

Chapter 2
Databases: Access to Structured Data

CA557
Information Access

Alan Smeaton & Cathal Gurrin © 2001-2008 - 1 - DCU

What is a Database?

- Database vs DBMS...
 - A DBMS is the software, a database is the data.
 - There are many different DBMSs, some free, some very expensive.
 - A database is usually a disk file or group of files.
- So, a DBMS is a software package than runs on a mainframe, server, desktop, or mobile device which at least provides fast access to structured data stored in databases.

Alan Smeaton & Cathal Gurrin © 2001-2008 - 2 - DCU

DBMS & databases

```

    graph TD
      DBMS[DBMS  
(MS SQLServer)] --- DB1[Database 1]
      DBMS --- DB2[Database 2]
      DBMS --- DB3[Database 3]
      DB1 --- T1[TABLE: Supplier]
      DB1 --- T2[TABLE: Customer]
      DB2 --- T3[TABLE: Product]
      T2 --- C1[Customer.Name]
      T2 --- C2[Customer.Address]
      T2 --- C3[Customer.City]
    
```

Alan Smeaton & Cathal Gurrin © 2001-2008 - 3 - DCU

Data operations

- With a Databases (managed by the DBMS) support information management by supporting the following operations on data:
 - **Insertion** of new data
 - **Deletion** of existing data
 - **Modification** of existing data
 - **Retrieval** of existing data, where this retrieval is usually based on some filtering arguments (e.g. retrieval of suppliers where their location is Dublin, ... , ...)

Alan Smeaton & Cathal Gurrin © 2001-2008 - 4 - DCU

Why use a Database at all?

- Scale
- Speed
- Flexibility
- Concurrency
- Backup & Recovery
- Transaction Management
- Data Integrity
- Distributed databases
- Hence, using a DBMS makes the **development** of applications faster, less prone to errors, and more flexible.

Alan Smeaton & Cathal Gurrin © 2001-2008 - 5 - DCU

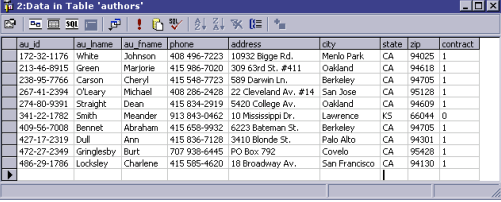
What can be stored in a DB?

- Anything really, even binary data
- Each database vendor will support different types of underlying datatypes:
 - Text
 - Char
 - Varchar
 - Text
 - Number
 - Float
 - Integers {int, smallint, tinyint}
 - Others {decimal, smallmoney, money, real, numeric}
 - Other
 - Binary {binary, varbinary}
 - Bit
 - Time {smalldatetime, datetime, timestamp}
 - image

Alan Smeaton & Cathal Gurrin © 2001-2008 - 6 - DCU

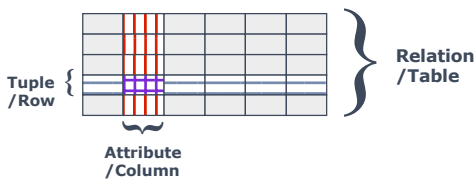
Important Concepts of the Relational Model

- **Table** – a structured repository of data of a specific type.
 - The table is the underlying data storage device in a relational database.
 - Important here is the rule that the data stored in a table is all of the same type.



Alan Smeaton & Cathal Gurrin © 2001-2008 - 7 - DCU

... more concepts ...



Alan Smeaton & Cathal Gurrin © 2001-2008 - 8 - DCU

... more concepts ...

- **/ row** – the data of a specific type stored in a table.
 - E.g. suppliers table, if there are five suppliers, then there are five tuples / rows.
 - There must be no significance in the order of tuples.
- **Column / attribute** – Rows are composed of columns or attributes.
 - Each column contains a single unit of data, stored in one of the supported datatypes.
 - For example, any row of the suppliers table may contain supplier name, address and phone number, each as individual columns.

