COMP 4522 Winter 2018 - Assignment 1
Data Warehousing - Part I

Due: Feb 27, 2018 before the end of the day (11:59pm)

Weight: 5%

You can do this assignment individually, or as a pair or as a trio.

In this assignment you will build a data warehouse using dimensional modeling (create a star schema). In Part II, you will then perform some data mining to extract "knowledge" from the data warehouse data.

Scenario

The dataset you will be using contains over 80,000 reports of UFO sightings over the last century. The Kaggle link is: https://www.kaggle.com/NUFORC/ufo-sightings. You will need a login to use all the contents of the website.

The column metadata has the following attributes:
`dateTime, city, state, country, shape, duration(seconds), duration(hours/min), comments, date posted, latitude and longitude of sighting`.

Study the data to understand it. You can also preview the first 100 rows of the data on the website.

The data is loaded on to a single table `ufo` on schema `ufodb` on an Oracle database for your use. In order to complete this assignment, you will need to fully understand the data in the table. You may also need to clean/scrub the data after finding out what is in the table, how it is formatted and what the columns mean. Do not include your investigative queries in your submission.

Create a Dimensional Model

Remember that there are four steps in dimensional modeling:

1. Choose the business process.
2. Determine the grain to use.
3. Identify the dimensions.
4. Identify the facts.

In this case, the business process is just the UFO sightings.

People record an UFO sighting, its location (city, state, country), the date and time of sighting, its shape, its duration (in seconds), and its latitude and longitude.

You will design the DW based on the types of queries you will need to write.

When building the dimensions, try to imagine all of the different ways the data could be aggregated, sorted or queried.

For example, one of the dimensions will be the Date. We may want to query by year, month, day of the week, season, zodiac sign. Another one would be to query based on short duration events, medium duration ones and the lengthy ones. We would like to query based on continent (for example, Europe or North America). You can think of other dimensions and other interesting queries you may want to write.

If anything is ambiguous or otherwise not clear, we are your business people, so you can ask us for clarification.

Populate the Data Warehouse / Extract Transform Load

In the data warehousing world, we speak of the Extract, Transform, Load (ETL) step. This is a step that takes place periodically and takes data from the operational systems, transforms it, and loads it into the data warehouse tables. The transform step is where the data is cleaned, if needed, where it is standardized and aggregated. The ETL process is often messy and difficult as the operational data is coming from multiple disparate systems that may have omissions and errors, different codes and data meanings. ETL takes place periodically, at whatever intervals - weekly, monthly, quarterly - are all appropriate for the application at hand.

For our system, we do not have any operational systems to draw from but we do some have some slightly messy source data table for you to work with.The `ufo` table stores data is a set of columns which are just strings. You will have to use functions to convert them into the appropriate data types.

When populating the data warehouse, you have two tasks: populate the dimension tables, and populate the fact tables.

Remember that each part of the key in the fact table is a foreign key to a dimension table, so you need to have the dimension tables populated first (else you will get a referential integrity error).

For all of the steps,, if something is ambiguous or otherwise unclear, ask us about it.

Deliverables

1. Create a design diagram for the data warehouse. You should use a tool (Gliffy, draw.io, MS Visio) to draw the diagram. Submit a PDF of this diagram. The diagram should fit in a one A4 size page.

2. Create a single self documenting Oracle SQL script called `A1.sql` to create the data warehouse.

Make note of the following:

- The script should be well documented. Each table should include a comment block that describes the table, what kind of table it is, what sort of facts it records (if it is a fact table), the grain, what design decisions you made while creating the table.
- From the descriptions above, you should be able to populate the dimension tables. Since you know that the data is available from January 1, 1910 to September 9, 2013, you may as well populate the date dimension table first. I would suggest inserting dates from January 1, 1910 to some date in the future, like December 31, 2020. Since this is a large number of entries (40,542 days to be exact), you will not want to simply do a series of inserts. That will suffice for some of the other tables, but you will want to write a PL/SQL loop for this one. You will need to populate the quarter, season and other fields at the same time, which will require that you use Oracle Date function, or even write some of your own functions.
- Now, populate your fact tables.
- Your script should create all primary and foreign keys.
- Appropriate indexes should be created.
- Submit your script to the Blackboard drop box provided.