University of Ottawa
School of Electrical Engineering and Computer Science
CSI4142 Introduction to Data Science

Assignment 1 2018

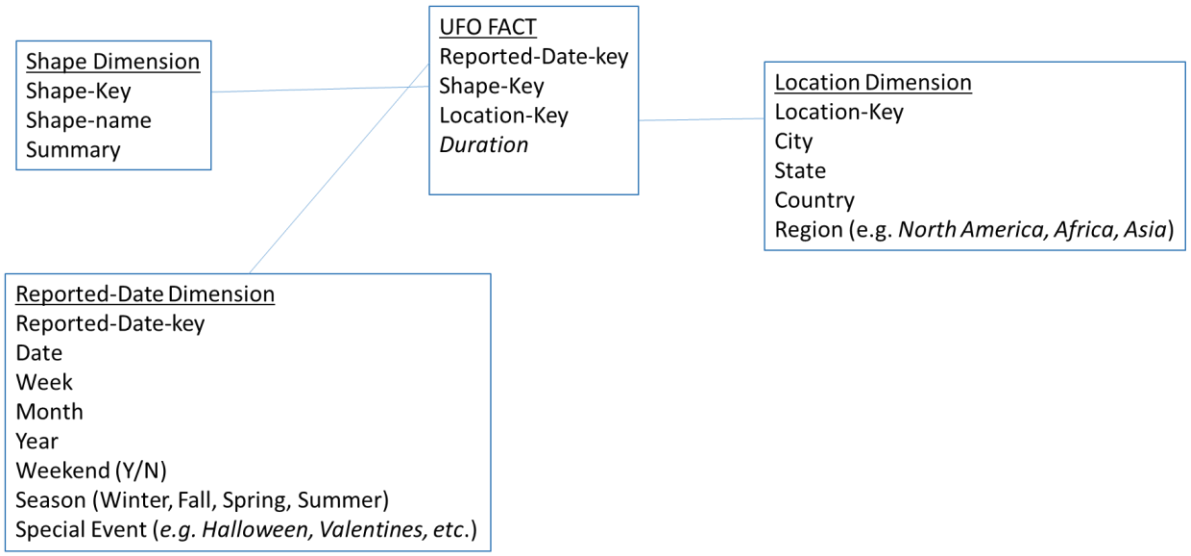
Total marks: 50

This is an individual assignment. Submit your answer using the Virtual Campus.

The National UFO reporting center (http://www.nuforc.org/index.html) maintains information about unidentified flying objects (UFOs) as reported since 1400. The database was first created in 1998 and may be accessed at this link: http://www.nuforc.org/webreports.html

The data include information such as the reported shape(s), the date of observation and reporting, the duration, the location and a short textual description of the UFOs, amongst others. The CEO of the National UFO reporting center asks you to build a data analytics application in order to explore this data.

You are given the task to create a data mart for the National UFO reporting data, in order to convert the data into a format that facilitates Online Analytical Processing (OLAP) queries. The following is an initial dimensional model that is submitted for your review.



Answer all the following questions.

1. Declare the grain of this dimensional model. (2)
2. There is a dimension missing from this model. Provide this dimension and motivate your answer. (4)
3. Define what a measure (or fact) is and suggest one more measure that may be added to this design. (2)
4. Explain what a concept hierarchy is and provide an example of a concept hierarchy in the UFO data mart. (2)
5. Aggregation is one way that may be used to speed up OLAP queries against a data mart. Give an examples of an aggregate that you would potentially create against the UFO dimensional model. (4)
6. The data contain some missing values. Give an example of an attribute with missing values and explain how you would handle such data during the analytical process. (2)
7. Draw a scatter plot in order to determine whether there is a correlation between the time of occurrence and the duration. (4)
8. Consider a query that returns the names of the States that reported the eight viewings of DELTA-based shapes. Show how you would use bitmap indexes to answer this query. (2)
9. Consider a query that returns the total number of sightings of DIAMOND shapes in Florida, in 2017. Show how you would use semi-joins and bitmaps to optimize this query. (4)
10. Create the UFO data mart in PostgreSQL and insert all the data into your tables. (You will notice that some of the durations are uncertain. In this case, you should record the minimum value. That is, a duration of 10-15 seconds should be converted to 10 seconds.) (6)

11. Provide the SQL statement to list the number of sightings by Month. (2)
12. Provide the SQL statement to list the number of sightings by Month and State. (4)
13. Provide the SQL statement to list the names and average durations of the five shapes that have the most sightings, overall. (4)
14. Provide the SQL statements to list, for each shape, the State with the longest recorded duration of sighting, together with the average duration for each shape. *For example, for the CRESCENT shape, the average duration was 47.5 seconds, the longest duration was 90 seconds and this occurred in WI (Wisconsin).* (4)
15. Provide the SQL statement to list the total the number of sightings over the weekends, for every different shape, as recorded in California versus Florida. (4)