MONASH University
Information Technology

www.it.monash.edu

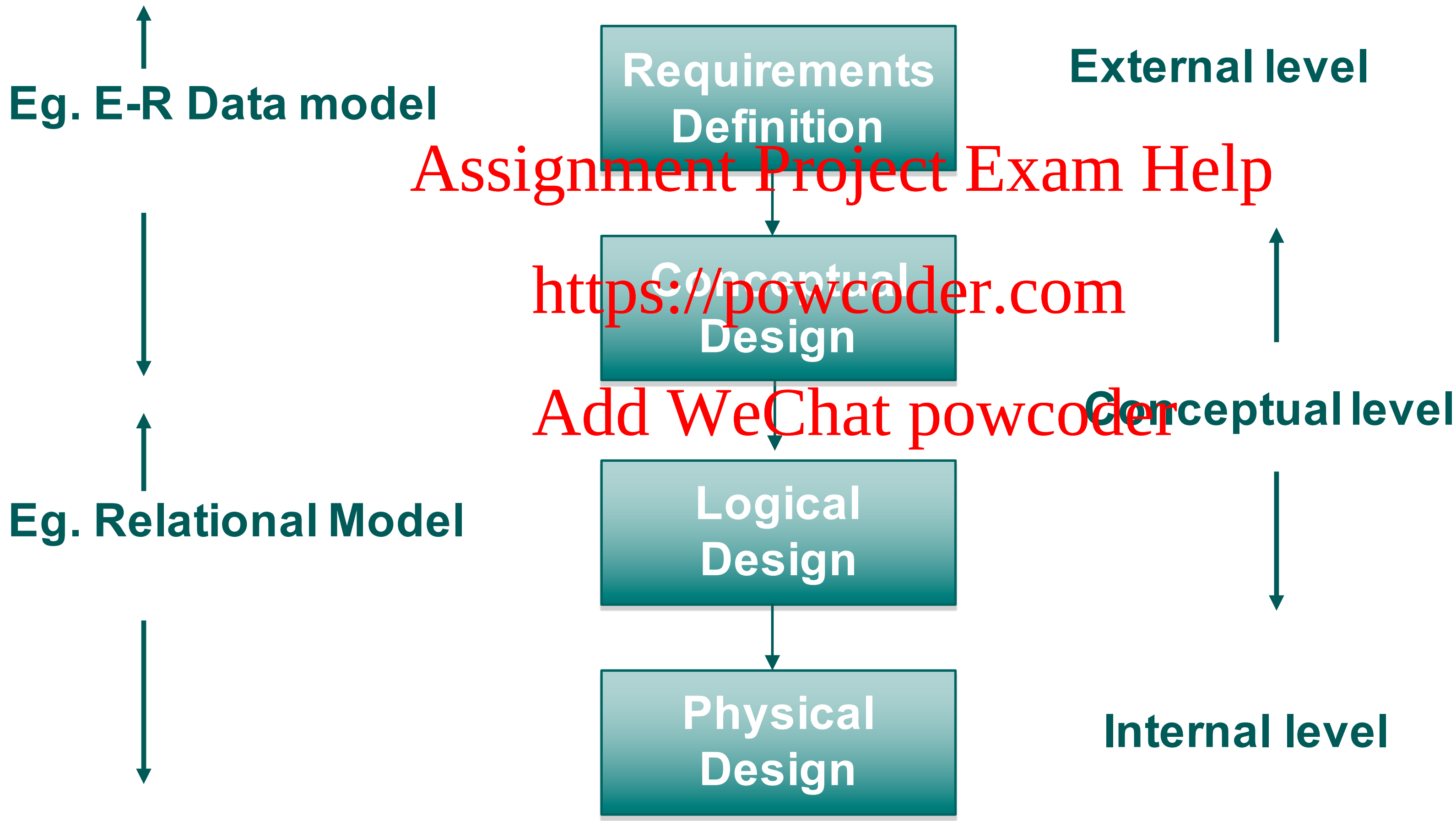
FIT9132 Introduction to Database

Normalisation of Relational Models

MARS code

HYT80D

Database Design Life Cycle - Recap



Properties of Relations - Recap

- Relations exhibit several fundamental properties:
  - Each row is unique - i.e. duplicate tuples are not allowed.
  - Each column has a (meaningful) name.
  - All the values in a column are values of a single attribute.
  - The order of attributes is immaterial.
  - The order of tuples is immaterial.
  - The entries are single-valued - each cell contains a single entry
  - Any value is addressable by specifying the name of the table, the primary key value for the relevant row, and the name of the column.

Relational Database

- A **relational database** is a collection of normalised relations.
  - Relational databases deal solely with named relations or two dimensional tables containing columns and rows.
  - Normalisation is part of the logical design phase of the relational database.

Example of a relational database:

```
order(order-id,order-date,order-total)
order-line(order-id,product-id,quantity)
product(product-id, description, unit-price)
```

Data Normalisation

- Relations should be normalised in order to avoid certain anomalies which may occur when updating data.
- Normalisation is a systematic series of techniques for progressively refining the data model.
- A formal approach to analysing relations based on their primary key (or candidate keys) and functional dependencies.
- A “treasure hunt” for hidden relations.

