kdeg
knowledge and data engineering group

Information Management and Data Engineering

CS4D2a – 4CSLL1 – CS3041

Introduction to Design Optimisation

Séamus Lawless

seamus.lawless@scss.tcd.ie

Database Design

- In previous lectures we have discussed various aspects of database design
  - The formal Relational Model
  - The Entity Relationship Model
    - Entity Relationship Diagrams
  - Mapping to a Logical Database Design
    - Relational Database Schema

Database Design

- Need of a formal method for analysing how the relations and attributes are grouped
- A measure of appropriateness or goodness, other than the intuition of the designer
  - To assess the quality of the design
- Measures
  - Design guidelines
  - Functional Dependencies
  - Normalisation

Design Guidelines

- A set of informal guidelines
  - Can be used as measures to determine the quality of a relation schema design
    - Attribute Semantics
    - Reduction of Redundancy
    - Reduction of NULLs
    - Generation of Spurious Tuples
- These measures are not always independent of one another

Attribute Semantics

- Attributes belonging to a relation have certain real-world meaning
- Semantics of a relation
  - refers to its meaning resulting from the interpretation of attribute values in a tuple
- Careful entity relationship modeling and accurate mapping to logical design help to ensure that a relational schema design has clear meaning

Guideline 1

- Design a relation schema so that it is easy to explain its meaning
- Give relations and attributes meaningful names
- Do not combine attributes from multiple entity types and relationship types into a single relation
  - Straightforward to interpret
  - Easy to explain its meaning

Violating Guideline 1

EMP_DEPT

Ename	Ssn	Bdate	Address	Dnumber	Dname	Dmgr_ssn

EMP_PROJ

Ssn	Pnumber	Hours	Ename	Pname	Plocation

- These relations violate guideline 1 by mixing attributes from distinct real-world entities
  - Employee and Department
  - Project and Employee

Reduction of Redundancy

- One goal of database schema design is to minimise the storage space used
- Grouping attributes into relation schemas has a significant effect on storage space

EMPLOYEE

Ename	Ssn	Bdate	Address	Dnumber

DEPARTMENT

Dname	Dnumber	Mgr_ssn

versus

EMP_DEPT

Ename	Ssn	Bdate	Address	Dnumber	Dname	Dmgr_ssn

Reduction of Redundancy

- Storing merged entities in single relations leads to another problem, *update anomalies*
- Update anomalies can be classified into:
  - Insertion anomalies
  - Deletion anomalies
  - Modification anomalies

Insertion Anomalies

EMP_DEPT

Ename	Ssn	Bdate	Address	Dnumber	Dname	Dmgr_ssn

- To insert a new employee into EMP_DEPT it is necessary to include either:
  - all attribute values for the department that the employee works for
  - NULLs, if the employee is not yet assigned
- Consistency becomes an issue
- Inserting a new department is difficult

Deletion Anomalies

EMP_DEPT

Ename	Ssn	Bdate	Address	Dnumber	Dname	Dmgr_ssn

- Deletion of Employees and Departments inextricably linked
  - If we delete the last employee currently assigned to a particular department, the information related to department is lost from the database
- This problem does not occur if using separate relations

Modification Anomalies

EMP_DEPT

Ename	Ssn	Bdate	Address	Dnumber	Dname	Dmgr_ssn

- Modification makes consistency an issue
- If the manager of a department is changed
  - It is necessary to update the tuples of every employee who works for that department
  - Records can easily get out of sync
- This problem does not occur if using separate relations

Guideline 2

- Design the relation schemas so that no insertion, deletion or modification anomalies are present
  - if anomalies are present, note them clearly and ensure all application programs operate correctly
- This second guideline is consistent with guideline 1

Reduction of NULLs

- If many attributes do not apply to all the tuples of a relation, you end up with many NULL values in those tuples
  - Waste storage space
  - Can make understanding attribute meanings more difficult
  - Leads to difficulty with Joins
  - Difficulty with aggregate functions
    - COUNT and SUM

