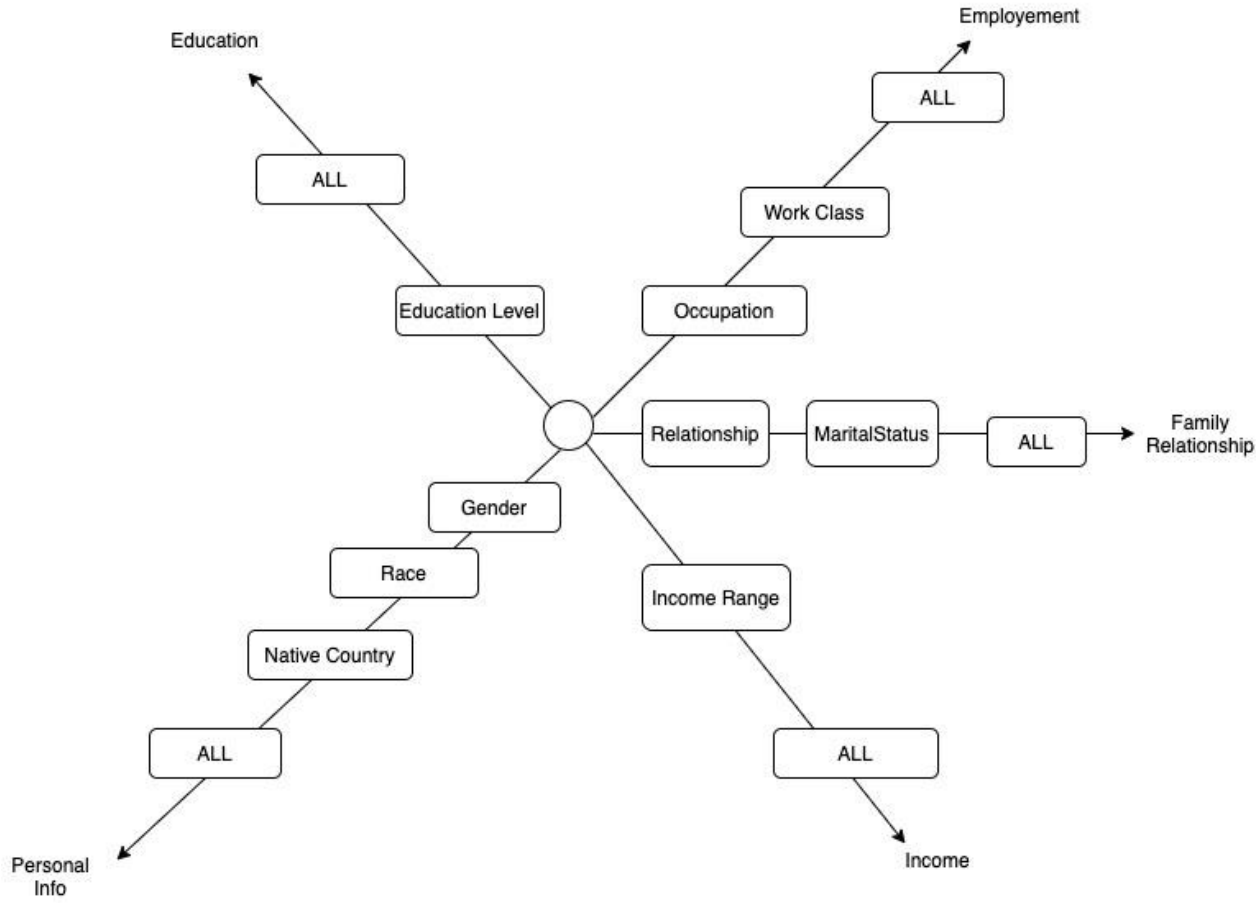
MID-PROJECT SUBMISSION
CITS3401

CHAMIKA KARIYAWASAM GAMAGE
22508087

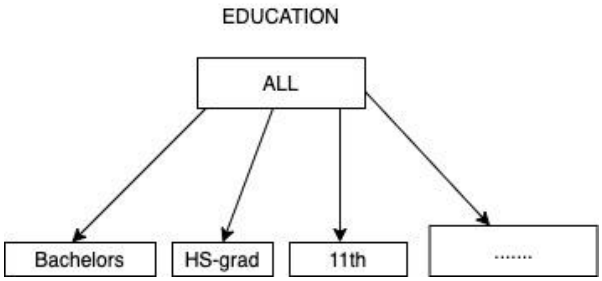
Starnet



The Starnet reflects the concept hierarchies within each dimension.

