**Advanced Databases CA**

|  |  |
| --- | --- |
| Student Name: Gavin O’Neill | Student ID: D15127205 |
| Current Date: 16/12/2016 | Subject Area: Advanced Databases |

*The original aim of the process was to enable Fingal County Council to look at enrolment numbers in schools and decide if there were then enough amenities such as parks, pitches, libraries, etc. for children to enjoy. Due to problems with the enrolment number data, I have instead concentrated on membership numbers in the libraries, specifically how they change from year to year.*

I. ASSOCIATED FILES

The project consists of six data files obtained from the Dublinked website and five program files.

**Data files:**

1. Dublinlocalauthorities.csv
2. Fcclibraries.csv
3. Fccparks.csv
4. Fccplayingpitches.csv
5. Fccschools.csv
6. Librarymembers.csv

**Program files:**

1. BulkInsert.sql
2. DataConversion.sql
3. ERtables.sql
4. StarSchema.sql
5. Queries.sql

*Dublinlocalauthorities.csv* is a file that I created to have a common attribute for the different entities in the database. The schools, libraries, parks, etc. are all located in Fingal so will be connected through this attribute.

|  |  |
| --- | --- |
| **Column** | **Description** |
| LA\_ID | This is an identifier for each local authority |
| Name | Name of the local authority |

*Fccparks.csv* gives a lot of information on all parks in the Fingal area. The table below describes the information provided:

|  |  |
| --- | --- |
| **Column** | **Description** |
| ParkID | Identifier for each park |
| Name | Name of the park |
| Address1, 2, 3 | Each line of the park’s address |
| Phone | Phone number |
| Email | Email address |
| Website | Website address |
| Type | Whether regional or local park |
| Parking | Parking spaces available or not |
| Bicycle\_Parking | Bike parking available or not |
| Toilets | Toilets available or not |
| LA\_ID | Gives identifier for local authority in charge |

*Fcclibraries.csv* contains information on all of the libraries in the Fingal area. The table below provides a description of the data it holds:

|  |  |
| --- | --- |
| **Column** | **Description** |
| Library\_ID | Identifier for each library |
| Name | Name of the library |
| Address1, 2, 3 | Each line of the library’s address |
| Phone | Phone number |
| Email | Email address |
| Website | Website address |
| Internet | Internet available or not |
| Wifi | Wifi available or not |
| Selfservice | Selfservice or not |
| LA\_ID | Gives identifier for local authority in charge |

*Fccplayingpitches.csv* contains data on various playing pitches in the Fingal area. The table below gives a description of the data:

|  |  |
| --- | --- |
| **Column** | **Description** |
| Facility\_ID | Identifier for facility |
| Facility\_Type | Type of facility |
| Facility\_Name | Name of the facility |
| Area | Area where it’s located |
| LA\_ID | Gives identifier for local authority in charge |

*Fccschools.csv* contains data on all school in the Fingal region. The table below describes the data included in the file:

|  |  |
| --- | --- |
| **Column** | **Description** |
| School\_Roll\_No | Identifier for each school |
| Name | Name of the school |
| Address1, 2, 3 | Each line of the school’s address |
| Phone | Phone number |
| School\_Level | Primary or secondary |
| Mixed\_Status | Boy/Girl school |
| Fee\_paying | Fees or not |
| LA\_ID | Gives identifier for local authority in charge |

*Librarymembers.csv* contains a list of the membership figures for each library in the Fingal region. The figures span a few years and are described below:

|  |  |
| --- | --- |
| **Column** | **Description** |
| Library\_ID | Identifier for each library |
| Year | Year of membership information |
| Members | Total membership number |
| New | New members this year |
| Lost | Members lost this year |

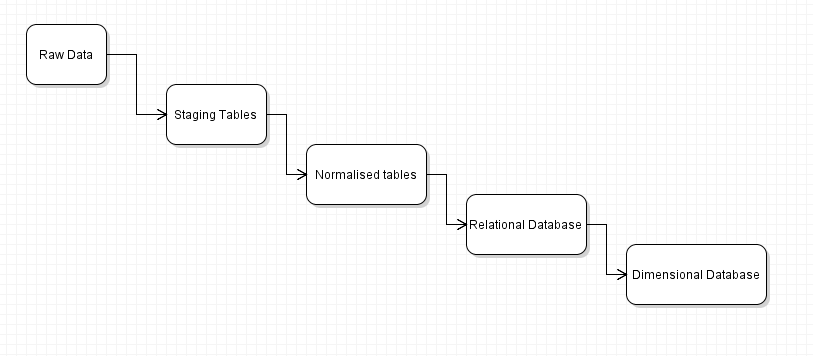
II. Structure of the Processing Pipeline

Figure 1 below shows the processing pipeline followed in this project. There are four main stages

involved: raw data is uploaded from the csv files to staging tables, cleaned and normalised, a relational

database is created for the data and finally it’s loaded into a dimensional model. Stored procedured are used

at each stage of this process and I will discuss this more below.



