EE4791 Database Systems

1. Consider a disk with block size B=512 bytes. A file has r=30,000 EMPLOYEE records of fixed-length. Each record has the following fields: NAME (30 bytes), SSN (9 bytes), DEPARTMENTCODE (9 bytes), ADDRESS (40 bytes), PHONE (9 bytes), BIRTHDATE (8 bytes), SEX (1 byte), JOBCODE (4 bytes), SALARY (4 bytes, real number). An additional byte is used as a deletion marker.

   (a) Calculate the record size R in bytes.
   (b) Calculate: (i) the blocking factor bfr and the number of file blocks b assuming an unspanned organization; (ii) the total disk space needed (no of bytes).
   (c) If the file is unordered and each record has a unique value of SSN, calculate the average number of block accesses needed for searching a record based on SSN value.
   (d) If the file is ordered by its key field SSN and no indexes are defined for the file, calculate the average number of block accesses needed for searching a record based on SSN value.

2. For the same file described in question one, assuming that the file is ordered by the key field SSN and a primary index on SSN is defined for the file. Suppose that the disk's block size B is still 512 bytes, and a block pointer is P=6 bytes long. Calculate:

   (a) The index blocking factor (fan-out) bfr_i for the primary index.
   (b) The number of first-level index entries and number of first-level index blocks.
   (c) The number of levels needed if we make it into a multi-level index.
   (d) The total number of blocks required by all the index entries in the multilevel index structure.
   (e) The number of block accesses needed to search for and retrieve a record from the file, given its SSN value, using the primary index if it is implemented as:
      (i) single-level index structure
      (ii) multilevel index structure
   (f) The percentage of improvement on number of block accesses achieved by:
      (i) 1(d) from 1 (c), and
      (ii) Implementing the above-mentioned primary index as a single-level index from 1(d)
      (iii) Implementing the above-mentioned primary index as multilevel index from implementing as a single-level index

3. Suppose the file is not ordered by the key field SSN and we want to construct a secondary index on SSN. Repeat the calculation from 2(a) to 2(e) for the secondary index.

4. For the same file described in question one, suppose the file is not ordered by the non-key field DEPARTMENTCODE and we want to construct a secondary index on DEPARTMENTCODE with one level of indirection that stores record pointers. Assume there are 1000 distinct values of DEPARTMENTCODE and that the EMPLOYEE records are evenly distributed among these values. Suppose that the disk's block size B is still 512 bytes, a block pointer is P=6 bytes and a record pointer is P_R =7 bytes long. Calculate:

   (a) The index blocking factor $bfr_{i.}$
   (b) The number of blocks needed by the level of indirection that stores record pointers.
   (c) The number of first-level index entries and the number of first-level index blocks.
   (d) The approximate number of block accesses needed to search for and retrieve all records in the file having a specific DEPARTMENTCODE value using the secondary index implemented in single-level index structure.
   (e) The number of levels needed if we make it into a multi-level index.
   (f) The total number of blocks required by all the index entries including the blocks required by the level of indirection level in the multilevel index structure.
   (g) The number of block accesses needed to search for and retrieve a record from the file, given its DEPARTMENTCODE value, using the secondary index implemented as multilevel index structure with the one level of indirection.

5. For the same file described in question one with block size B = 512, block pointer size P = 6 bytes and record pointer size Pr = 7 bytes, assuming that the file is not ordered by the key field SSN and we want to construct a B+-tree access structure (index) on SSN. Calculate:

   (a) The orders p and p_{leaf} of the B+-tree.
   (b) The number of leaf-level blocks needed if blocks are approximately 69% full.
   (c) The number of levels needed if internal nodes are also 69% full.
   (d) The total number of blocks required by the B+-tree.
   (e) The number of block accesses needed to search for and retrieve a record from the file--given its SSN value--using the B+-tree.

6. In Question 5, if we want to construct a B-tree access structure (index) on SSN instead of B+-tree, calculate the following:

   (a) The order p of the B-tree.
   (b) The number of leaf-level blocks needed if the B-tree is approximately 69% full.
   (c) The number of levels needed if the B-tree is 69% full.
   (d) The total number of blocks required by the B-tree.
   (e) The maximum number of block accesses needed to search for and retrieve a record from the file--given its SSN value--using the B-tree.
   (f) Can we calculate the number of block assesses instead of maximum number of block accesses for 6(e)? If not, why?

