Problem Solving
Assignemnt-3

1. Suppose that a data warehouse consists of the three dimensions' time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit.

   a). Draw a schema diagram for the above data warehouse using one of the schemas. [star, snowflake, fact constellation]
   b). Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2004?
   c). To obtain the same list, write an SQL query assuming the data are stored in a relational database with the schema fee (day, month, year, doctor, hospital, patient, count, charge)
2. Suppose that a data warehouse for Big-University consists of the following four dimensions: student, course, semester, and instructor, and two measures count and avg_grade. When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the avg_grade measure stores the actual course grade of the student. At higher conceptual levels, avg_grade stores the average grade for the given combination.

   a). Draw a snowflake schema diagram for the data warehouse
   b). Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year) should one perform in order to list the average grade of CS courses for each BigUniversity student

3. Suppose that a data warehouse consists of the four dimensions, date, spectator, location, and game, and the two measures, count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate.

   Draw a star schema diagram for the data warehouse.

4. Design a data warehouse for a regional weather bureau. The weather bureau has about 1,000 probes, which are scattered throughout various land and ocean locations in the region to collect basic weather data, including air pressure, temperature, and precipitation at each hour. All data are sent to the central station, which has collected such data for over 10 years. Your design should facilitate efficient querying and online analytical processing, and derive general weather patterns in multidimensional space.
5. Suppose a company would like to design a data warehouse to facilitate the analysis of moving vehicles in an online analytical processing manner. The company registers huge amounts of auto movement data in the format of (Auto ID, location, speed, time). Each Auto ID represents a vehicle associated with information, such as vehicle category, driver category, etc., and each location may be associated with a street in a city. Assume that a street map is available for the city.

   a) Design such a data warehouse to facilitate effective online analytical processing in multidimensional space.

   (b) The movement data may contain noise. Discuss how you would develop a method to automatically discover data records that were likely erroneously registered in the data repository.

(c) The movement data may be sparse. Discuss how you would develop a method that constructs a reliable data warehouse despite the sparsity of data.

(d) If one wants to drive from A to B starting at a particular time, discuss how a system may use the data in this warehouse to work out a fast route for the driver.

6. A data cube, C, has n dimensions, and each dimension has exactly p distinct values in the base cuboid. Assume that there are no concept hierarchies associated with the dimensions.

   (a) What is the maximum number of cells possible in the base cuboid?

   (b) What is the minimum number of cells possible in the base cuboid?

   (c) What is the maximum number of cells possible (including both base cells and aggregate cells) in the data cube, C?

   (d) What is the minimum number of cells possible in the data cube, C?

7 RFID (Radio-frequency identification) is commonly used to trace object movement and perform inventory control. An RFID reader can successfully read an RFID tag from a limited distance at any scheduled time. Suppose a company would like to design a data warehouse to facilitate the analysis of objects with RFID tags in an online analytical processing manner. The company registers huge amounts of RFID data in the format of (RFID, at location, time), and also has some information about the objects carrying the RFID tag, e.g., (RFID, product name, product category, producer, date produced, price).

   (a) Design a data warehouse to facilitate effective registration and online analytical processing of such data.

   (b) The RFID data may contain lots of redundant information. Discuss a method that maximally reduces redundancy during data registration in the RFID data warehouse.

   (c) The RFID data may contain lots of noise, such as missing registration and misreading of IDs. Discuss a method that effectively cleans up the noisy data in the RFID data warehouse.

   (d) One may like to perform online analytical processing to determine how many TV sets were shipped from the LA seaport to BestBuy in Champaign, Illinois by month, by brand, and by price range. Outline how this can be done efficiently if you were to store such RFID data in the warehouse.

   (e) If a customer returns a jug of milk and complains that is has spoiled before its expiration data, discuss how you could investigate such a case in the warehouse to find out what could be the problem, either in shipping or in storage.