Alan Smeaton & Cathal Gurrin © 2001-2008 - 9 - DCU

... more concepts ...

- **Keys...** we will see later
- **Indexes** and fast access
 - Speed is a vital aspect of any DBMS.
 - A slow DBMS is unacceptable.
 - Indexes provide fast access to data (at the cost of disk space efficiency).
 - Many different types exist and their task is to ensure that the database is structured in such a way as to support fast retrieval performance.

Alan Smeaton & Cathal Gurrin © 2001-2008 - 10 - DCU

More on Indexes

- Similar to the back of a book index, they maintain a sorted list to support fast retrieval.
- They are NOT created by default:
 - Searching through large columns of data is usually not efficient, unless an index exists...
 - the DBMS searches the index to locate the required data and retrieves those specific tuples (rows).
- Great, so why not every column?
 - Indexes degrade performance of data insertion, modification and deletion.
 - Indexes are memory and disk space hungry.
 - Not all data is suitable for indexing, data must be sufficiently unique.

Alan Smeaton & Cathal Gurrin © 2001-2008 - 11 - DCU

Complexity of Representing Data

- Complexity can stem from apparent simplicity...
 - June 1, 2004
 - 1 June 2004
 - 01/06/2004
 - 06/01/2004
 - 20040601
 - 01062004

Alan Smeaton & Cathal Gurrin © 2001-2008 - 12 - DCU

Another Example...

- John F Kennedy
- Kennedy, John F.
- JFK
- The 35th President of the USA
- John Fitzgerald Kennedy

Alan Smeaton & Cathal Gurrin © 2001-2008 - 13 - DCU

So... ..

- If we define a structure for the data we know what each element is...
 - XML
 - XML Databases
 - SGML
 - HTML ???
 - Relational Databases
 - Which allows us to query for precise data:
 - Stored using precise data types
 - We know what we are getting is the one and only correct answer to our query...
 - So, how do we query?

Alan Smeaton & Cathal Gurrin © 2001-2008 - 14 - DCU

... more concepts ...

- **SQL** (Structured Query Language)
 - a query language for structured data in a database
 - Artificial language... opposed to SE queries in natural language..
 - Simple and efficient, it has a number of key benefits
 - Firstly, it is not proprietary, but every relational DBMS vendor supports SQL (and have their own extensions).
 - Secondly it allows for performing complex and sophisticated database operations..
 - Finally it is fairly easy to understand and use.

Alan Smeaton & Cathal Gurrin © 2001-2008 - 15 - DCU

What can we do with SQL

- Data Manipulation
 - Select – query for data
 - Insert – add new data
 - Delete – remove data
 - Update – update data
- Data Definition
 - Create {table, view}
 - Alter {table, view}
 - Drop {table, view}
 - Other database commands

Alan Smeaton & Cathal Gurrin © 2001-2008 - 16 - DCU

4.2 The Select Statement (basic form)

SELECT <something> FROM <somewhere> WHERE <some limitation> ---->

*
name
age
address
...
gender

attribute names
or *

people
employees
authors
...
movies

tables or views

age > 25
name="smith"
Address="Dublin"
...
Salary > 25,000

constraints based on
attribute values
or entity presence

Everything is structured... no ambiguity !!!

Alan Smeaton & Cathal Gurrin © 2001-2008 - 17 - DCU

A Sample Database

```

    graph LR
        movie --> stars
        movie --> movieStar
        movie --> movieExec
        studio --> movie
        stars --> movieStar
        stars --> movieExec
    
```

SELECT * FROM movie WHERE year = 2004

SELECT title, length FROM movie WHERE year = 2004

Alan Smeaton & Cathal Gurrin © 2001-2008 - 18 - DCU

INSERT Operations

So we can INSERT a new MovieStar like this:

- `INSERT INTO MovieStar (name, address, gender, birthdate)`
- `values (Tom Hanks', 'New York', 'm', '12/31/58')`

But we can INSERT a new MovieStar like this, not giving the column names – but it's not recommended, can lead to errors:

- `INSERT INTO MovieStar`
- `values ('Tom Hanks', 'New York', 'm', '12/31/58')`

Alan Smeaton & Cathal Gurrin © 2001-2008 - 19 - DCU

DELETE Operation

- This is self-explanatory, but be careful with the WHERE<condition>
- So to delete a given actor from the MovieStar Table:
`DELETE FROM MovieStar WHERE name = 'Jennifer Lopez'`
- or
`DELETE FROM sales WHERE cost <= 100.00;`

Alan Smeaton & Cathal Gurrin © 2001-2008 - 20 - DCU

UPDATE Operation

- For example, we can update the MovieStar relation to reflect a change in Hollywood's name to 'Hollywood'.
`UPDATE MovieStar`
`SET address = 'Hollywood'`
`WHERE address = 'Hollywood'`
- Or to proportionally increase marks in an exam.
`update Students`
`set CA557Mark = CA557 * 1.1`
`where class = 'MEC'`

Without the WHERE clause, all tuples will be updated!

Alan Smeaton & Cathal Gurrin © 2001-2008 - 21 - DCU

UPDATE Operation

- OK, so compare this to a Search Engine...
 - In a Search Engine the data is a 'bag of words' on a web page.
 - Match bag of words against bags of words
 - In a DBMS, all the data is structured and we know what each piece of data is referring to.
 - If the WWW were a DBMS, then the result of a search would always be correct.
- So, how does one know how to structure data in a database?

Alan Smeaton & Cathal Gurrin © 2001-2008 - 22 - DCU

... more concepts ...

- **ER models** – the logical design of a database, called the database schema.. Relational model
 - Try to model the complexity of real-world data
- ER models have the following properties:
 - An entity is an instance of a physical object in the real world.
 - An entity (set) is a group of objects of the same type.
 - An entity has properties or *attributes* to describe its characteristics.
 - Entities can be associated via *relationships*, and each relationship can have properties or attributes.
 - Entities become tables and relationships become tables.

Alan Smeaton & Cathal Gurrin © 2001-2008 - 23 - DCU

Entities (cont)

Model a customer buying a product... example here focuses on the customer entity only...

```

    graph LR
      Customer[Customer Entity] --- address((address))
      Customer --- contact((contact))
      Customer --- status((status))
      Customer --- phone((phone))
      Customer --- name((name))
      Customer --- orders{orders}
      orders --- Product[Product Entity]
      style Customer fill:#f96
      style Product fill:#ccc
      style address fill:#fff,stroke:#333
      style contact fill:#fff,stroke:#333
      style status fill:#fff,stroke:#333
      style phone fill:#fff,stroke:#333
      style name fill:#fff,stroke:#333
      style orders fill:#fff,stroke:#333
      style Product fill:#ccc,stroke:#333
  
```

Alan Smeaton & Cathal Gurrin © 2001-2008 - 24 - DCU

Attributes

- Attributes represent properties that 'adequately' describe an entity.
- Should have a descriptive name...
- May be of many types ... supported by DBMS... {varchar, char, int, float, bit,...}
- An object represented by an entry may have a value for each attribute...
 - EMPLOYEE : { fname, lname, age, room, phone }

Cathal Gurrin 29 LI.10 5442

- Values may be blank...
 - These are NULL values... may cause problems later...
 - Can disallow NULLs in a particular attribute
- May have default values... if none is entered by a user.

Alan Smeaton & Cathal Gurrin © 2001-2008 - 25 - DCU

... more concepts ...

- Joins** – the ability to join two or more tables together to create a short-lived (life of the query) temporary virtual table for the purpose of complex SQL queries.
- Views** – a virtual table based on an underlying table or number of tables (or subset of both). Views are often used to limit access for certain users to certain data... instead of giving them access to the whole table (incl. salary details for example) you can give them access to a view, which may contain all employee information except the salary details.