EE4791 Database Systems -Tutorial 8

1. Consider the following three relations for a College:

   STUDENT (StudentID, StudentName, Major, Age, MaritalStatus, PostalCode)
   REGISTRATION (StudentID, CourseID, Mark)
   COURSE (CourseID, CourseName)

   a) Choose Oracle data types for the attributes in table STUDENT based on their common characteristics.
   b) Suppose you are designing a default value for the age field in the database. What possible values you will consider and why? How might the default value vary by other characteristics about the student, such as school within the university or degree sought?
   c) When a student has not chosen a major at a university, the university often enters a value of "Undecided" for the major field. What is the most suitable default value for major? What is the difference between "Undecided" and "null" as default value?
   d) In addition to the information shown in the relations and discussed in b) and c), we further assume that StudentID and CourseID are foreign keys of the table REGISTRATION. Using SQL to define the tables for the three relations. Specify primary key and referential integrity constraints.

2. Consider the EER diagram shown in Figure 1. The estimates on the usage for this EER diagram are as follows:

   - 5,000 parts in which 3% are manufactured parts and 100% are purchased parts
   - 2000 suppliers
   - an average of 4 supplies per supplier
   - 50,000 direct accesses of part per hour
   - 20,000 direct accesses of purchased part per hour
   - 50% of accesses to purchased part will lead to the accesses of all the related supplies and their suppliers.
   - an average of 3 supplies associated with each purchased part
   - 8,000 direct accesses to supplier
   - 1% of accesses to supplier will lead to the access of all the associated supplies and purchased part associated

   Based on these estimates, draw a usage map for this EER diagram. If all these accesses are carried out interactively, identify two most useful ways to support the performance of these assesses without using keys or indexes.

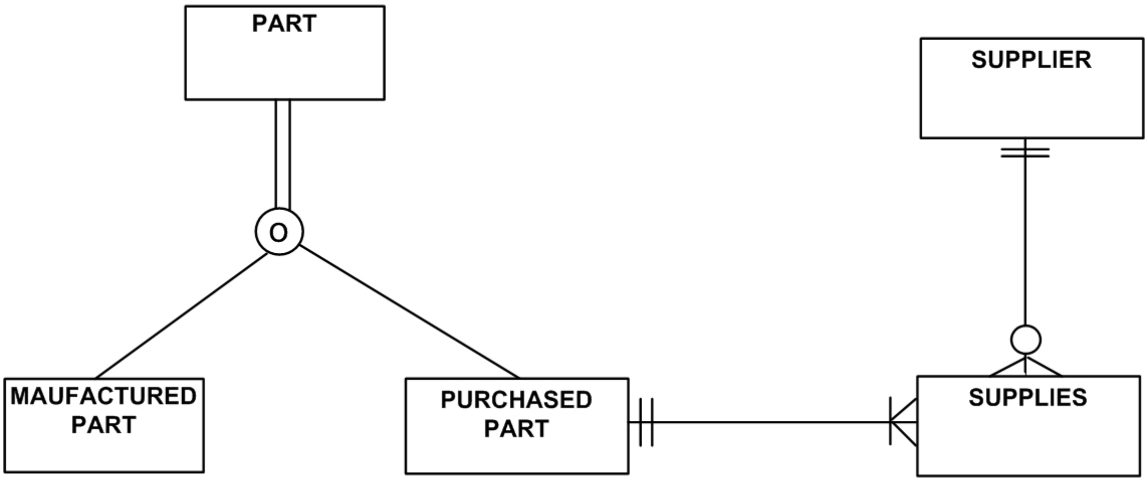

Figure 1. An EER Diagram

3. Consider the EER diagram shown in Figure 2. Based on the following estimates on the average usage of the system:

   • There are 50,000 customers, and of these, 80 percent represent regular accounts and 30 percent represent national accounts.
   • Currently, the system stores 800,000 orders, although this number is constantly changing.
   • There are 3,000 products.
   • Approximately, 500 orders are placed per hour and each of them has an average of 20 products. When an order is placed, a customer record is accessed directly and all the product records involved are also accessed directly.