Sample Data

```
Sales(Cus-id,Name,Ord-no,O-date,
    Prod-id,Descrip,Qty-ord)
```


Cus-id	Name	Ord-no	O-date	Prod-id	Descrip	Qty-ord
C23	G.Gold	O56	15/4	P92	Chisel	6
C23	G.Gold	O56	15/4	P38	Plane	14
C75	R.Red	O57	16/4	P19	Saw	3
C19	B.Blue	O58	16/4	P33	Punch	24
C19	B.Blue	O58	16/4	P38	Plane	9
C23	G.Gold	O59	17/4	P19	Saw	10

What's Wrong with the Sales Relation?

- It contains . . .
  - Customer, Order and Product Data, as well as . . .
  - Sales Data
- Highly redundant
  - multiple recording of facts that:
    - G.Gold has Cus-id 23
    - Order O56 was lodged on 15 April
    - Product P19 is a Saw
  - wasteful as regards storage space
  - inconsistent data - multiple versions of data items

The Update Anomalies of the "Sales" File

- *Modification anomaly* - the need to make multiple changes when a single data item changes, e.g.
  - > G. Gold should have been H. Gold
- *Creation (or insertion) anomaly* - unnecessary delay adding new data, e.g.
  - > new customer who hasn't yet lodged an order,
  - > a product that hasn't yet been ordered
- *Deletion anomaly* - loss of data when deletions occur, e.g.
  - > loss of customer name when all orders for that customer have been filled, and the records are therefore deleted

Another Example

(McFadden et al, 1998)

- Consider a relation showing training courses which employees have enrolled in:

- TR-COURSE (Emp-ID, Course-ID, Fee)

- Assume,
  - each employee can enrol in many courses and
  - each course has a standard fee

  What is the primary Key?

Emp-ID	Course-ID	Fee
E130	C200	75
E200	C300	100
E250	C200	75
E245	C400	150
E500	C300	100
E575	C500	50

Redundancy – each course fee is repeated for each employee taking a course

If a new course (say C600) is offered, it cannot be added to the table until at least one employee enrolls in it
Creation anomaly

If employee E245 withdraws from course C400, all info for C400 is lost
Deletion anomaly

If the fee for C200 changes, multiple modifications must be made
Modification anomaly

MARS

Given a functional dependency A->B,
A is called the ________________.

Necessary Terminology

- Functional Dependence:
  - Is a relationship between two attributes
  - An attribute B is functionally dependent on an attribute A (written A→B) if each value of A is associated with exactly one value of B, i.e. if at a given point in time, the value of A determines the value of B.
  - Note that A and B may be composite attributes.
  - We can also say that A *determines* B, and that A is therefore a *determinant* of B.

e.g. Student-ID → Student-Name (Student-ID is a determinant)

Reverse is not true

Student-Name **–x->** Student-ID (students may have same name)

Normalisation

- Normalisation consists of a series of steps.
- Each step refines the data model further and makes update anomalies less likely.
- Formally we proceed through:
  - First Normal Form (1NF)
  - Second Normal Form (2NF)
  - Third Normal Form (3NF)
  - Boyce-Codd Normal Form (BCNF)
  - Fourth & Fifth Normal Form (4NF, 5NF)
  - Higher Normal Forms

First Normal Form (1NF)

- **Remove Repeating Groups ….**
- Every value in the relation is atomic (single-valued)

- First Normal Form is designed to eliminate repeating groups of attributes. Repeating groups may be indicated by { } or () in a table Note - this is not a relation by definition, it is more a collection of attributes or a so-called “un-normalised form”.


```
CUSTOMER(cus-id,surname,initials,
                    (order-no,order-date))
```

- The combination **order-no, order-date** may occur 0, 1 or m times for a given customer.

First Normal Form (1NF)

- Form a new relation from the repeating items plus the key of the original.


```
CUSTOMER(cus-id,surname,initials,
(order-no,order-date))
```


- Becomes:


```
CUSTOMER(cus-id,surname,initials)
ORDER(ord-no,cus-id,order-date)
```

- **The relations are now in 1NF.**
  - A relation in which the intersection of each row and column contains one and only one value.

Second Normal Form (2NF)

- **Remove Partial Dependencies ….**
- Second Normal Form aims to eliminate **partial key dependencies**.
- A partial key dependency arises when:
  - The primary key of the relation is composite AND
  - There exists a non-key attribute (or attributes) which is (are) functionally dependent on part of the primary key only.
- Partial key dependencies lead to update anomalies.
- IMPORTANT - a dependency A->B where A is ***only part of a candidate key*** and B is a non-key attribute is a partial dependency and should be removed.
- An attribute that forms part of any candidate key is termed a ***key attribute***. Other attributes are called ***non-key attributes*****.**

Second Normal Form (2NF)

Example of partial key dependency:

`ORDER-LINE(ord-no,prod-code,description,number-ordered)`

- In this relation, `prod-code` is all that is required to determine `description`. That is, we do not need to know the entire primary key in order to know the value of description. This is a partial key dependency.