Alan Smeaton & Cathal Gurrin © 2001-2008 - 26 - DCU

A View

Underlying Table

1	2	3	4	5	6

A View over the Table

1 (2)	2 (3)	3 (4)

Alan Smeaton & Cathal Gurrin © 2001-2008 - 27 - DCU

Another View

Underlying Table

1	2	3	4	5	6

A View over the Table

1 (2)	2 (3)	3 (4)

Alan Smeaton & Cathal Gurrin © 2001-2008 - 28 - DCU

Views for Security

Table A
(personal data concerning employees)

Managers

SELECT
UPDATE
DELETE
INSERT

View VA
(derived from Table A)

Employees

SELECT
UPDATE
DELETE
INSERT

DENIED!

"Where do we store the access right information for users?"

Alan Smeaton & Cathal Gurrin © 2001-2008 - 29 - DCU

Database Integrity

- Huge benefit of using a DBMS
- Essential that data remains in a state of integrity
- How do we support this?
 - Transactions
 - Concurrency
 - Backup & Recovery
 - Database Replication
 - Normalisation
 - AND KEYS & REFERENTIAL INTEGRITY

Alan Smeaton & Cathal Gurrin © 2001-2008 - 30 - DCU

Keys

- Every tuple/row in a table should have an attribute, or group of attributes that uniquely identifies it.
 - A **Key** is an unique identifier of any row in a table...
 - A **Candidate Key** is an attribute or combination of attributes which is a unique row identifier.
 - One candidate key is chosen as the **Primary Key** and the others are alternate keys.
 - A **Compound key** is a concatenated key... a concatenation of attributes is required for uniqueness of the key... e.g. first name & last name and age
 - A **Simple key** is a key that is not a compound key, if a single row is sufficient to identify a row.
 - A **Foreign key** is a (combination of) attribute(s) in one table whose values are required to match those of the primary key of another table.
 - Foreign keys are not necessarily part of the primary key and foreign-to-primary matches represent references.

- 31 -

A Key is an Attribute

-EMPLOYEE : { f_name, l_name, age, room, phone, pps }

- 32 -

Candidate Key Examples

Employee#	f_name	m_i	l_name	phone	pps	age
12	Neil	G	Howie	7023212	123-232	42
13	Rowan	M	Morrison	7035622	431-221	29
14	Alder	E	MacGregor	3025214	451-123	48
16	Ash	G	Buchanan	1024877	043-299	26
18	Willow	R	MacGregor	3021264	563-232	25
..
..
23	Alder	T	MacGregor	3021268	239-941	29

- 33 -

Compound Key

Employee#	f_name	m_i	l_name	phone	pps	age
12	Neil	G	Howie	7023212	123-232	42
13	Rowan	M	Morrison	7035622	431-221	29
14	Alder	E	MacGregor	3025214	451-123	48
16	Ash	G	Buchanan	1024877	043-299	26
18	Willow	R	MacGregor	3021264	563-232	25
..
..
23	Alder	T	MacGregor	3021268	239-941	29

A Unique Identifier?

- 34 -

Compound Key

Employee#	f_name	m_i	l_name	phone	ssn	age
12	Neil	G	Howie	7023212	123-232	42
13	Rowan	M	Morrison	7035622	431-221	29
14	Alder	E	MacGregor	3025214	451-123	48
16	Ash	G	Buchanan	1024877	043-299	26
18	Willow	R	MacGregor	3021264	563-232	25
..
..
23	Alder	T	MacGregor	3021268	239-941	29

A Unique Identifier now?