   Based on these assumptions, draw a usage map for this EER diagram.

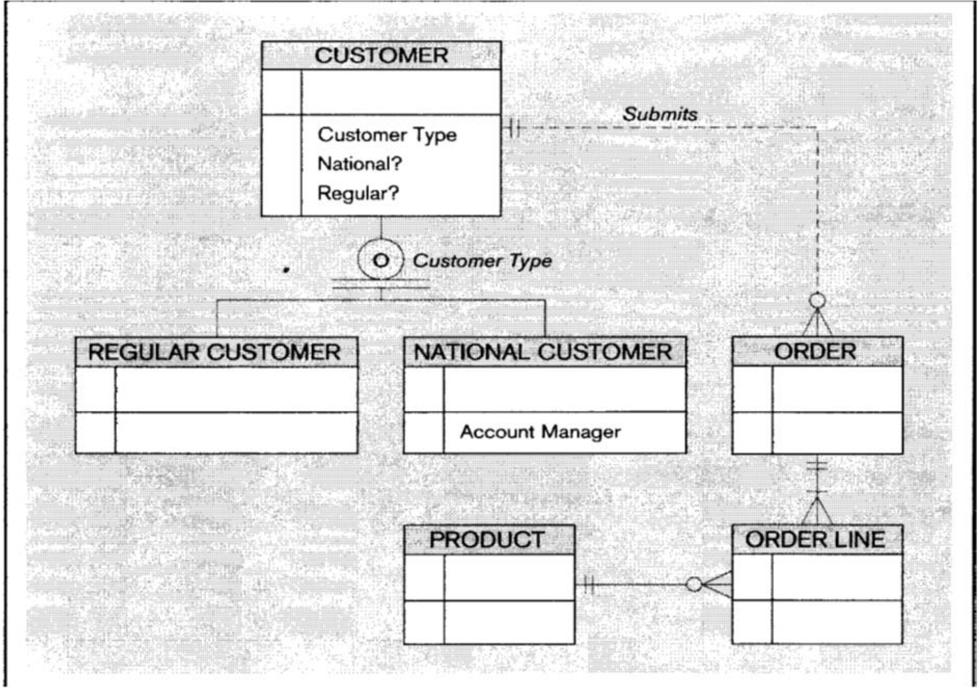

Figure 2. An EER Diagram

4. Consider the following normalized relations from a database in a large retail chain:

EMPLOYEE (EmployeeID, Employee_Name, EmployeeAddress, PhoneM)
SCHEDULE (DepartmentID, EmployeeID, Date)

Primary keys are underlined.

a) Explain why normalized relations may not be efficient physical records. What opportunities might exist for denormalizing these relations when defining the physical records for this database? Under what circumstances would you consider creating such denormalized records here?

b) Denormalize the tables assuming employee data is only accessed via schedule.

c) List drawbacks of denormalization.

EE4791 Database Systems -Tutorial 9

1. Consider the following normalized relations for a sports league:

   TEAM(TeamID, TeamName, TeamLocation)
   PLAYER(PlayerID, PlayerFirstName, PlayerLastName, PlayerDateOfBirth, PlayerSpecialtyCode)
   SPECIALTY(SpecialtyCode, SpecialtyDescription)
   CONTRACT(TeamID, PlayerID, StartTime, EndTime, Salary)
   LOCATION(LocationID, CityName, CityState, CityCountry, CityPopulation)
   MANAGER(ManagerID, ManagerName, ManagerTeam)

   What recommendations would you make regarding opportunities for denormalization? What additional information would you need to make fully informed denormalization decisions?

2. Consider the 3NF relational schema shown in Figure 1. Assume that the most important reports that the organization needs are as follows:

   a) A list of project assignments of a given developer
   b) A list of the total costs for all projects
   c) For each team, a list of its membership history
   d) For each country, a list of all projects, with projected end dates, in which the country's developers are involved
   e) For each year separately, a list of all the assignments that were completed during that year.

   Based on the limited information, make a recommendation regarding the indexes that you would create for this database. Choose two relations and provide the SQL command that you would use to create secondary indexes for the relations.

