Databases and Info Systems

Introduction to Databases

Dr. Seán Russell
sean.russell@ucd.ie,

School of Computer Science,
University College Dublin

February 19, 2020

Table of Contents

Why Study Databases?

- Databases are an essential component of life in modern society: most people do several activities every day that involve some interaction with a database
- E.g. Searching for or buying something on Taobao or buying items in a supermarket
- As software engineers, we will have to at times work on systems that interact with or mange these data storage systems
- We must have an understanding of how they work and how we can interact with them to find/update the information we need

Managing the Data

- The systems discussed in the example all represent huge amounts of information
- we learned in OOP to create programs that could store all of the information in files
- But managing the information would be a very difficult task
- Because of these difficulties, we usually use a **D**ata**B**ase **M**anagement **S**ystem (DBMS) when we need to store large amounts of information

What is a Database?

- The most basic definition is: "A database is a collection of **related** data"
- For example, consider the names and phone numbers of the people you know
  - You may this data recorded in the contacts in your phone
  - This is a collection of **related** data with an implicit meaning and hence is a database

Properties of a Database

- There are some more specific properties that we can use to better define what a database is
  - A database represents some aspect of the world
  - A database is a logically connected collection of data with some meaning
  - A database is designed, built and populated with data for a specific purpose

Database Size

- A database can be of any size and complexity
- Small databases like our phone contacts are databases
- Large databases like the tax records of everyone in China are databases

Table of Contents

Database Management Systems

- A **D**ata**B**ase **M**anagement **S**ystem (**DBMS**) is a collection of programs that enables users to create and maintain a database
- The DBMS is a general-purpose software system that allows the processes of **defining**, **constructing**, **manipulating**, and **sharing** databases among various users and applications

DBMS Models

- Most databases used in the world are based on the **relational** model
- Other models were tried first
  - Hierarchical model
  - Network model
- Newer models are popular in some uses
  - **NOSQL** (**N**ot **O**nly **SQL**)

Keeping Records

- Governments and businesses have always kept records of **important** information
- For centuries this was on **paper**
- To find a piece of information, large files or books would have to be **searched**
- This would take longer the more information that was recorded


wiseGEEK

Electronic Documents

- When computers became more widely used, this made it practical to store information in electronic storage
- Early attempts to achieve this usually tried to create electronic version of the documents being stored
- The major use that was tried first was storing financial information

Spreadsheets

- Accountants have been using spreadsheets for centuries to record financial information
- In the 1960s, electronic versions of these spreadsheets were developed for use on mainframe computers
- However, modern computerised spreadsheets began with the development of VisiCalc in 1978
- VisiCalc became the basis for all electronic spreadsheets since (like Excel)

Visicalc

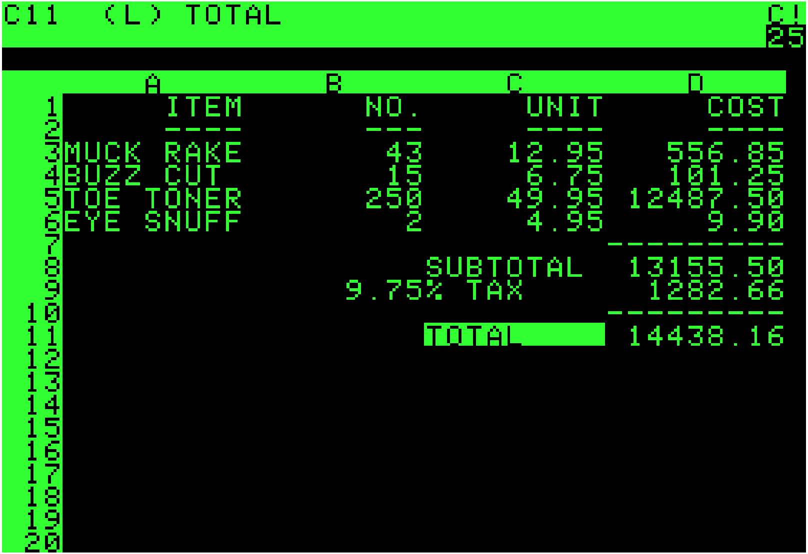

C11 (L) TOTAL
C!
25
A
B
C
D
ITEM
NO.
UNIT
COST
MUCK RAKE 43 12.95 556.85
BUZZ CUT 15 6.75 101.25
TOE TONER 250 49.95 12487.50
EYE SNUFF 2 4.95 9.90
SUBTOTAL 13155.50
9.75% TAX 1282.66
TOTAL 14438.16

Table of Contents

Electronic Documents

- Storing the same data as an electronic copy of an existing document is not good enough
- How do we process or search for information in these files?
- For example, we often want to know the answers to complex questions like:
  - How many Stage 2 BSc students do we have?
  - Is Sean Russell in the list?
  - What modules is he registered for?