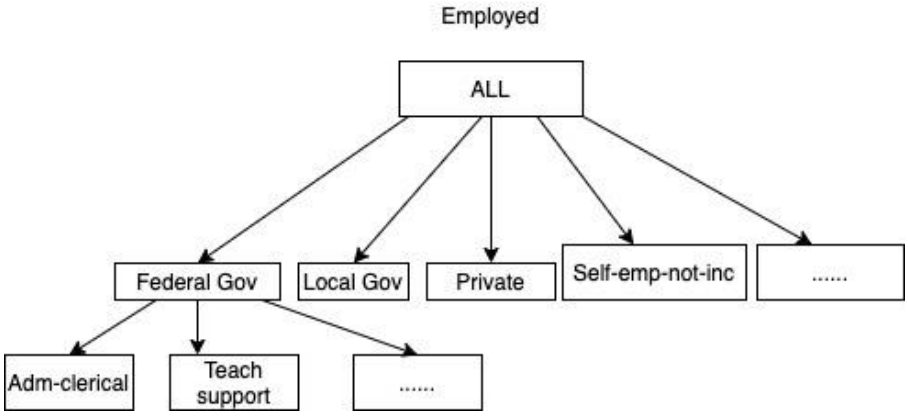
HIERACHIES

EDUCATION



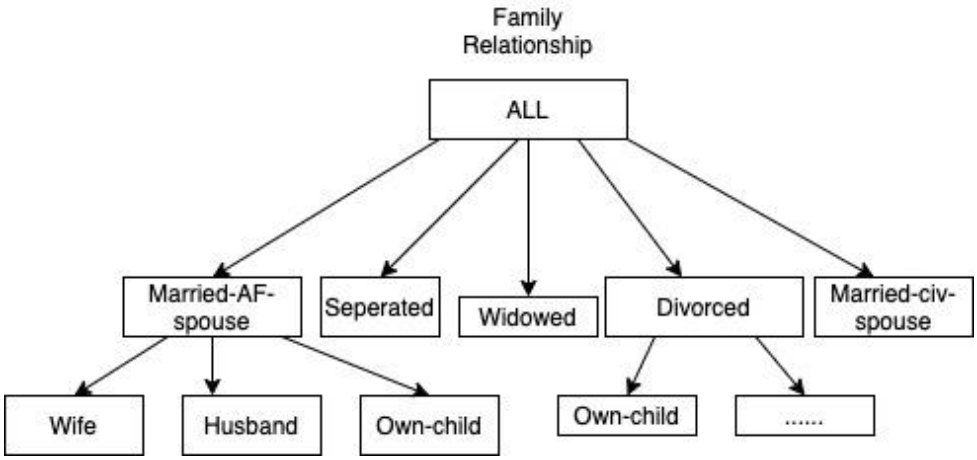
The education level has ALL and then branches out to specific levels of education. There is positional to group the education into primary and secondary. By this I mean group all the education up to HS-grad into primary and then group the rest into higher education. There could be argument made here that education,emplyement could be grouped into one. I have seen this in economics terms but I believe for this project its best to keep education alone.