- 35 -

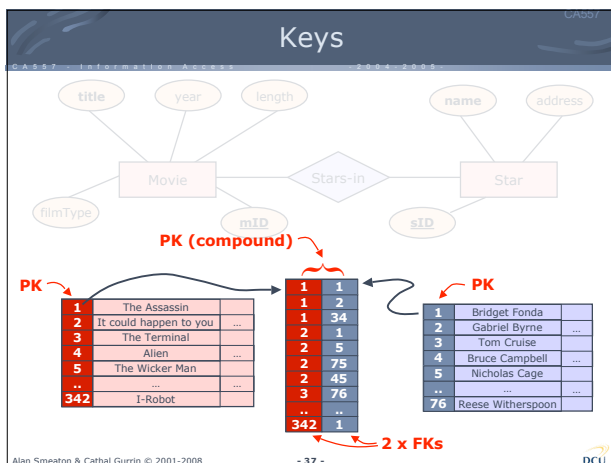
Keys

1	2	3	4	5	..	342
The Assassin	It could happen to you	The Terminal	Alien	The Wicker Man	..	I-Robot

1	2	3	4	5	..	342
1	34	1	5	75	45	76
..	1

1	2	3	4	5	..	76
Bridget Fonda	Gabriel Byrne	Tom Cruise	Bruce Campbell	Nicholas Cage	..	Reese Witherspoon

- 36 -



The Alternative?

1	The Assassin	Bridget Fonda
1	The Assassin	Gabriel Byrne
1	The Assassin	Harvey Keitel
2	It could happen to you	Brigette Fonda
2	It could happen to you	Nicholas Cage
2	It could happen to you	Rosie Perez
...
...
...
...

- ### KEYS: The Integrity Rules...
- In the relational model there are two integrity rules:
 - Entity Integrity:** No attribute forming part of the primary key of a base table is allowed to have NULL values.
 - Referential Integrity:** If table T_2 includes a foreign key FK matching the primary key PK of some base table T_1 , then every value of FK in T_2 must:
 - be equal to the value of the PK in some tuple of T_1 ; or
 - be wholly NULL, i.e. each attribute in that FK must be NULL.

- ### Normalisation
- So... how do I know how to structure my data in a database?
 - NORMALISATION
 - A formalism of simple ideas with a practical application in logical database schema design...
 - essentially that tables should not contain repeating groups.
 - No repetition means less chance of erroneous data being entered.
 - Many stages of the normalisation process, called normal forms
 - unf, 1st, 2nd, 3rd, ..., 7th...

- ### Why is Normalisation Necessary?
- Keeps data accurate by reducing redundancy (duplication).
 - Duplication may lead to errors as we will see.
 - Saves space by limiting the amount of redundant information stored.
 - Obvious that storing data many times is less desirable than storing data only once.
 - Reducing redundancy allows for faster processing of data.
 - Having to update data once is faster than many times.
 - Provides a framework within which design of Relational Databases should be structured.
 - Supports good DB design.
- Normalisation theory should allow us to recognise relations with undesirable properties, tell us what is "wrong" and how to "correct" it.*

Example of un-normalised data

Suppliers Table

SID	STATUS	CITY	PID	QTY
S1	20	London	P1	300
S1	20	London	P2	200
S1	20	London	P3	400
S1	20	London	P4	200
S1	20	London	P5	100
S1	20	London	P6	100
S2	10	Paris	P1	300
S2	10	Paris	P2	400
S3	10	Paris	P2	200
S4	20	London	P2	200
S4	20	London	P4	300
S4	20	London	P5	400

PK

3NF, minimum normalisation

Purchases = {SID,PID,Qty}

SID	PID	Qty
S1	P1	300
S1	P2	200
S1	P3	400
...		
S4	P5	300

SC = {SID, CITY}

SID	CITY
S1	London
S2	Paris
S3	Paris
S4	London
S5	Athens

CS = {CITY, STATUS}

CITY	STATUS
London	20
Paris	10
Athens	30

Alan Smeaton & Cathal Gurrin © 2001-2008 - 43 - DCU

ODBC

- ODBC, Open DataBase Connectivity is a standard that is used to enable applications to interact with different back-end DBMSs
- It is a wrapper around DBMSs that makes all databases operate in a clearly defined and consistent fashion.
- For example, the same code written could interact (assuming use of ODBC) with SQL Server, DB2, and Oracle databases without the programmer having to examine how each works.

Alan Smeaton & Cathal Gurrin © 2001-2008 - 44 - DCU

