Databases Notes (with even more memes)

SQL Tings (not necessary for the exam, just nice to know)

- DML: Data Manipulation Language
    - SELECT; retrieves data from a table
    - UPDATE; updates data into a table
    - DELETE; deletes all records from a table
    - INSERT; inserts data into a table
- DDL: Data Definition Language
    - DROP; deletes the objects from the table (or the whole table) completely
    - CREATE; creates objects
    - ALTER; alters objects
    - TRUNCATE; deletes records from a table and resets the table to initial value (the length stays the same)
- DCL: Data Control Language
    - GRANT; gives user permission
    - REVOKE; takes away user permission
- TCL: Transactional Control Language
    - BEGIN; begins a transaction
    - COMMIT; saves work done
    - ROLLBACK; restores data before the last commit
    - SAVEPOINT X; sets an intermediate point within the transaction
    - ROLLBACK TO X; undoes everything before X

A.1

- Data are just fax. Raw fax. Numbers or text only.
- Information is a piece of data put into context
- Computers process data into information
- An information system is any combination of IT and people using the technology to support operations, management, and decision-making
    - Store, collect, and process data
    - Transform data into information
- A database is an organized pool of logically related data that can easily be accessed
    - Data is stored within the data structures of the database
    - A typical information system has somewhere to store data, which is a database
    - Microsoft Access is an example of a database management system (DBMS, more like BDSM), which is a software that handles databases
- Numerical data types require more memory than text, so using the text data type for numbers that don't require calculations saves memory
- Data processing is what transforms data into information.
    - You give raw data to an information system and it does its stuff, such as summarizing, calculating averages, graphs, whatever really
    - The output of data processing is information.

- Why do we use databases?
    - ~~Because our teachers hate us and we should have done OOP~~
    - ~~—~~Can store very large numbers
    - ~~—~~Quick and easy to find information
    - ~~—~~Easy to add new data or delete old data
    - ~~—~~Data can be searched and sorted easily
    - ~~—~~Data can be imported into other applications

- A database transaction is a sequence of operations/single unit of work which is done to update a piece of data in a database.
- It may or may not work.
- It is a sequence of steps that satisfy a client's requests.
- When moving money between accounts, one account must be debited and the other must be credited. Unless both operations are carried out successfully, the transaction will be rolled back.

- Concurrency is defined as the ability for multiple processes to access or change shared data at the same time.
- The more processes that can be done at the same time without blocking each other, the better the concurrency of the system.
- Problems with concurrency include: lost update, temporary update problem (uncommitted data), incorrect summary problem (inconsistent retrieval)
    - Lost update: Problem occurs when two transactions accessing the same item have their operations interleaved in a way that makes some items incorrect.
        - Safe: Begin, Update, Commit, Begin, Update, Commit (Both updates work)
        - Error: Begin, Begin, Update, Update, Commit, Commit (Blue update only)
    - Uncommitted data: Problem occurs when one transaction on an item is rolled back and the item needs to be updated again, but that item is accessed before the rollback. Basically, you rolled back too late.
        - Safe: Begin, Update, Commit, ***Rollback***, Begin, Update, Commit (Both work)
        - Error: Begin, Update, Commit, Begin, Update, ***Rollback***, Commit (The information from red is not restored)
    - Inconsistent retrieval: Problem occurs when one transaction is trying to calculate a summation of items in a table and the other is trying to alter those values.

- The ACID properties of a database transaction, which all transactions must conform to:
    - **A**tomicity – the transaction must be completed fully. If not, then the transaction won't be recorded. All or nothing.
    - **C**onsistency – the data remains consistent at the start and end of a transaction. For example, when transferring funds from one bank account to another, consistency ensures that the total funds in both accounts does not change.

    - **I**solation – multiple transactions can occur at the same time such that each transaction does not affect the outcome of the others.
    - **D**urability – once the transactions are complete, the new data values are stored to the hard disk.

- A deadlock is a situation where two transactions (on different tables) are waiting for each other to give up their locks. Both of them have locked each other because they both need information from each other.
    - The DBMS has to detect a deadlock and abort it.
- Row locking prevents two sessions from updating the same data at the same time. When is a row is locked, another session cannot update data from that row until the lock is released. If the entire database is locked then only one session can apply updates.
- The difference between data validation and data verification:
    - Validation happens first, to check if inputted data matches the conditions (correct data type, number of characters, max-min numbers, etc.)
    - Verification happens next, which checks that the data entered exactly matches the source.