EMPLOYEMENT



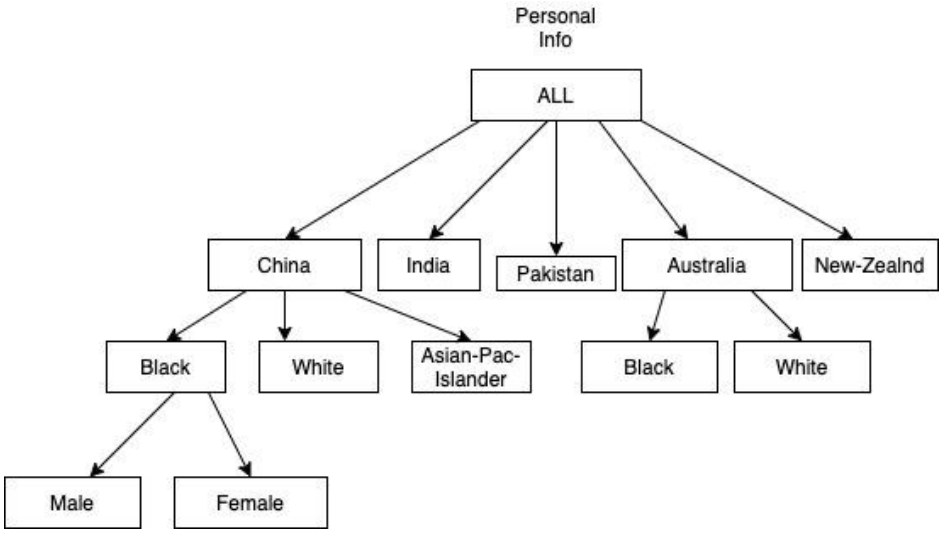
The hierarchy for employment starts off with ALL. From All it goes to work class. Work class corresponds to a general class of workers. From work class it goes to individual occupation of the workers. Note that I have not drawn each occupation that's under each work class. This will take too much space but the above picture is a general idea. There could be an argument made here as to including hours worked in here. However including hours worked as a hierarchy makes the dimension table far too big and ends up with the table having almost 1000 rows and ids.

FAMILY RELATIONSHIP



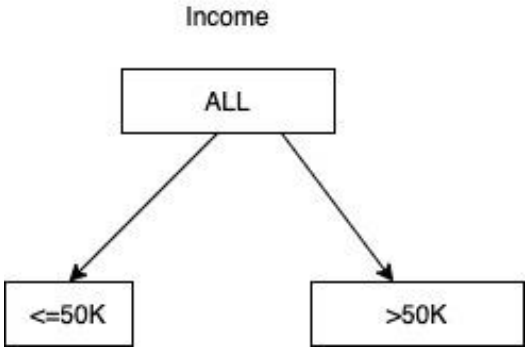
The hierarchy of Family relationship extends from ALL to marital status then relationship. This hierarchy could be changed the other way around as it is very subjective. The hierarchy in the image makes sense because you have to be married then you have a relationship within that marriage. This is highly subjective as vice versa could also make sense.

PERSONAL INFO



The hierarchy for personal information goes from ALL to Native country and then your race and then your gender. This hierarchy is also very subjective to whoever is making the hierarchies. Gender could be left out and made into a separate dimension table but it makes sense to include gender as personal information because it is categorized under there.

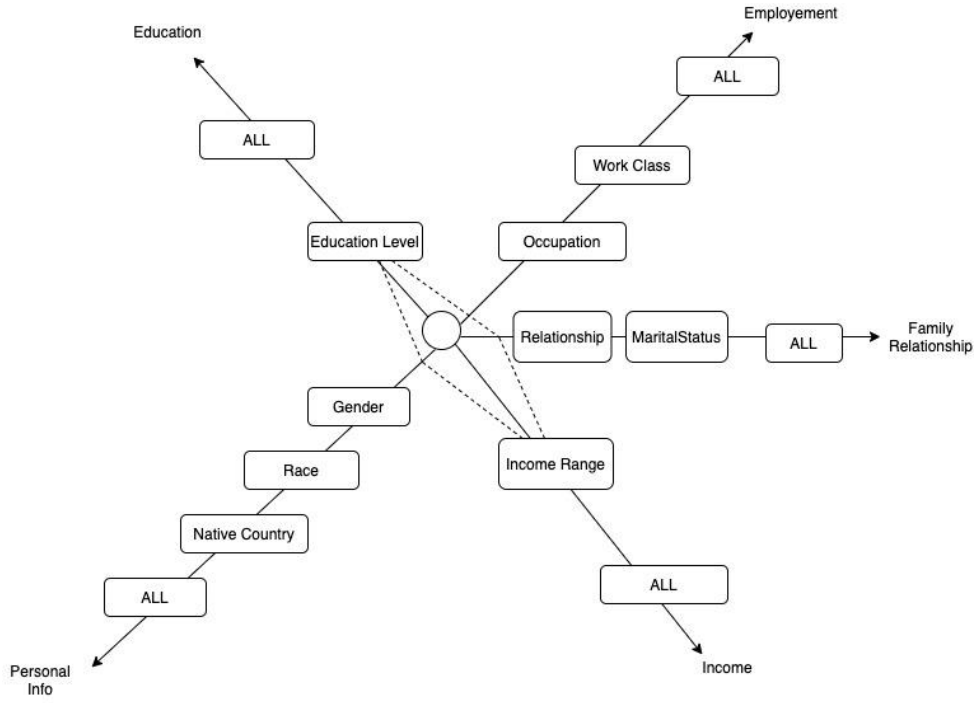
INCOME



Income hierarchy extends from ALL to earning less than <=50k and greater than 50k.
The argument could be made as to why this dimension even exists. The reason why this is as it helps with when doing analysis.

