Data Warehousing & Data Mining

Mid I Exam (Section A and B) Fall 2016

Date: September 19, 2015 Marks: 25 Time: 60 min.

Name: ----------

Roll No: ---------- Section: ----------

Q1. *(4 points)*
Consider the following normalized data structure:

Student (Roll No, Name, Address, Batch)
Course (Course Code, Title, CrHrs, School)
Grade (Roll No, Course Code, GPA, Letter Grade)

Assume there are 1000 students, 100 courses, and 40,000 grades. Each column size is 10 bytes. Suppose a query requires data from all these three tables frequently. You are required to improve the performance of the query using pre-join de-normalization technique. Show your de-normalized data structure and evaluate storage cost (in MB) for the normalized and de-normalized data structure respectively.

Ans:
Storage cost of normalized structure: 1.644MB (40,000+4000+1,600,000= 1,644,000)
Storage cost of de-normalized structure: 4MB (100 x 40,000= 4,000,000)

Q2. *(2 points)*
What is the concept of virtual cube? When will you consider to forming virtual cube?

Ans: Similar to a relational view; two (or more) cubes are linked along common dimension(s).
Virtual cubes are used when there is a need to join information from two dissimilar cubes that share one or more common dimensions. Often used to save space by eliminating redundant storage of information.

Q3. *(2 points)*
What is meant by slice-and-dice? Give an example.

Q4. *(8 points)*
In a University case study, we have following dimensions:

Term Year (Term Key, Term Description, Academic Year, Season)
Student (Student Key, Student Id (Production/Natural Key), Student Description, Student Batch)
Course (Course Key, Course Name, Course Credit hours, Course School)
Faculty (Faculty Key, Faculty Employee Id (Production/Natural Key), Faculty Name, Faculty School)

Facts in a registration fact table are RegistrationCount (always = 1), GPA and letter grade. The grain of the fact table is one row for each registered course by student and term. Fact table has the following dimensionality: term, student, course, and faculty.

Following queries are also made most frequently:

Query#1: Average GPA by term by batch by course school by faculty
Query#2: Total number of registered students by academic year by course school

Draw a star schema that includes registration base fact table and aggregate fact tables for the above requirements. Take appropriate assumption, if required. Show the primary keys, foreign keys and all the relationships between the dimensions and fact tables.

Q5. *(2 points)*
Refer to the faculty dimension of above star schema. Suppose University has1000 faculty members and 50 faculty schools. How many rows reside in physical data model (for faculty dimension only), if we use Star Schema and Snow Flake Schema respectively?

Ans: 1000 & 1050

Q6. *(2 points)*
Estimate the size (in number of rows) of aggregate fact table for above query#2. Assume 45 terms, 15 academic years, 200,000 students, 25 batches, 5000 courses, 50 schools, and 1000 faculty members.

Ans: 15 x 50 = 750 rows

Q7. *(3 points)*
Identify the full-additive, semi-additive, and non-additive facts, if any, in the above registration base fact table.

Ans: registration count (fill-additive), GPA & letter grade (non-additive)

Q8. *(2 points)*
Refer to the student dimension of above star schema. Show the revised student dimension schema that also preserves the history of changes to the student.

Ans: only add effective date attributes (i.e. start date and end date). Surrogate key is already here.