Data Warehousing Project - Report

Student Number: 17377463
Student Name: Nathan Crone
Domain Area of this Data warehouse: Education

1. Background, Idea, Motivation

My first step when starting this project was thinking of an idea to run with. I had seen an example of a star schema in the first few lectures we had; however, I was not sure of another idea of a star schema that would be as good as the example used in the slides.
I had thought about taking a sports route for this project and doing my data warehouse on football as this is my favourite sport, but I didn't really want to do this as I wanted to be original and had heard from others that they were doing projects on various different sports.
After about a week of thinking and running through ideas in my head, the idea of doing a data warehouse on education and grades in school popped into my head.
I came up with this idea as my father is a teacher and I often hear him wonder about the overall performance of his students. I wanted to solve this problem by creating a data warehouse focused on analysing the students' grades. I figured that if I could do this for this project, this data warehouse would also have some practical use in the world outside of college after this module as I could easily rearrange it to suit my father's students' school grades.
While I wasn't sure of the exact dimensions that I would use, I was determined to devise dimensions that would enable me to analyse students' grades in great detail.
After I went through the process of designing the star schema, I decided that what I wanted my data warehouse to do was to help me understand how students' grades are affected depending on the subject they take, their personal attributes, where they live, the school they go to, and the year they do their exams in.
I thought that if I could represent the database in this way, I would have a vast number of dimension columns to use when designing OLAP queries for querying the data warehouse.

2. Warehouse schema design (description of why you designed as you did). Build in MySQL.

When I officially started my project, the natural first step was to create the star schema. I started doing this on the website “Vertabelo.com”, which one of my class members had recommended to me due to its ability to turn a completed star schema into an SQL database for you.

I started by building a fact table in the diagram and then had to decide on dimensions that this table would link to.

My first thoughts were to create a “School” and “Time” dimension as this would allow me to compare how the average leaving cert points of a year group vary by the year, and what school they go to.

I then followed this by creating a “Student” dimension along with a “Residence” dimension. I figured that this would make the data warehouse more refined and would allow me to go through every person within a particular year group to see how each individual’s leaving cert points are affected by the year, the students' personal attributes, the school they go to, and the area that they live in.

To make the data warehouse even more refined, I decided to add in a “Subject” dimension and to change the dependent variable to the individual grade per subject as opposed to the students overall leaving cert points. This allowed me to see how the particular subject that a student takes can also affect the students results.

After this, I was satisfied that I had enough detail in my schema that I could perform detailed analyses of the students’ grades in my queries once the database was created.

With those dimensions, I would be able to see how a students grade in a particular subject is affected by the particular subject, where the student lives, the school the student goes to, the year and the personal attributes of the student.

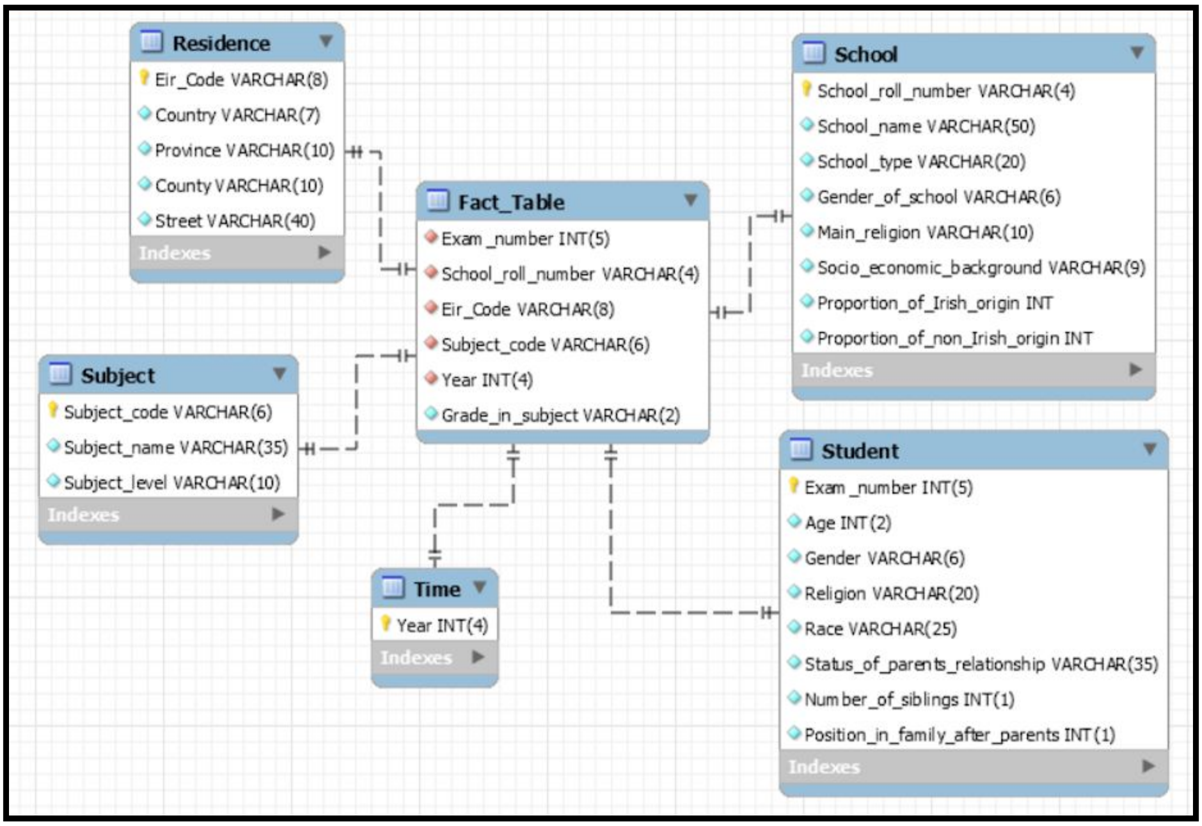
As I was satisfied, I then proceeded to fill in all of the columns of each dimension along with the type of data that would be allowed to be entered into each column.