BUSINESS QUESTIONS

1.What percentage of individuals with a bachelors earn over 50k?

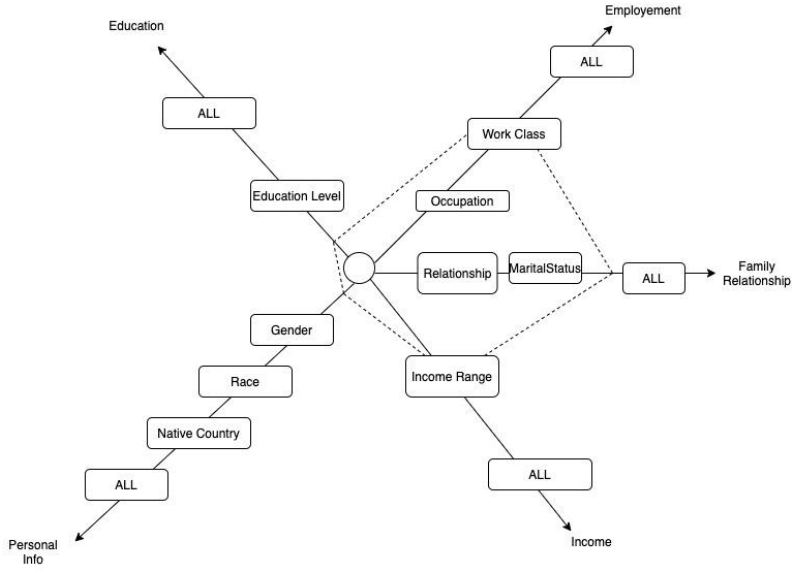


Measure here is fact individual count

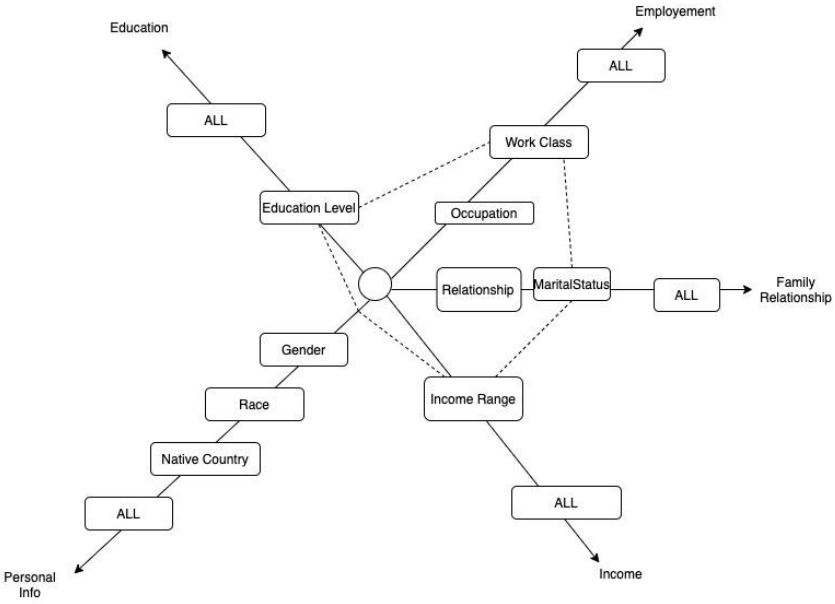
2.Which level of education is the most common with people earning over 50k?
Same Starnet as above but different breakdown of answer.
Measure here is fact individual count

3.What level of government has the smallest wage gaps amongst people?