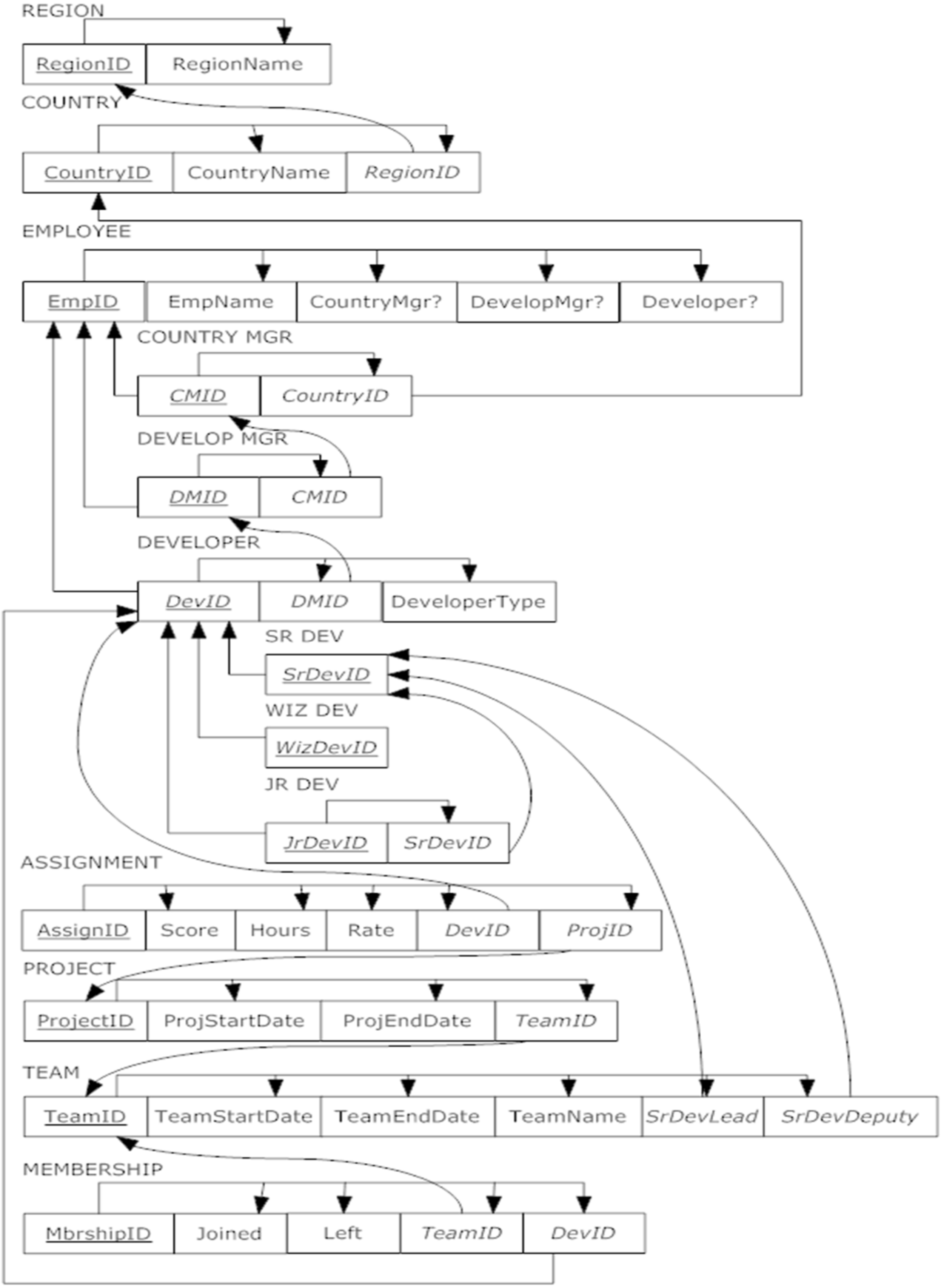

Figure 1. A 3NF Relational Schema

3. Create a join index on the CourseCode fields of the Course_T and Registration_T tables in Figure 2.

Course_T

RowID	CourseCode	CourseTitle
2001	EE4791	Database Systems
2002	EE4717	Web Application Design
2003	EE4190	Introduction to Modern Radar