File Based Databases

Advantages

- More lightweight than a DBMS
- Good enough for small data

Disadvantages

- No query language
- No scalability
- Hard to update schema or modify data
- Recovery Issues

Hierarchical Database Model

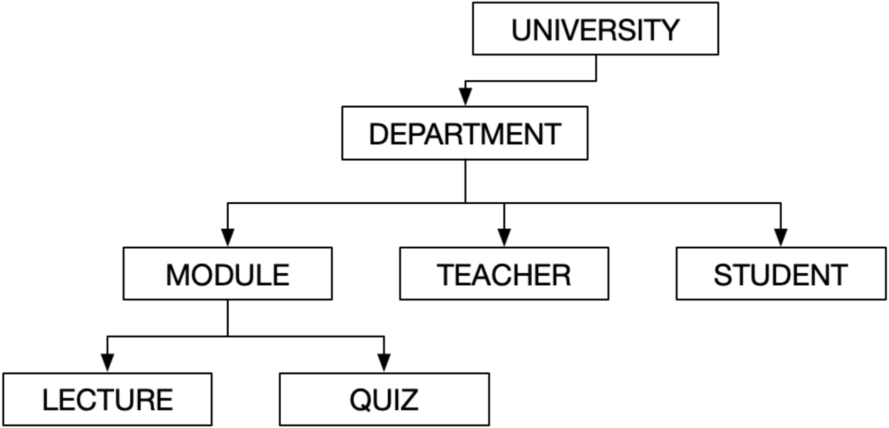
Advantages

- It was based on a tree structure consisting of nodes, branches and roots
- It allowed for 1 to many relationships between different types of data

Disadvantages

- There was no standard of the hierarchical model so implementations were different in how they worked and stored data.
- Relationships can only be between parent and child

Hierarchical Database Model

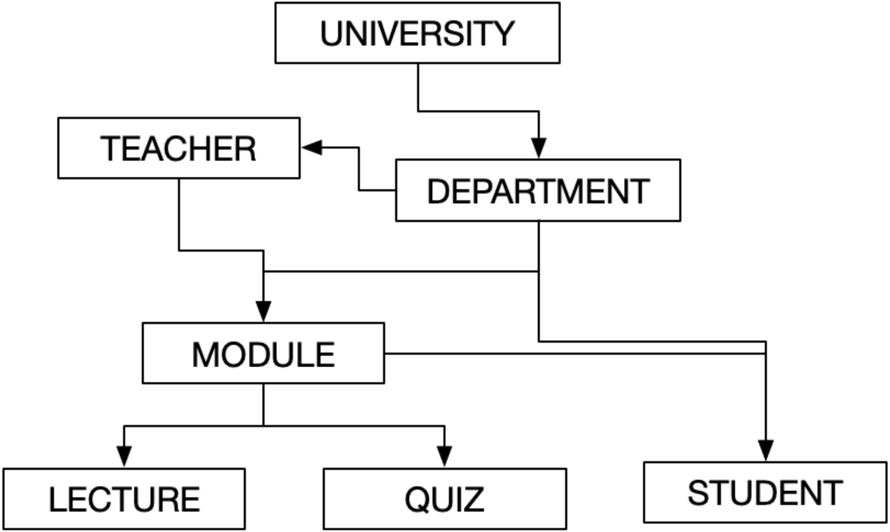


Network Database Model

Advantages

- It was based on a graph structure consisting of nodes, and edges
- The Database Task Group (DBTG) of CODASYL (COnference on DAta SYstems Languages) developed the model so that it could be standardised
- It allowed for many to many relationships between different types of data

Network Database Model



Relational Database Model

- Formulated by Edgar Codd of IBM in 1970
- Commercial RDBMS in 80s
- Codd's 12 Rules (actually 13) that all RDBMSs should follow
- Most widely used model at present
    - Access, Oracle, MySQL, MariaDB, MS SQL Server, DB2, Sybase ASE, PostgresSQL etc.

Relational Database Concepts

- Data is represented as collections of **relations**
- Each relation is **table** of values
- Each table consists of **rows** and **columns**
- Each **row** represents an **entity** or **record**
- Rows are **unordered**

Relational Database Concepts

- No duplicate rows are allowed
- Each relation has a **primary key**, the value of which uniquely identifies the **record/entity**
- Each column represents an **attribute**
- Table name and column names are used to help interpret the values

Advantages of Relational Approach

- Data can be shared
- Redundancy can be reduced
- Integrity can be maintained
- Security can be enforced
- Conflicting requirements can be balanced
- Standards can be enforced

NoSQL

- NoSQL databases were invented in the 2000s to deal with some new challenges that traditional relational databases struggled with
- Especially useful for large, unstructured data, and data that does not need to be updated instantly
- There are different types of NoSQL database

NoSQL Types

Document databases - pair each key with a complex data structure known as a document. Documents can contain many different key-value pairs, or key-array pairs, or even nested documents. Typical example: *MongoDB*

Graph stores - used to store information about networks of data, such as social connections. Graph stores include *Neo4J* and *Giraph*.

NoSQL Types

Key-value stores - the simplest NoSQL databases. Every single item in the database is stored as an attribute name (or 'key'), together with its value. Examples of key-value stores are Riak and Berkeley DB. Some key-value stores, such as *Redis*, allow each value to have a type, such as 'integer', which adds functionality.

Wide-column stores - such as *Cassandra* and *HBase* are optimized for queries over large datasets, and store columns of data together, instead of rows.

Table of Contents

Database Terminology

- **Relation** is a mathematical term for a **table**
- **Row** is called a **Tuple**
- **Column** is called an **Attribute**
- **Domain** is used to describe the types of values that can appear in a column
- **Degree** is the number of attributes
- **Cardinality** the number of tuples/rows in a relation
- **Atomic Value** precisely one value at each row intersection
- **Null Value** Missing, not known or irrelevant data (not the same as zero or blank)

Relational Example (students relation)

student_num	name	major	year_of_entry
17206777	Sean Russell	SE	2017
18205333	David Lillis	IOT	2016
16205777	Brett Becker	EIE	2016

Relational Example

- The degree of `students` is 4 (there are 4 columns/attributes)
- The attributes of `students` are named `student_num`, `name`, `major` and `year_of_entry`
- The cardinality of `students` is 3 (there are three rows/tuples)
- The domain of the attribute `major` is one of the values in SE, IOT, EIE, and FIN and cannot be any other values.
- The primary key is the attribute `student_num` as this uniquely identifies each row in `students`