per_week	Fnlwgt
2		23	1	4	39	13	1	5	<=50K	2174	0	40	77516
3		34	1	4	39	13	4	2	<=50K	0	0	13	83311
4		22	1	4	39	9	6	4	<=50K	0	0	40	215646
5		37	1	3	39	7	6	2	<=50K	0	0	40	234721
6		12	2	3	5	13	10	2	<=50K	0	0	40	338409
7		21	2	4	39	14	4	2	<=50K	0	0	40	284582
8		33	2	3	23	5	8	3	<=50K	0	0	16	160187
9		36	1	4	39	9	4	2	>50K	0	0	45	209642
10		15	2	4	39	14	10	5	>50K	14084	0	50	45781
11		26	1	4	39	13	4	2	>50K	5178	0	40	159449
12		21	1	3	39	10	4	2	>50K	0	0	80	280464
13		14	1	2	19	13	10	2	>50K	0	0	40	141297
14		7	2	4	39	13	1	5	<=50K	0	0	30	122272
15		16	1	3	39	12	12	5	<=50K	0	0	50	205019
16		24	1	2	42	11	3	2	>50K	0	0	40	121772
17		18	1	1	26	4	14	2	<=50K	0	0	45	245487
18		9	1	4	39	9	5	5	<=50K	0	0	35	176756
19		16	1	4	39	9	7	5	<=50K	0	0	40	186824
20		22	1	4	39	7	12	2	<=50K	0	0	50	28887
21		27	2	4	39	14	4	4	>50K	0	0	45	292175
22		24	1	4	39	16	10	2	>50K	0	0	60	193524
23		38	2	3	39	9	8	6	<=50K	0	0	20	302146

adult-training processed_header

Step 3: Load

The data warehouse is created and loaded through executing two SQL files.

Creating Database and Tables

- The first file is called Project01_CreateTables.sql.
- When the file is executed in SQL Server Management Studio, the Database is created along with all Dimension Tables and the Fact Table.
- Relations are set between the Fact Table foreign keys and Dimension Table primary keys

Loading Data into the Data Warehouse

- Variables are set for the source path of the data for the Dimension Tables and Fact Table
- Bulk insert statements are executed to load data into each table from the corresponding CSV File
- For example,
  BULK INSERT [dbo].[FactAdultIncome] FROM '$(Sql_FactTable_SourceDataPath) adult-training processed.csv'

inserts data into the 'FactAdultIncome' table from the 'adult-training processed.csv' file located in the path set by the variable 'Sql_FactTable_SourceDataPath'

NOTE: please change the source path to where the Dimension and Fact Table CSV files are located after extracting the ZIP file. On my computer they are located in my Dropbox folder.