Registration_T

RowID	StudentID	CourseCode	Grade
1001	A001	EE4791	A
1002	A001	EE4717	A
1003	B001	EE4791	A+
1004	C001	EE4791	B

Figure 2. An instance of Course_T and Registration_T

4. Consider the following normalized relations for sports leagues in global scope:

LEAGUE(LeagueID, LeagueName, LeagueLocation)
TEAM(TeamID, TeamName, TeamLocation, TeamLeague)
PLAYER(PlayerID, PlayerFirstName, PlayerLastName, PlayerDateOfBirth, PlayerSpecialtyCode)
SPECIALTY(SpecialtyCode, SpecialtyDescription)
CONTRACT(TeamID, PlayerID, StartTime, EndTime, Salary)
LOCATION(LocationID, CityName, CityState, CityCountry, CityPopulation)
MANAGER(ManagerID, ManagerName, ManagerTeam)

The following database operations are frequently used:

a) Maintenance (insert, update and delete) of these relations
b) Reporting players by team
c) Reporting players by team and specialty
d) Reporting players ordered by salary (the highest)
e) Reporting teams and players by city

Based on the above-mentioned information:

a) Identify the foreign keys
b) Make a recommendation regarding the indexes that you would create for this database.

Explain how you used the list of operations to arrive at your recommendation.

5. The following table shows some simple student data as of the date 06/20/2010:

Key	**Name**	**Major**
001	Amy	Music
002	Tom	Business
003	Sue	Art
004	Joe	Math
005	Ann	Engineering

The following transactions occur on 06/21/2010:

- Student 004 changes major from Math to Business.
- Student 005 is deleted from the file.
- New student 006 is added to the file: Name is Jim, Major is Phys Ed.

The following transactions occur on 06/22/2010:

- Student 003 changes major from Art to History.
- Student 006 changes major from Phys Ed to Basket Weaving.

Do the following:

a. Construct tables to represent transient data for 06/21/2010 and 06/22/2010, reflecting these transactions.

b. Construct tables represent periodic data for 06/21/2010 and 06/22/2010, reflecting these transactions.

6. Compare and contrast star schema with normalized relational data model.

EE4791 Database Systems -Tutorial 9

1. Identify the key differences between logical database design and physical database design.

2. State the major tasks to be carried out in physical database design and the methods that may be applied in each task.

3. Explain the difference between a static Web page and a dynamic one.

4. Contrast the following terms:
    a) Two-tier architecture; three-tier architecture.
    b) Fat client; thin client.
    c) ODBC; JDBC.
    d) SQL; Java.

EE4791 Database Systems -Tutorial 11

1. In a college, the following are four groups of data used to analyze grades achieved by students:

    - CourseSection Attributes: CourseID, SectionNumber, CourseName, Units, RoomID, and RoomCapacity. During a given semester, the college offers an average of 500 course sections.
    - Professor Attributes: ProfID, ProfName, Title, DepartmentID, and DepartmentName. There are typically 200 professors at the college at any given time. On the average, 2 professors teach one course section.
    - Student Attributes: StudentID, StudentName, and Major. Each course section has an average of 40 students, and students typically take five courses per period.
    - Period Attributes: SemesterID, and Year. The database will contain data for 30 periods (a total of 10 years).

    The only fact that is to be recorded in the fact table is CourseGrade. Do the following:

    a. Design a basic star schema for the loading the above-mentioned data in a data mart.
    b. Estimate the number of rows in the fact table, using the assumptions stated previously.
    c. Estimate the total size of the fact table (in bytes), assuming that each field has an average of 5 bytes.

2. In additional to the information given in Question 1, we further assume that:

    - RoomID uniquely identifies RoomCapacity.
    - CourseID uniquely identifies CourseNumber, CourseName and Units.
    - DepartmentID uniques identifies DepartmentName

    Discuss the advantages and disadvantages in using normalized dimension tables in star schemas for data marts. Based on these further assumptions, normalize the dimensional tables in the star schema designed in Question 1.