- Changing a product description must happen in multiple places
- Relations are in 2NF when:
  - They are in 1NF AND
  - All partial key dependencies have been removed.

Second Normal Form (2NF)

- To convert the following to 2NF:

```
ORDER-LINE(ord-no,prod-code,description,
number-ordered)
```


Form a new relation from the items that do not require the full key:


```
ORDER-LINE(ord-no,prod-code,number-ordered)
PRODUCT(prod-code,description)
```


- Now the relations are in 2NF.

A relation that is in first normal form and every non-key attribute is fully functionally dependent on the entire primary (or candidate) key.

Third Normal Form (3NF)

- **Remove Transitive Dependencies ….**
- Third Normal Form aims to eliminate transitive dependencies.

- A transitive dependency arises when:
  - An attribute in a relation might be more immediately identified by another non-key attribute rather than by the primary/candidate key.
- Transitive key dependencies lead to update anomalies.

Third Normal Form (3NF)

- Example of transitive key dependency:

ORDER(<u>ord-no</u>,ord-date,cus-id,name,street,suburb)

name, street and suburb are only transitively, (through cus-id), identified by ord-no.

- To convert to 3NF:

Form a new relation from the items identified by the non-key identifier:

ORDER(<u>ord-no</u>,ord-date,cus-id)
CUSTOMER(<u>cus-id</u>,name,street,suburb)

Third Normal Form (3NF)

- Relations are in 3NF when:
  - They are in 2NF

  AND
  - All transitive key dependencies have been removed.

Higher Normal Forms

- Often, converting the relations to 3NF is sufficient to remove update anomalies and proceed with the implementation phase. HOWEVER, there are certain circumstances where 3NF is not sufficient and we may need to further normalise the data.
- Higher normal forms include:
  - Boyce-Codd Normal Form (an alternative 3NF)
  - Fourth Normal Form
  - Fifth Normal Form

Boyce-Codd Normal Form

- A relation is in BCNF if and only if every determinant is a candidate key.
- If there is only one candidate key, BCNF and 3NF are equivalent.
- BCNF may be considered a stronger form of 3NF.

Fourth and Fifth Normal Form

- 4NF: No nontrivial multi-valued dependencies
- 5NF: No join dependencies

- We will not cover these normal forms.
- The situations that they cover are not as frequent.
- (Covered in text books ….)

MONASH University
Information Technology

GO Boldly.

PROJ_NUM	PROJ_NAME	EMP_NUM	EMP_NAME	JOB_CLASS	CHG_HOUR	HOURS
15	Evergreen	103	June E. Arbough	Elect. Engineer	84.50	23.8
		101	John G. News	Database Designer	105.00	19.4
		105	Alice K. Johnson *	Database Designer	105.00	35.7
		106	William Smithfield	Programmer	35.75	12.6
		102	David H. Senior	Systems Analyst	96.75	23.8
18	Amber Wave	114	Annelise Jones	Applications Designer	48.10	24.6
		118	James J. Frommer	General Support	18.36	45.3
		104	Anne K. Ramoras *	Systems Analyst	96.75	32.4
		112	Darlene M. Smithson	DSS Analyst	45.95	44.0
22	Rolling Tide	105	Alice K. Johnson	Database Designer	105.00	64.7
		104	Anne K. Ramoras	Systems Analyst	96.75	48.4
		113	Delbert K. Joenbrood *	Applications Designer	48.10	23.6
		111	Geoff B. Wabash	Clerical Support	26.87	22.0
		106	William Smithfield	Programmer	35.75	12.8
25	Starflight	107	Maria D. Alonzo	Programmer	35.75	24.6
		115	Travis B. Bawangi	Systems Analyst	96.75	45.8
		101	John G. News *	Database Designer	105.00	56.3
		114	Annelise Jones	Applications Designer	48.10	33.1
		108	Ralph B. Washington	Systems Analyst	96.75	23.6
		118	James J. Frommer	General Support	18.36	30.5
		112	Darlene M. Smithson	DSS Analyst	45.95	41.4

MARS

If there are no repeating groups in a relation and all the determinants in the relation are candidate keys, then the relation is in

A.1NF

B.2NF

C.3NF

D.All of the above.

MARS

Consider a relation R(A,B,C,D).
The composite primary key of the relation is (A,B).
There is another candidate key (B,C).

If there exists a functional dependency B->C,
this functional dependency is said to be a _________ dependency

A. Partial

** Can (B,C) really be a candidate key?*

B. Prime

C. Transitive

D. None of the above

MARS

Consider a relation R(A,B,C).
The composite primary key of the relation is (A,B).
(A,B) is the only candidate key.

Suppose also that there is a functional dependency B->C.

This is a(n) __________ dependency.

MARS

Which of the following is false?

A. A relation with a single non-composite candidate key with no repeating groups is guaranteed to be in second normal form.

B. Foreign keys must have the same name as the primary key in another relation.

C. Transitive dependencies can lead to modification anomalies.

D. A relation in third normal form is guaranteed to be in first normal form.