Measure here is fact individual count

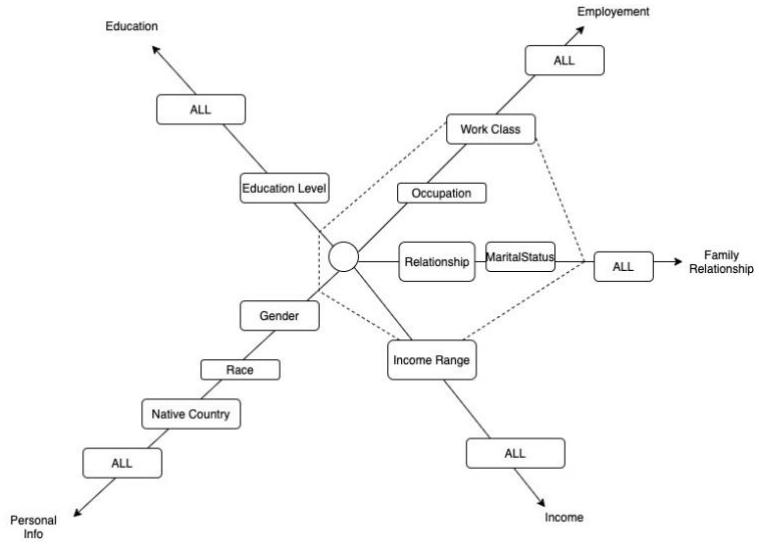


4. How many HS-grads within the work class of private are Divorced and earn over 50k?



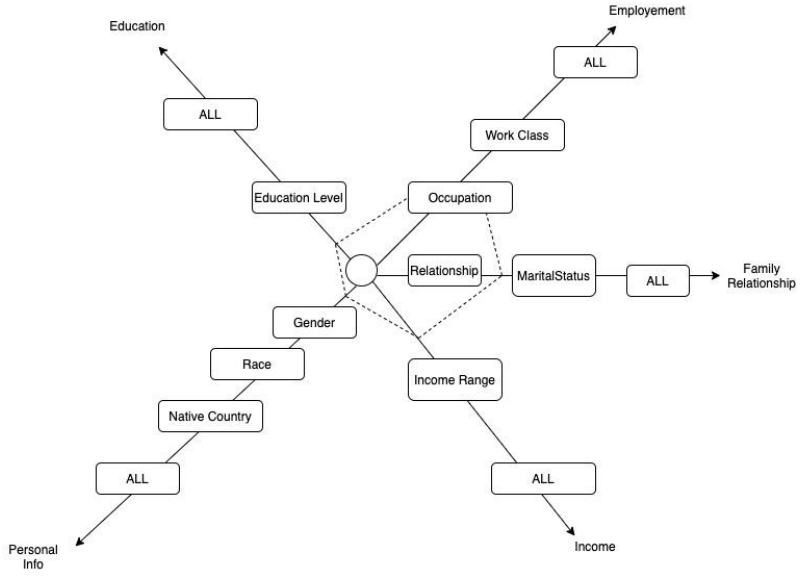
Measure is fact individual count

5. Which work class works the most to earn more than >50k



The measure here is hours instead of individual count

6. Which occupation has the oldest workers



Measure here is age instead of fact individual count

ETL PROCESS

The etl process for this project was done through python. The script attached in this zip folder will demonstrate how the IDS are matched from dimension table to fact table. This piece of code is a snippet to as how the matching works.
First we open the master data and group the rows. Here row 5 and 7 are under family.

```
##          Family
          family_row = [row[5].strip(), row[7].strip()]
          FamilyID = Family(family_row)
```

Then we open the family csv which has all unique combination of marital status and relationship. If the data that read in the master matched the data in dimension, then we transfer that id.