Reduction of NULLs

- A NULL value may typically have two interpretations
  - missing but inapplicable
    - Zip Code for Irish Addresses (although we do now have EirCode!)
  - missing but applicable
    - an employees date of birth is empty
      - unknown
      - known but absent

Guideline 3

- Avoid placing attributes in a relation schema whose values may frequently be NULL
  - If NULLs are unavoidable, ensure they apply in exceptional cases and not the majority of tuples
- Using space efficiently and avoiding joins on NULL values are the main criteria for deciding upon attribute inclusion or exclusion
  - if excluded, create a separate relation for that attribute

Violating Guideline 3

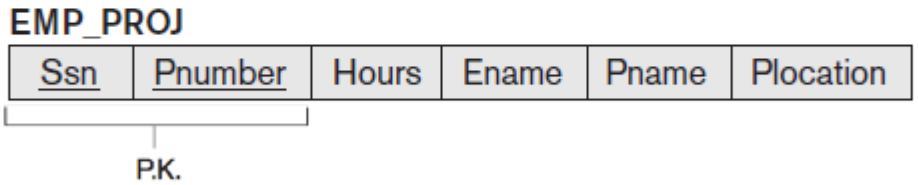
- If only 15% of employees have an office, then including an Office_Number attribute in the EMPLOYEE relation would violate guideline 3
  - Instead, create an EMPLOYEE_OFFICE relation
  - This could contain the attributes
    - Ssn, Office_Number
  - A tuple is entered in the relation for all employees with an office

Generation of Spurious Tuples

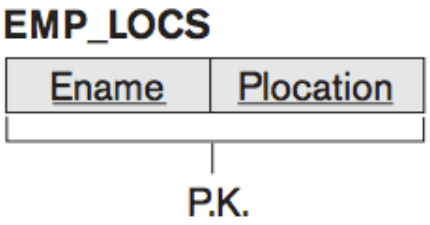
- Joins across relations should only be performed on *Primary Key – Foreign Key* pairs of attributes
- If joins are performed on attributes which are not a Primary Key – Foreign Key pairing, spurious tuples are generated as a result
  - These tuples represent information which is not valid

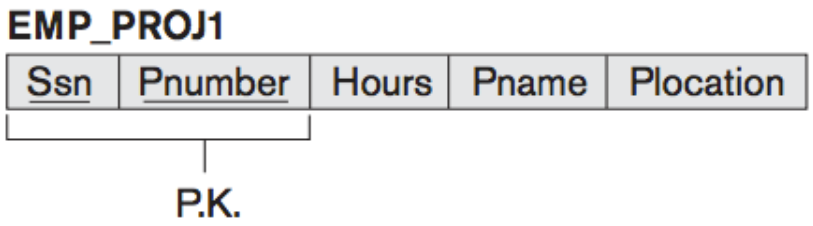
Generation of Spurious Tuples

- Suppose, instead of the EMP_PROJ relation



- We had defined two separate relations





Generation of Spurious Tuples

- The original relation contained the following:

EMP_PROJ

Ssn	Pnumber	Hours	Ename	Pname	Plocation
123456789	1	32.5	Smith, John B.	ProductX	Bellaire
123456789	2	7.5	Smith, John B.	ProductY	Sugarland
666884444	3	40.0	Narayan, Ramesh K.	ProductZ	Houston
453453453	1	20.0	English, Joyce A.	ProductX	Bellaire
453453453	2	20.0	English, Joyce A.	ProductY	Sugarland
333445555	2	10.0	Wong, Franklin T.	ProductY	Sugarland
333445555	3	10.0	Wong, Franklin T.	ProductZ	Houston
333445555	10	10.0	Wong, Franklin T.	Computerization	Stafford
333445555	20	10.0	Wong, Franklin T.	Reorganization	Houston
999887777	30	30.0	Zelaya, Alicia J.	Newbenefits	Stafford
999887777	10	10.0	Zelaya, Alicia J.	Computerization	Stafford
987987987	10	35.0	Jabbar, Ahmad V.	Computerization	Stafford
987987987	30	5.0	Jabbar, Ahmad V.	Newbenefits	Stafford
987654321	30	20.0	Wallace, Jennifer S.	Newbenefits	Stafford
987654321	20	15.0	Wallace, Jennifer S.	Reorganization	Houston
888665555	20	Null	Borg, James E.	Reorganization	Houston