I did not want to create an extra column in each dimension to host a unique attribute for the sole purpose of being the primary key, so once I had filled in all the columns of each dimension, I went through each dimension and found an already existing column that held a unique attribute to each different element in the dimension such as “School_roll_number” for the School dimension and “Exam_number” for the Student dimension and made this the primary key for that dimension.

As the dimensions of the schema were then set up, it was just a matter of filling in the fact table with the foreign keys and then also adding the dependent variable which would be the grade achieved in the subject.

When I realised that MySQL has a section where you can create star schemas like I had done in the website Vertabelo, I re-did the final version of my star schema in MySQL. This was easy to do as I was already familiar with the process.

The final version of my Star Schema is as follows:



3. Two OLAP Queries & the problems they solve.

A). OLAP Query #1

What my first query does is it finds out the number of times males pass more higher level subjects than females in mixed schools, for all school types, for all counties, in all years, where the socio-economic background of the school is ‘Non-DEIS’.

The reason I created this is that I wanted to find out which school types and which counties, per year, get better results for males than females. I thought that this would be a useful result as in general, females tend to do better in exams in school than males which is why I wanted to find out whether there was a correlation between males passing more higher level subjects than females and the county they live in or the type of school they attend.

For this query I restricted the results to schools where the socio-economic background is ‘Non-DEIS’; however, I could have restricted it to DEIS schools or by provence either.

The SQL code to execute this query is as follows:

```
SELECT COUNT(Grade_in_subject) AS Male_passes_Gr8r_than_female,
           IFNULL(County, 'All Counties' ) AS County, IFNULL(Year, 'ALL years') AS Year,
           IFNULL(School_type, 'All school types') AS School_type

FROM (SELECT COUNT(CASE WHEN Gender = 'Male' THEN 1 ELSE NULL END) AS
         Number_of_male_students, Subject_level, Province, Grade_in_subject,
         COUNT(CASE WHEN Gender = 'Female' THEN 1 ELSE NULL END) AS
         Number_of_female_students, School_type, County, Year, Gender_of_school

          FROM fact_table JOIN residence USING (Eir_code)
                          JOIN school     USING (School_roll_number)
                          JOIN student    USING (Exam_number)
                          JOIN subject    USING (Subject_code)
                          JOIN time       USING (Year)

          WHERE Subject_level = 'Higher' AND Gender_of_school = 'Mixed'
                   AND Socio_economic_background = 'Non-DEIS' AND
                   Grade_in_subject NOT LIKE '%7' AND Grade_in_subject NOT LIKE '%8'

          GROUP BY School_roll_number) AS Passes_in_higher_in_mixed_in_non_deis

WHERE Passes_in_higher_in_mixed_in_non_deis.Number_of_male_students >
          Passes_in_higher_in_mixed_in_non_deis.Number_of_female_students

GROUP BY County, Year, School_type WITH ROLLUP;
```

B). OLAP Query #2

I created my second query in order to find out whether the percentage of non-Irish born people who achieved grades above 60% (ie a 1,2,3 or 4) was greater than the percentage of Irish born people who achieved grades above 60%. This was done for all subjects, for all subject levels, in the year 2017.

My thoughts behind this query was that I wanted to find out in which subjects, and their corresponding subject level, Irish and non-Irish people tend to do better in.

I wanted to find out whether there is a particular set of subjects that Irish born people perform better in or are not as good at when compared with non-Irish born people.

I had to restrict the query to the year 2017 to get the final dataset to be a manageable size for this document.

This query is flexible and if required could also be used to output the percentage of Irish born people and the percentage of non-Irish born people that achieve a grade of over any particular percentage in each subject, in each level, in any year specified.