A.2

- A database contains tables, which contain records, which contain fields, which contain characters~~, which contain pixels, which contain subpixels, which contain…~~
- A table is a collection of **related** records
- A field is the smallest unit of data – has a name, size (length), and data type
- A foreign key is a primary key from one table that is accessed from another table – it may not be a primary key in the new table
- A candidate key is a key that **can** be a primary key but **isn't**
- A composite key is a primary key made up of multiple attributes
- A schema is the plan or layout for a database. It dictates which tables (a type of entity) are present and how they are linked.
    - Conceptual schema: a model that contains entities (concepts) and their relationship between them. It is a basic model.
    - Logical schema: shows and classifies the entities as tables, fields, or objects. It shows arrows connecting different entities, and each table has the fields inside it.
    - Physical schema: even more details. Shows details about data types and restrictions/conditions.
- A data dictionary is a file or set of files that contains metadata. It contains:
    - The definitions of records/entities in the database,
    - How much space has been allocated,
    - Default values for columns,
    - Providing information about the privileges/access rights of certain users,
    - NOT actual data. Just metadata and information for managing the database. Without a data dictionary a DBMS cannot access data.

- Data normalization is the process of simplifying a database and reducing redundancy.

- Unnormalized form has all the data in one table, and honestly, it's quite a mess
    - Changing things would require updating them many times
    - Normalization mainly splits the table into smaller and more well-structured tables

- Update anomalies – issues that normalization solves:
    - Insertion anomaly – if you want to add new records you have to add the repeated data
    - Deletion anomaly – if all of the data is deleted you lose the repeated data
    - Modification anomaly – if one record of the repeated data is changed it will have to be changed everywhere

- Referential integrity ensures that every record in a related table is related to the primary table. A record without a matching relation is known as an orphaned record.

- 1st Normal Form (1NF):
    - Each cell must be a single value (atomic)
    - Entries in each column must be the same data type
    - Rows are uniquely identifiable

- 2nd Normal Form (2NF):
    - Must be in 1NF
    - All non-key columns must depend on the primary key
    - No partial dependencies
    - Tables must contain data about only one type of entity

- 3rd Normal Form (3NF):
    - Must be in 2NF *(NO KIDDING)*
    - No transitive dependencies (a non-key depends on another non-key)

A.3

- The database administrator:
    - Protects data
    - Installs the database software
    - Configures the database to the way that is required
    - Upgrades the software if needed
    - Migrates the data to a new machine if needed
    - Deals with backup and recovery
    - Handles troubleshooting
    - Works with software developers
    - ~~Has no life~~

- The other people that use a database include:

    - Application programmers – developers for the software. They retrieve information and create/change information. They use DML calls and make requests to the DBMS.
    - System analysts – people that analyze the system *(NO KIDDING [2])* and make predictions for costs and economic/technical feasibility.
    - End users – casual users (can and cannot use queries respectively). They are just trained to press a couple buttons. Such n00bs.

- Approaches to recover data:
    - Manual reprocessing: the data is regularly backed up and if the system crashes, the latest backup is restored. The users will have to reapply transactions to bring the database to how it was just before the crash. Reapplying transactions will take time and may cause concurrency issues.
    - Automated recovery: When a transaction happens, it is recorded in a log or journal separate from the database. The information that is logged includes: what the database looked like before the transaction, what it looked like after, who made the transaction, and when it happened. Rollback applies before images and rollforward applies after images.

- Integrated database systems: the connection of several databases built into another application.
    - Data integration involves combining data from different sources. This can be for commercial purposes (when two similar companies need to merge their data) or scientific purposes (when two similar research institutes need to compare their results) or many other purposes.
    - For example, many hospitals that are undergoing cancer research can combine their results to identify a trend.
    - Collaboration is encouraged (sounds like myp bs to me).

- A warehouse is a large collection of data.
    - Companies can identify their best-selling product.
    - Information about every sale can be accumulated to make predictions of what people will buy.
    - They can place offers and sales on products that are selling well.
    - They will achieve ***S T O N K S*** .

- Computer misuse – accessing material without permission, accessing material with criminal intentions, altering data without permission
- Data protection – data must be collected with the subject's consent, data must only be kept if necessary, data must be accurate and kept up-to-date, data cannot be transferred
- Measures to protect data:
    - Encryption – scrambling data so that only people with the key can see it
    - Auditing – checking to see who is using the database and what they're doing
    - Authentication – used to grant users the right to see information

- Views – giving users specific views so that they can only see what they need

- In some situations security is more important than privacy. Authorities can access anyone's personal data if there are any criminal suspects. (still sounds like bs to me)

- Data mining is the process of analyzing ***B I G*** Data (it best be capitalized) and summarizing it into useful information for decision making. Data matching (link analysis) is the process of analyzing records of the same individual/product in different resources.

Mining	Matching
Use of patterns to identify trends	Comparing two sets of collected data
Extracting unknown pieces of information from large databases	Finding errors in data
Developing techniques for discovering knowledge in large databases	Identifying records that correspond to the same entities across different databases