CMPS 566 **Assignment #2**, Spring 2021
Dr. Vijay Raghavan
Assigned: March 15, 2011
Due: March 29, 2021 by 11:55 pm
Total Marks: 100

(Note that any sort of cheating will **NOT** be tolerated. More information can be found on class Web page on cheating policy)

Q1. [60 points] Implement a star schema Data Warehouse for the hurricane data (data and the helper document with data format is attached, the columns you need are highlighted in the data format file. You can download the assignment and data from this link: https://drive.google.com/drive/folders/1di8nGgUkPQQTBtT4LZpXzs5LcAk66vnP?usp=sharing). Import the above downloaded data into your database. The measure can be count (quantity of hurricanes). All dimensions should be hierarchical. The following dimensional tables are recommended, as reasonable options, for the DW design:

- Time dimension for Hurricane-Start-Date
- Time dimension for Hurricane-End-Date
- Time dimension for Hurricane-Peak-Status-Date
- Dimension for Event-Peak-State
- Dimension for Latitude
- Dimension for Longitude
-

The concept hierarchy for Latitude and Longitude dimensions should be based on creating bins, consisting of different interval ranges, starting from all the distinct values of the Lat-Lon values that actually occur in the data.

A hurricane's Hurricane-Peak-Status-Date is the date on which it first reached its maximum winds speed or minimum pressure value. The Lat-Lon values associated with the hurricane is the location at which that peak status was experienced. Thus, exactly one Lat-Lon pair of values represents each individual event.

The Event-Peak-State dimension will have all possible types of events in its dimension. The possible event types include:
TD – Tropical cyclone of tropical depression intensity (< 34 knots)
TS – Tropical cyclone of tropical storm intensity (34-63 knots)
HU – Tropical cyclone of hurricane intensity (> 64 knots) EX – Extratropical cyclone (of any intensity)
SD – Subtropical cyclone of subtropical depression intensity (< 34 knots)
SS – Subtropical cyclone of subtropical storm intensity (> 34 knots)
LO – A low that is neither a tropical cyclone, a subtropical cyclone, nor an extratropical cyclone (of any intensity)
WV – Tropical Wave (of any intensity)
DB – Disturbance (of any intensity)

The second level groups of members of this dimension in the concept hierarchy can be Non-hurricane, Medium and Severe, where Non-hurricane group is {TD, SD, LO, WV, DB}, Medium group is {SS, TS} and the Severe group is {HU}. Further categorization of HUs are done using maximum wind speeds, as follows:

Wind speed (miles/hour)	Category
74 - 95	1
96 - 110	2
111 - 129	3
130 - 156	4
157	5

The following are two example events and the various attributes associated with those two events:

Hurricane Gordon
Hurricane-Start-Date = 09/14/2000; 15Z
Hurricane-End-Date = 09/18/2000; 15Z
Hurricane-Peak-Status-Date = 09/17/2000; 9Z
Event-Peak-State = "HU-1"
Latitude = 26.90
Longitude = -84.70

Hurricane Bret
Hurricane-Start-Date = 08/18/1999; 21Z
Hurricane-End-Date = 08/23/1999; 21Z
Hurricane-Peak-Status-Date = 08/22/1999; 13Z
Event-Peak-State = "HU-4"
Latitude = 26.30
Longitude = -96.20

The Fact table will have a key attribute value for each dimension and just a single measure that needs to be monitored, which is called COUNT. Within the base cuboid all these counts will have the initial value of "1". There is an option in DW context whereby the design can use what is known as a 'Factless Fact Table'. In such a case the Fact table has no measures (i.e. just the key attributes). In the former design, you will use SQL Sum ( ) function over the COUNT attribute to get aggregated values for the number of hurricanes associated with cuboids at various aggregation levels. In the latter design (Factless Fact table), you can simply use the SQL Count ( ) function to compute the number of hurricanes associated with tuples in an aggregate level cuboid. Note that the latter design saves space in the base fact table, since you don't have to store the initial COUNT attribute values of "1" for each base tuple. However, the COUNT attribute will be needed, if the designer decides to materialize some or all of the aggregate level cuboids.

Implement a program that allows you to do the following:

a) Roll-up on count by each dimension
b) Drill-down on count by each dimension

Have a simple user interface to allow activation of the above functions. As a minimum, your interface should allow the specification of a particular aggregate level cuboid, specify layout requirements of the display and the submission of a request for counts associated with each tuple of that cuboid to be displayed. In addition, with respect to the cuboid currently displayed, the user should able to select any one dimension and indicate, for the selected dimension, the concept hierarchy level to which either a drill-down or a roll-up operation should be performed, specify layout requirements of the display and be able to obtain the counts associated with each tuple of the newly specified cuboid.
Hints:

--Use SQL statements to implement each roll-up, drill-down.
--All the calculation of the response of each OLAP operation (SQL statements) should be done from the base cuboid.

--See Example 3.7 in 2nd edition of the text book for reference.
-- The hurricane data is from Atlantic basin (AL). You need to extract columns specific to hurricane date and time, longitude and latitude, maximum wind speed and minimum pressure value those fields are highlighted in the format file.

You can implement your design using Python, C++ or Java.

Deliverables:
--The schema for the data warehouse.
--Source code.
--Demo.

Q2. [20 points] Suppose that a data warehouse for U.S. Traffic Ticket System consists of the following four dimensions: *driver, violation, location,* and *date,* and two measures count and amount (fine amount). For *location,* the concept hierarchy involves "street < city < state" and for *date,* the hierarchy consists of "date < month < quarter < year." Design the hierarchies (each having exactly two levels) for the *driver* and *violation* dimensions by yourself.

a) Draw a snowflake schema diagram for the data warehouse.

b) Starting with the base cuboid [*driver, violation, location,* and *date*], what specific OLAP operations (e.g., roll-up from street to city) should one perform in order to list the total amount of fines for the year of 2010 in Lafayette Louisiana.

c) How many cuboids will this cube contain (including the base and apex cuboids)?

d) Suppose there are four cuboids (including the Base cuboid) materialized:

cuboid 1: {year, street}
cuboid 2: {year, city}
cuboid 3: {state} where year =2010

Which of the above cuboids would you select for the query in part b)? Explain your reasons.

Q3. [20 points] Suppose that the following table is derived by attribute-oriented induction.

Class	Place	Year	Count
Programmer	USA	2000	180
		2001	200
.	. Others	2000	120
.		2001	150
.	.		
DBA	USA	2000	20
.	.	2001	20
.	Others	2000	80
.	.	2001	85

a) Transform the table into a crosstab showing the associated t-weights and d-weights.

b) Map the class Programmer into a (bidirectional) quantitative descriptive rule with attribute place,
For Example, For_All (X), Programmer(X) ↔ (place(X) = “USA”)
[t : x%, d : y%]…∨ (…)[t : w%, d : z%].

c) Map the class DBA into a (bidirectional) quantitative descriptive rule with attributes place and year,
For Example, For_All (X), DBA(X) ↔ (place(X) = “USA”∧ year(X)=”2000”)
[t : x%, d : y%]…∨ (…)[t : w%, d : z%].