

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.

Cathal Gurrin © 2003-2005 - 12 - 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

Cathal Gurrin © 2003-2005 - 13 - DCU

Another Example...

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

Cathal Gurrin © 2003-2005 - 14 - 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?

Cathal Gurrin © 2003-2005 - 15 - DCU

CA557

... more concepts ...

CA557 - INFORMATION ACCESS - 2004 - 2005 -

- **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.

Cathal Gurrin © 2003-2005 - 16 - DCU

CA557

What can we do with SQL

CA557 - INFORMATION ACCESS - 2004 - 2005 -

- **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

Cathal Gurrin © 2003-2005 - 17 - DCU

4.2 The Select Statement (a 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 !!!

Cathal Gurrin © 2003-2005
- 18 -
DCU

A Sample Database

```

    graph LR
      movie(movie) --- starsIn(starsIn)
      movie --- movieExec(movieExec)
      starsIn --- movieStar(movieStar)
      movieExec --- movieStar
      studio(studio) --- movie
  
```

SELECT * FROM movie WHERE year = 2004
 SELECT title, length FROM movie WHERE year = 2004

Cathal Gurrin © 2003-2005
- 19 -
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')`

Cathal Gurrin © 2003-2005 - 20 - 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;`

Cathal Gurrin © 2003-2005 - 21 - 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!

Cathal Gurrin © 2003-2005 - 22 - DCU

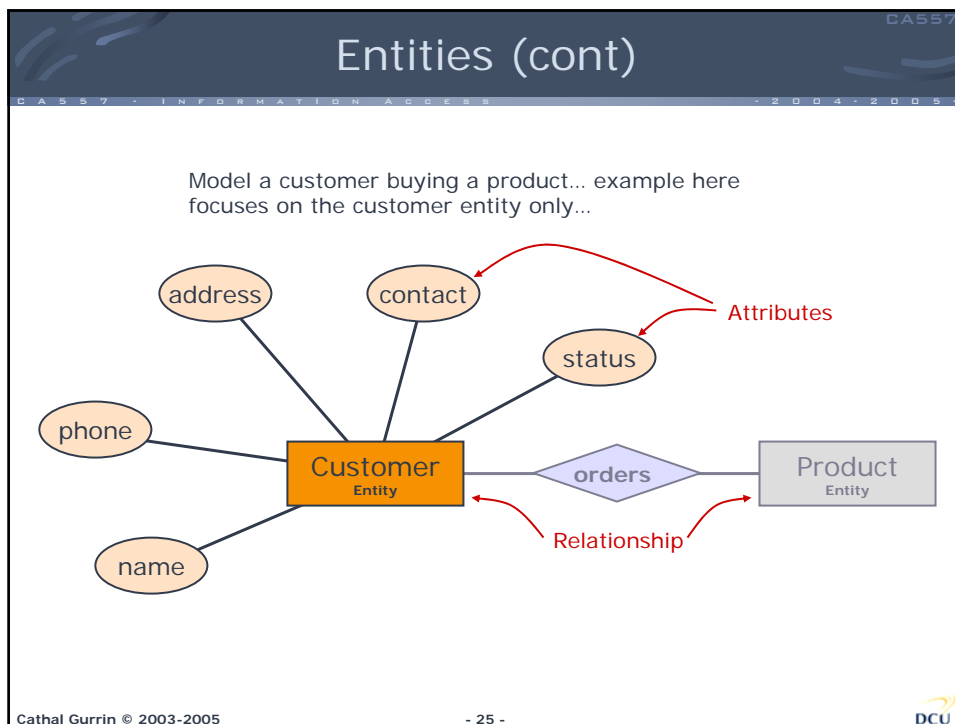
- OK, so compare this to a Search Engine...
 - In a Search Engine the data is a 'bag of words' on a web page.
 - Match bag of words against bags of words
 - In a DBMS, all the data is structured and we know what each piece of data is referring to.
 - If the WWW were a DBMS, then the result of a search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

Cathal Gurrin © 2003-2005 - 23 - 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.