Generation of Spurious Tuples

- In the two relation version, this becomes

EMP_LOCS

Ename	Plocation
Smith, John B.	Bellaire
Smith, John B.	Sugarland
Narayan, Ramesh K.	Houston
English, Joyce A.	Bellaire
English, Joyce A.	Sugarland
Wong, Franklin T.	Sugarland
Wong, Franklin T.	Houston
Wong, Franklin T.	Stafford
Zelaya, Alicia J.	Stafford
Jabbar, Ahmad V.	Stafford
Wallace, Jennifer S.	Stafford
Wallace, Jennifer S.	Houston
Borg, James E.	Houston

EMP_PROJ1

Ssn	Pnumber	Hours	Pname	Plocation
123456789	1	32.5	ProductX	Bellaire
123456789	2	7.5	ProductY	Sugarland
666884444	3	40.0	ProductZ	Houston
453453453	1	20.0	ProductX	Bellaire
453453453	2	20.0	ProductY	Sugarland
333445555	2	10.0	ProductY	Sugarland
333445555	3	10.0	ProductZ	Houston
333445555	10	10.0	Computerization	Stafford
333445555	20	10.0	Reorganization	Houston
999887777	30	30.0	Newbenefits	Stafford
999887777	10	10.0	Computerization	Stafford
987987987	10	35.0	Computerization	Stafford
987987987	30	5.0	Newbenefits	Stafford
987654321	30	20.0	Newbenefits	Stafford
987654321	20	15.0	Reorganization	Houston
888665555	20	NULL	Reorganization	Houston

Generation of Spurious Tuples

- If we do a join on the two relations:

	Ssn	Pnumber	Hours	Pname	Plocation	Ename
	123456789	1	32.5	ProductX	Bellaire	Smith, John B.
*	123456789	1	32.5	ProductX	Bellaire	English, Joyce A.
	123456789	2	7.5	ProductY	Sugarland	Smith, John B.
*	123456789	2	7.5	ProductY	Sugarland	English, Joyce A.
*	123456789	2	7.5	ProductY	Sugarland	Wong, Franklin T.
	666884444	3	40.0	ProductZ	Houston	Narayan, Ramesh K.
*	666884444	3	40.0	ProductZ	Houston	Wong, Franklin T.
*	453453453	1	20.0	ProductX	Bellaire	Smith, John B.
	453453453	1	20.0	ProductX	Bellaire	English, Joyce A.
*	453453453	2	20.0	ProductY	Sugarland	Smith, John B.
	453453453	2	20.0	ProductY	Sugarland	English, Joyce A.
*	453453453	2	20.0	ProductY	Sugarland	Wong, Franklin T.
*	333445555	2	10.0	ProductY	Sugarland	Smith, John B.
*	333445555	2	10.0	ProductY	Sugarland	English, Joyce A.
	333445555	2	10.0	ProductY	Sugarland	Wong, Franklin T.
*	333445555	3	10.0	ProductZ	Houston	Narayan, Ramesh K.
	333445555	3	10.0	ProductZ	Houston	Wong, Franklin T.
	333445555	10	10.0	Computerization	Stafford	Wong, Franklin T.
*	333445555	20	10.0	Reorganization	Houston	Narayan, Ramesh K.
	333445555	20	10.0	Reorganization	Houston	Wong, Franklin T.

* * *

Guideline 4

- Design relation schemas so that they can be joined using equality conditions on primary key, foreign key pairs
  - This guarantees that no spurious tuples are generated by the join
- Avoid relations that contain matching attributes that are not foreign key, primary key combinations

Design Guidelines

- Informal measures used to determine the quality of a relational schema design
  - Ensure that attribute semantics are easily understood
  - Reduce the redundant information in tuples
  - Reduce the number of NULL values in tuples
  - Ensure that spurious tuples are not generated by enforcing primary key, foreign key matching