National University of Computer and Emerging Sciences, Lahore Campus

Course Name:	Data Warehousing & Data Mining	Course Code:	CS409
Program:	BS(CS)	Semester:	Fall 2018
Duration:	60 Minutes	Total Marks:	25
Paper Date:	Wed 03-Oct-2018	Weight	12.5%
Section:	CS	Page(s):	3
Exam Type:	Midterm 1		

Student : Name:_____ Roll No._____
Section: CS

Instructions/Notes:
1. Scratch sheet can be used for rough work however, all the questions and steps are to be shown on question paper. No extra/rough sheets should be submitted with question paper.
2. You will not get any credit if you do not show proper working, reasoning and steps as asked in question statements.

Q1. *(8 points)* Give the appropriate answers of the following questions very briefly:

a. What are essential differences between the dependent data mart and independent data mart?
b. What kind of situations are there where you might want to use degenerated dimensions (DD)? Give an example of DD.
c. What are the various data sources for the data warehouse?
d. What is the concept of factless fact table? Give an example.

Ans:

a. In the top-down approach of dw implementation, it represents a centralized repository along with a set of dependent data marts. These data marts provide a single version of the truth.
In the bottom-up approach of dw implementation, it may represent a set of independent data marts. These data marts are likely to have inconsistent data definitions and standards. Such variances hinder analysis of data across data marts.

b. A degenerate dimension (DD) acts as a dimension key in the fact table, however does not join to a corresponding dimension table because all its interesting attributes have already been placed in other analytic dimensions. Sometimes people want to refer to degenerate dimensions as textual facts, however they're not facts since the fact table's primary key often consists of the DD combined with one or more additional dimension foreign keys.
Degenerate dimensions commonly occur when the fact table's grain is a single transaction (or transaction line). Transaction control header numbers assigned by the operational business process are typically degenerate dimensions, such as order, ticket, credit card transaction, or check numbers. These degenerate dimensions are natural keys of the "parents" of the line items.

c. Production data, archived data, internal data, external data

d. A factless fact table is a fact table that does not have any measures. It is essentially an intersection of dimensions (it contains nothing but dimensional keys). There are two types of factless tables: One is for capturing an event, and one is for describing conditions. Example: Student Attendance factless fact table with dimensions: Student, Course, Time, Semester, Teacher.

Consider the following University case study for next questions:

Dimensions are
Semester (Semester Key, Semester Description, AcademicYear)
Course (Course Key, Course Description, Offering School)
Student (Student Key, Student ID (Production/Natural Key), Student Description, Batch)

Assume: 60 Semesters, 20 academic years, 200 courses, 10 schools, 500,000 students, and 25 batches.

Facts in a registration fact table are GPA, LetterGrade and RegistrationCount (always=1). The grain of the fact table is one row for each registered course by student in a semester. Fact table has the following dimensionality: semester, course, and student.
Following queries are also made most frequently:

Query 1. Average GPA by offering school by batch by semester.
Query 2. Total number of registered students by semester by offering school.

Q2. *(9 points)*
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.

Ans: Do-your-self.

Q3. *(3 points)*
Identify the full-additive, semi-additive, and non-additive facts, if any, in the above registration base fact table.

Ans:
Fully-additive: RegistrationCount
Non-additive: GPA and LetterGrade

Q4. *(3 points)*
Estimate the size (in number of rows) of the above base fact table and aggregate fact tables.

Ans:
Base Fact Table: 6billions (6,000,000,000) rows
Aggregate Fact Table1: 10 (Offering Schools) x 25 (Batches) x 60 (Semesters) = 15,000 rows
Aggregate Fact Table2: 60 (Semesters) x 10 (Offering Schools) = 600 rows

Q5. *(2 points)*
Refer to the student dimension of above star schema. How many rows reside in physical data model (for student dimension only), if we use Star schema and Snow Flake schema respectively?

Ans:
Star Schema: 500,000
Snow Flake Schema: 500,000 + 25 = 500,025