Cathal Gurrin © 2003-2005 - 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 L1.10 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.

Cathal Gurrin © 2003-2005 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.

Cathal Gurrin © 2003-2005 DCU

A View

Underlying Table

1	2	3	4	5	6

A View over the Table

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

Cathal Gurrin © 2003-2005 DCU

Another View

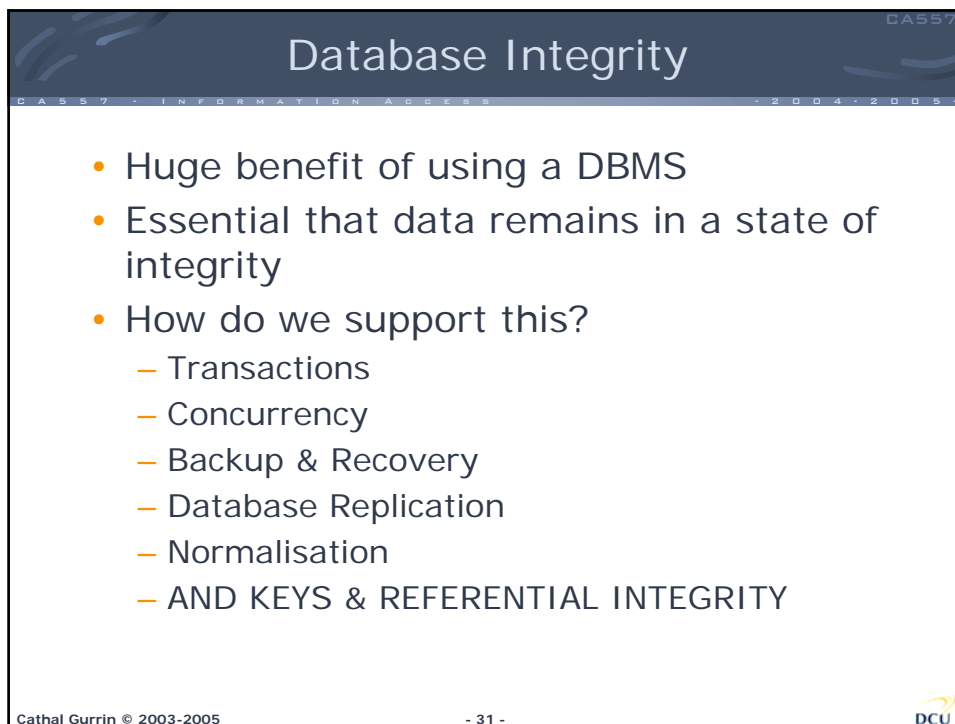
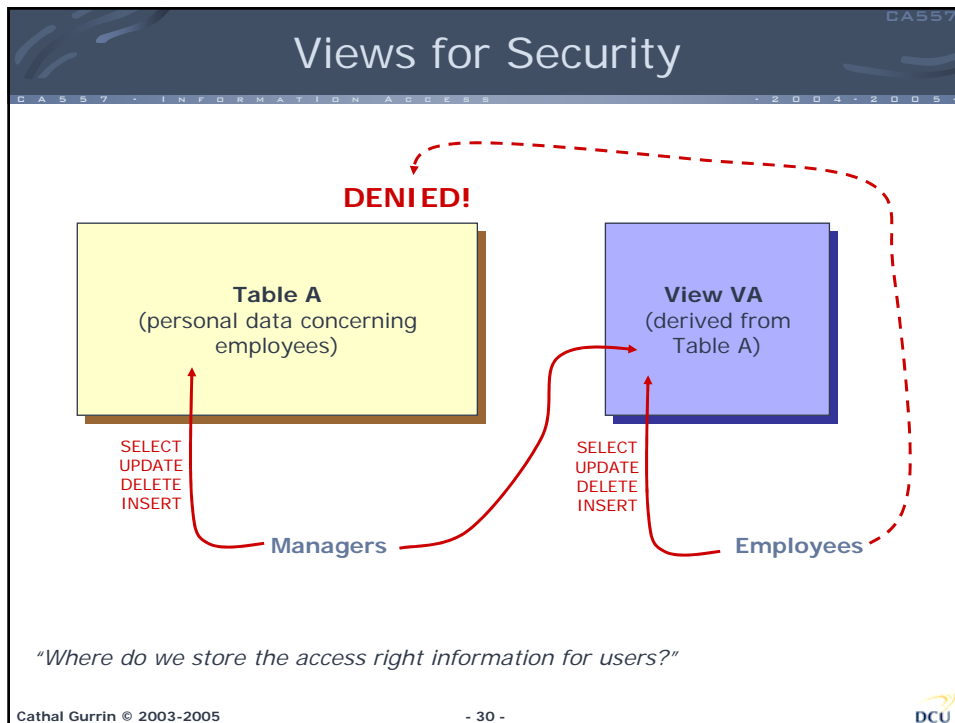
Underlying Table