3. Assuming that the college stated in Question 1 now wants to include the following new data about course sections:

    - The department offering the course
    - The school to which the department reports

    Change the star schema designed in Question 2 to cater for the new data.

4. A food manufacturing company needs a data mart to summarize facts about the following type of orders to move goods:

    a) Transfer goods internally, between plants and from plants to storage
    b) Sales to customers from storage locations

c) Purchases from vendors to plants

d) Returns of goods from customers to storage locations

The company needs to treat customers, vendors, plants, and storage locations as distinct dimensions that can be involved at both ends of a movement event. For each type of destination or origin, the company wants to know the type of location (i.e., customer, vendor, etc.), name, city, and state. Facts about each movement include dollar and volume moved, cost of movement, and revenue collected from the move (if any, and this can be negative for a return). Design a star schema to represent this data mart directly. Simplify the resulting star schema through generalization.

5. An international pharmaceutical company operates a network of 300 chain drug stores all over the world. The company is setting up a drug data warehouse to store information for drug sales analysis. The total sales of drugs (Total_Items_Sold and Total_Sales_Value) per day for each drug and for each store should be kept in the data warehouse. There is an average of 50 different drugs sold by each store per day. Data for the data warehouse are extracted from the company database. There are ee relevant tables in the database:

   DRUG(Drug_ID, Drug_Name, Package_Dosage, Price)
   SALES(Drug_ID, Store_ID, Sales_Date, Prescription_Details)
   STORE(Store_ID, Address, Comuntry)

   Do the following:

   a) Design and draw a schema to represent the data warehouse accurately for the company.
   b) Estimate the number of rows in the fact table in part 5(a).

EE4791 Database Systems -Tutorial 12

1. Whitlock Department Stores runs a multiuser DBMS on a LAN file server. Unfortunately, at the present time, the DBMS does not enforce concurrency control. One Whitlock customer had a balance due of 250.00 when the following three transactions related to this customer were processed at about the same time:

   - Payment of $250.00
   - Purchase on credit of $100.00
   - Merchandise return (credit) of $50.00

   Each of the three transactions read the customer record when the balance was $250.00 (i.e., before any of the other transactions were completed). The updated customer record was returned to the database in the order shown in the bulleted list above.

   a. Show the sequence of events for the above-mentioned situation. What balance will be included for the customer after the last transaction was completed?

   b. If proper X locking mechanism is implemented, show the sequence of events for processing the transactions in the order shown in the bulleted list above. What balance should be included for the customer after the three transactions have been processed?

   c. If versioning is implemented for concurrency control, show the sequence of events for processing the transactions in the order shown in the bulleted list above. What balance should be included for the customer after the three transactions have been processed (without restarting any transaction)?

2. For each situation described below, decide the appropriate recovery techniques to be applied:

   a. A phone disconnection occurs while a user is entering a transaction.

   b. A disk drive fails during regular operations.

   c. A lightning storm causes a power failure.

   d. An incorrect amount is entered and posted for a student tuition payment. The error is not discovered for several weeks.

   e. Data entry clerks have entered transactions for two hours after a full database backup when the database becomes corrupted. It is discovered that the journalizing facility of the database has not been activated since the backup was made.

3. For each of the situations described below, indicate which of the following security measures is most appropriate:

   a. A national brokerage firm uses an electronic funds transfer (EFT) system to transmit sensitive financial data between locations.

   b. An organization has set up an offsite computer-based training center. The organization wishes to restrict access to the site to authorized employees. Because each employee's use of the center is occasional, the center does not wish to provide the employees with keys to access the center.

   c. A manufacturing firm uses a simple password system to protect its database but finds it needs a more comprehensive system to grant different privileges (e.g., read, versus create or update) to different users.

d. A university has experienced considerable difficulty with unauthorized users accessing files and databases by appropriating passwords from legitimate users.

4. During the Sarbanes-Oxley audit of a financial services company, you note the following issues. Categorize each of them into the area to which they belong: IT change management, logical access to data, and IT operations:

   a. Five database administrators have access to the sa (system administrator) account that has complete access to all the databases.
   b. Several changes to database structures did not have appropriate approval by management.
   c. Some users continued to have access to the database even after having been terminated.
   d. Databases are backed up on ad hoc basis due to manpower problem.