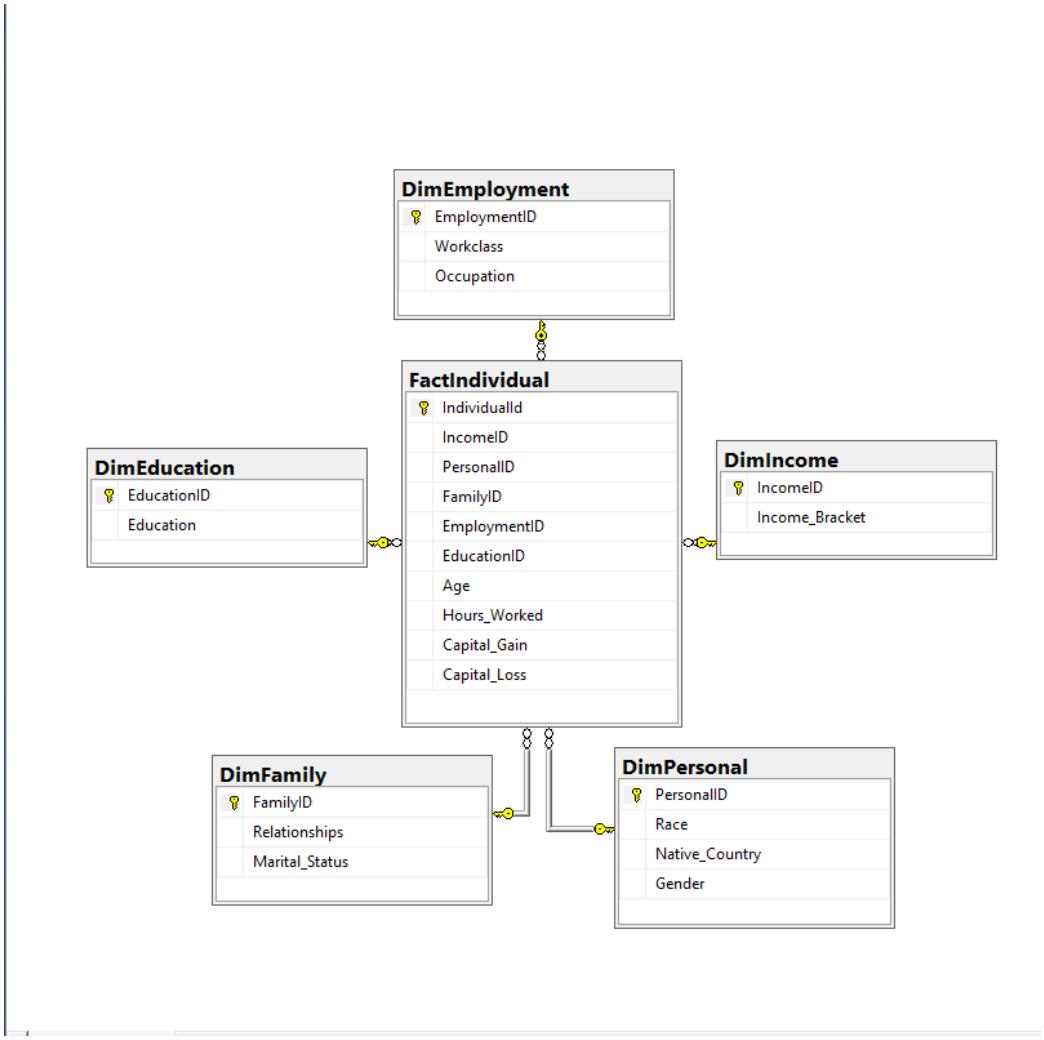
```
with open('Family.csv', 'r') as family_file:
    family_writer = csv.reader(family_file, delimiter=',')
    id = 0
    for row in family_writer:
      id += 1
      Dimension_family_row = [row[1].strip(), row[2].strip()]
      if family_row == Dimension_family_row:
        return id
```

The data for income range was manually manipulated because there are only 2 IDS. For example I just copied and pasted all the income range data into fact table, then I just did search and replace income >50K with 2 and so on.

In the master data file, data rows with null values were deleted all together. This was to help with the integrity of the data and it conflicted when bulk loading the data to have empty values.

There is another python script I used to get the family data because there are more than one way of doing the matching of IDs.

SCHEMA



The reasoning behind a star scheme is for pure simplicity sake. There is potential here to place the family dimension inside the personal dimension and make it a snowflake. However this makes no logical sense in terms of concepts hierarchies. I believed it is best to group data about personal background which is native country, race and gender and then group all the data about relationships and other aspects of life into a separate dimension. The data set is also not big enough to use snowflake. There are multiple other data sets in Kaggle such as Airbnb data and housing data which makes sense of using snowflake scheme.

The answers to the queries are on the powerBI pdf file.

Please note there is an error in this project regarding visual studio. The concept hierarchies for personal dimension and family dimension seem not to be working. What I mean is When I input the hierarchy for personal dimension as all-Country-Race-Gender, it converts the hierarchy to all-Race-Country-Gender. I cannot fix it. This doesn't effect when answering the queries because the 2 are mixed around. The same applies for family and the marital status and relationship. I cannot get the concept hierarchies to work but I did manage to answer the queries properly.

DATA CUBE