1	2	3	4	5	6

A View over the Table

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

Cathal Gurrin © 2003-2005 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.

Cathal Gurrin © 2003-2005 - 32 - DCU

A Key is an Attribute

–EMPLOYEE : { fname, lname, age, room, phone, pps }

Cathal Gurrin © 2003-2005 - 33 - DCU

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

Cathal Gurrin © 2003-2005 - 34 - DCU

Compound Key

f_name	m_i	l_name	phone	age
Neil	G	Howie	7023212	42
Rowan	M	Morrison	7035622	29
Alder	E	MacGregor	3025214	48
Ash	G	Buchanan	1024877	26
Willow	R	MacGregor	3021264	25
..
..
Alder	T	MacGregor	3021268	29

A Unique Identifier?

Cathal Gurrin © 2003-2005 - 35 - DCU

Compound Key

f_name	m_i	l_name	phone	age
Neil	G	Howie	7023212	42
Rowan	M	Morrison	7035622	29
Alder	E	MacGregor	3025214	48
Ash	G	Buchanan	1024877	26
Willow	R	MacGregor	3021264	25
..
..
Alder	T	MacGregor	3021268	29

A Unique Identifier now?

Cathal Gurrin © 2003-2005 - 36 - DCU

Keys

1	The Assassin	1	1
2	It could happen to you ...	1	2
3	The Terminal	1	34
4	Alien ...	2	1
5	The Wicker Man	2	5
..	...	2	75
..	...	2	45
342	I-Robot	3	76
..
342	..	342	1

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

Cathal Gurrin © 2003-2005 - 37 - DCU

Keys

CA557 - INFORMATION ACCESS - 2004 - 2005 -

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

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

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

PK (compound) PK PK 2 x FKs

Cathal Gurrin © 2003-2005 - 38 - DCU

The Alternative?

CA557 - INFORMATION ACCESS - 2004 - 2005 -

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
...
...
...
...

Cathal Gurrin © 2003-2005 - 39 - DCU

KEYS: The Integrity Rules...

CA557 - INFORMATION ACCESS - 2004-2005

- 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.

Cathal Gurrin © 2003-2005 - 40 - DCU

Normalisation

CA557 - INFORMATION ACCESS - 2004-2005

- 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
 - 1nf, 2nf, 3nf, ..., 7th...

Cathal Gurrin © 2003-2005 - 41 - DCU

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}

<u>SID</u>	<u>PID</u>	Qty
S1	P1	300
S1	P2	200
S1	P3	400
...		
S4	P5	300

SC = {SID, CITY}

<u>SID</u>	<u>CITY</u>
S1	London
S2	Paris
S3	Paris
S4	London
S5	Athens

CS = {CITY, STATUS}

<u>CITY</u>	<u>STATUS</u>
London	20
Paris	10
Athens	30

Cathal Gurrin © 2003-2005 - 44 - 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.

Cathal Gurrin © 2003-2005 - 45 - DCU