The SQL code to execute this query is as follows:

```
SELECT Year, Subject_name, Subject_level, IF((Class.Number_Irish_gr8r_than_60 = 0 AND
            Class.Number_non_Irish_gr8r_than_60 = 0), 'No grades over 60% in this class',
            IF((Class.Number_non_Irish_gr8r_than_60)/(Class.Number_non_Irish_in_class) >
            (Class.Number_Irish_gr8r_than_60) / (Class.Number_Irish_in_class) OR
            (Class.Number_Irish_in_class = 0 AND Class.Number_non_Irish_gr8r_than_60 != 0),
            'Non-Irish', IF((Class.Number_non_Irish_gr8r_than_60) /
            (Class.Number_non_Irish_in_class) = (Class.Number_Irish_gr8r_than_60) /
            (Class.Number_Irish_in_class), 'Same percentage of Non-Irish and Irish', 'Irish')))
            AS Group_w_more_grades_over_60

FROM (SELECT Subject_name, Subject_level, Year, COUNT(CASE WHEN Race LIKE
                      '%Non-Irish%' AND (Grade_in_subject LIKE '%4' OR Grade_in_subject
                      LIKE '%3' OR Grade_in_subject LIKE '%2' OR Grade_in_subject LIKE '%1')
                      THEN 1 ELSE NULL END) AS Number_non_Irish_gr8r_than_60,
                      COUNT(CASE WHEN Race NOT LIKE '%Non-Irish%' AND
                      (Grade_in_subject LIKE '%4' OR Grade_in_subject LIKE '%3' OR
                      Grade_in_subject LIKE '%2' OR Grade_in_subject LIKE '%1')THEN 1 ELSE
                      NULL END) AS Number_Irish_gr8r_than_60, COUNT(CASE WHEN Race
                      LIKE '%Non-Irish%' THEN 1 ELSE NULL END) AS
                      Number_non_Irish_in_class, COUNT(CASE WHEN Race NOT LIKE
                      '%Non-Irish%' THEN 1 ELSE NULL END) AS Number_Irish_in_class

        FROM fact_table JOIN residence USING (Eir_code)
                        JOIN school     USING (School_roll_number)
                        JOIN student    USING (Exam_number)
                        JOIN subject    USING (Subject_code)
                        JOIN time       USING (Year)

        GROUP BY year, Subject_code) AS Class

WHERE year = '2017'
```

```
GROUP BY year, Subject_name, Subject_level
ORDER BY year, Subject_name, Subject_level;
```

4. The final Datasets.

These results tables were just copied from the results tables that were generated by the SQL query.

A.) Dataset from OLAP Query #1

The result of my first query is the table below. I added empty rows between each county to make the table easier to read and understand.

Male_passes_gr8r_than_female	County	Year	School_type
1	Donegal	2018	Vocational School
1	Donegal	2018	All school types
1	Donegal	All years	All school types
1	Dublin	2017	Vocational School
1	Dublin	2017	All school types
2	Dublin	2018	Voluntary School
2	Dublin	2018	All school types
3	Dublin	All years	All school types
1	Limerick	2013	Vocational School
1	Limerick	2013	All school types
1	Limerick	2014	Community School
1	Limerick	2014	Voluntary School
2	Limerick	2014	All school types
3	Limerick	All years	All school types
2	Westmeath	2016	Community School
2	Westmeath	2016	All school types
2	Westmeath	All years	All school types
9	All Counties	All years	All school types

B.) Dataset from OLAP Query #2

The result of my second query is the table below. This is every subject in the year 2017 and the corresponding group that did better in that subject. Each subject has two levels and 3 subjects have 3 levels.

Year	Subject_name	Subject_level	Group_w_more_grades_over_60
2017	Accounting	Higher	Non-Irish
2017	Accounting	Ordinary	Irish
2017	Applied Maths	Higher	Non-Irish
2017	Applied Maths	Ordinary	Irish
2017	Art	Higher	Non-Irish
2017	Art	Ordinary	Irish
2017	Biology	Higher	Irish
2017	Biology	Ordinary	Non-Irish
2017	Business	Higher	Irish
2017	Business	Ordinary	Irish
2017	Chemistry	Higher	Non-Irish
2017	Chemistry	Ordinary	Irish
2017	Construction	Higher	Irish
2017	Construction	Ordinary	Non-Irish
2017	Design & Communication Graphics	Higher	Non-Irish
2017	Design & Communication Graphics	Ordinary	Non-Irish
2017	Economics	Higher	Irish
2017	Economics	Ordinary	No grades over 60% in this class
2017	Engineering	Higher	Irish
2017	Engineering	Ordinary	No grades over 60% in this class
2017	English	Foundation	Same percentage of Non-Irish and Irish
2017	English	Higher	No grades over 60% in this class
2017	English	Ordinary	Irish
2017	Geography	Higher	Same percentage of Non-Irish and Irish
2017	Geography	Ordinary	No grades over 60% in this class

2017	German	Higher	Irish
2017	German	Ordinary	Irish
2017	History	Higher	No grades over 60% in this class
2017	History	Ordinary	Irish
2017	Home Economics	Higher	Irish
2017	Home Economics	Ordinary	Irish
2017	Irish	Foundation	Irish
2017	Irish	Higher	No grades over 60% in this class
2017	Irish	Ordinary	Irish
2017	L.C.V.P	Higher	Irish
2017	L.C.V.P	Ordinary	Irish
2017	Maths	Foundation	Irish
2017	Maths	Higher	No grades over 60% in this class
2017	Maths	Ordinary	Irish
2017	Physics	Higher	No grades over 60% in this class
2017	Physics	Ordinary	Non-Irish