Figure

In the first stage the data is uploaded in bulk from the csv files into staging tables with no constraints and error files are used to gather information on errors as required. I have used just one stored procedure here, which is called repeatedly for each table.

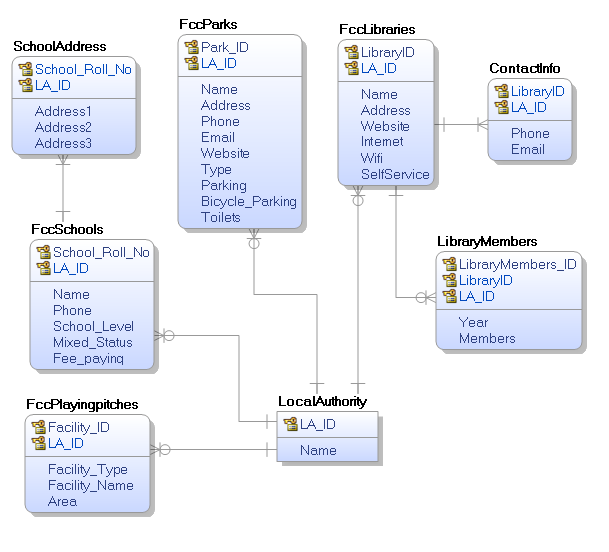
Stage two involves moving the data into normalised tables using primary key, foreign key and not null constraints. A separate stored procedure is used for each table in this stage.

Stage three involves moving the normalised data into a new relational database based on an ER model (see next section). As with the previous stage a separate stored procedure is used for each entity in the database.

Stage four is the final stage in the process and involves moving data from the relational database into a new model based on the Star Schema (see below for model). This dimensional database will allow us to carry out analytical queries on the data. Stored procedures are used for each table again.

III. ENTITY RELATIONSHIP MODEL

Figure 2 below shows an entity relationship diagram for the relational database. There are 8 entities in total, with two of the 6 csv files being broken into two entities to make that total.



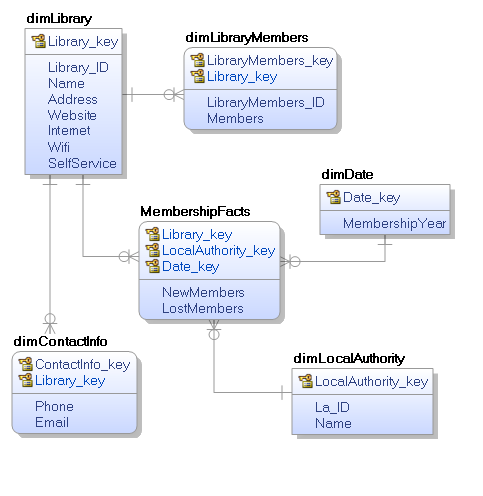
Figure

The LocalAuthority entity has one to many relationships with the FccSchools, FccParks, FccPlayingpitches and FccLibraries entities. These are identifying relationships. A local authority can be in charge of many of these amenities but an amenity can only have one local authority in charge of it.

FccLibraries has one to many, identifying relationships with ContactInfo and LibraryMembers. and the same goes for FccSchools.

IV. DIMENSIONAL MODEL

Figure 3 displays the star schema of the dimensional model. (Update: the actual dimension database is slightly different to this as it was not functional and the fact table is not working as it should)



Figure

V. FINDINGS OF DATA QUERY

There were issues with the fact table so queries couldn’t be done with it, only via the other dimension tables. The surrogate keys are not being referenced in the fact table so all of its values are NULL and unfortunately a solution could not be found. There are some queries in the queries.sql file that give some insights but not relating to the changing library membership numbers as hoped.

VI. REFLECTION

One of the most difficult parts of this assignment was finding good datasets to use from the Dublinked website. It’s more than likely due to the fact that I didn’t do a project like this before and so maybe I wasn’t sure exactly how to tell a good, usable dataset from the bad.

I found the work to be positive overall and definitely a good learning experience. Starting a project from scratch like this and researching all the errors is definitely the best way to get your head around every type of situation that can occur.

I still haven’t figured out how to make a perfectly working process from start to finish as there are issues with the fact table so hopefully this will be explained with any feedback and will allow me to improve in the future.

VIII. CONCLUSIONS

As mentioned previously, the fact table didn’t function as expected so no analytical queries of note could be made.

IX. README

Each sql file must be read in order going BulkInsert, DataConversion, ERtables, StarSchema. If there are errors running the files at first, there are some commented drop functions that can be used to get around it.

File paths may need to be edited at the BulkInsert.sql stage as these are set to my own working directory as they are.