

Database Management Systems

Lecture



Normalization





- Normalization and Normal Forms
- 1st Normal Form
- 2nd Normal Form
- 3rd Normal Form
- Boyce-Codd Normal Form
- 4th Normal Form
- 5th Normal Form
- Normalization review



Normalization

- The process of analyzing given relation schemas based on their FDs and keys to:
 - Minimize redundancy
 - Minimize insertion, deletion and modification anomalies
- Normalization aims at decomposing a relation into smaller relations which have the above desired properties
- A top-down approach
 - Design by analysis



Normal forms

- The process of normalization takes a given relation through steps
- At each step a certain test is performed to check if the relation satisfies certain conditions
 - If it does, then the relation is said to be in a certain Normal Form
- There are a total of 6 Normal forms
 - 1NF, 2NF, 3NF, BCNF (Boyce-Codd), 4NF, 5NF
- Each normal form is stricter than the previous
 - For a relation to be in a certain Normal Form, it must already be in the previous Normal Form
- The normal form of a relation refers to the highest NF condition it meets



Normal Forms

All relations (normalized and non-normalized)

1NF relations

2NF relations

3NF relations

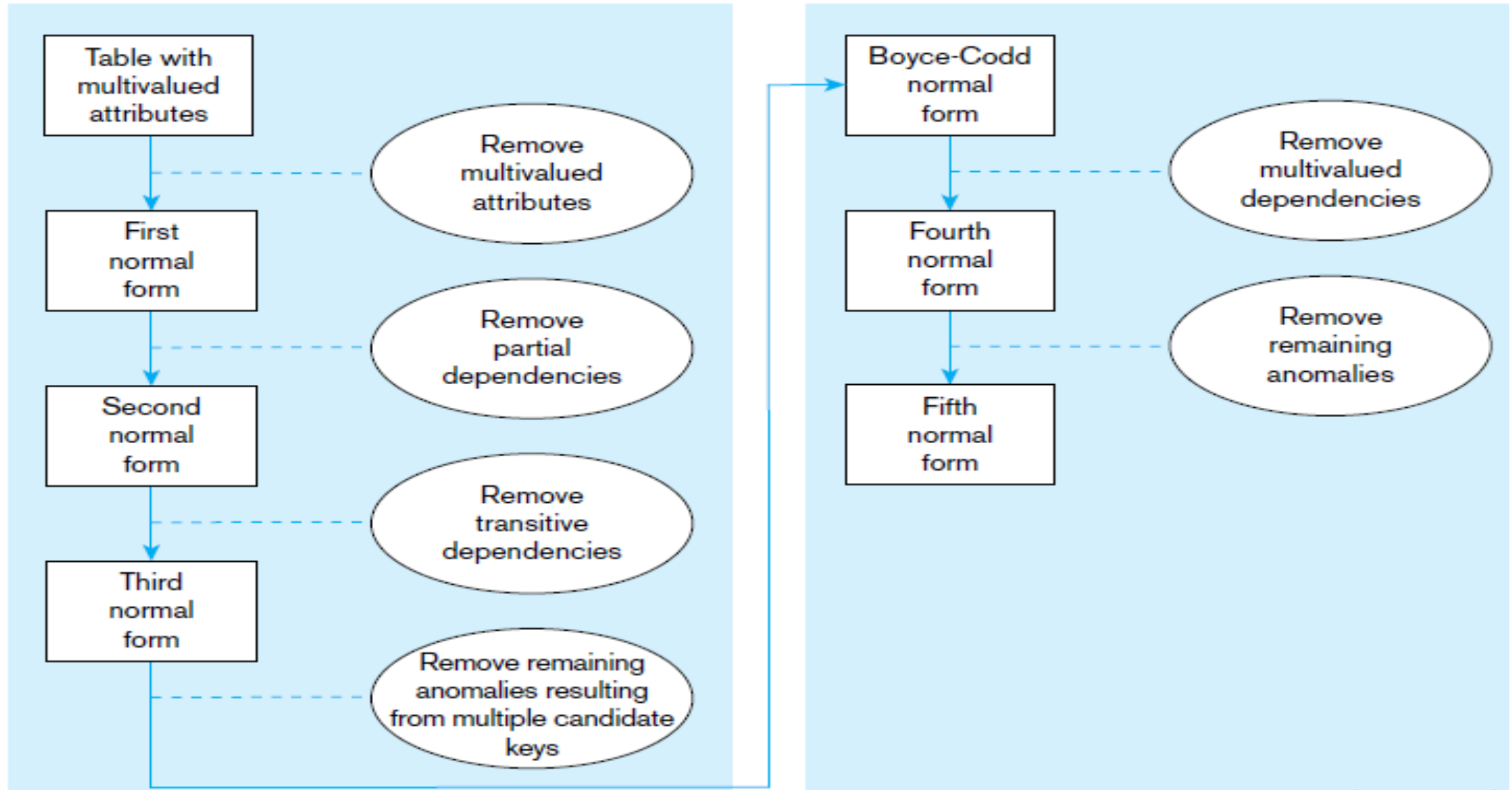
BCNF relations

4NF relations

5NF relations



Steps in Normalization





- A relation is said to be in 1NF if all of its attribute values are atomic
- That is, no cell of the relation can have multiple values, or a set of tuples (nested relation) as a value
- A relation that is NOT in 1NF is a non-normalized relation



DEPARTMENT

A non-normalized relation

Dname	<u>Dnumber</u>	Dmgr_ssn	Dlocations
Research	5	333445555	{Bellaire, Sugarland, Houston}
Administration	4	987654321	{Stafford}
Headquarters	1	888665555	{Houston}

How to convert in into 1NF? Three possible ways...



DEPARTMENT

A non-normalized relation

Dname	<u>Dnumber</u>	Dmgr_ssn	Dlocations
Research	5	333445555	{Bellaire, Sugarland, Houston}
Administration	4	987654321	{Stafford}
Headquarters	1	888665555	{Houston}

Method # 1: Repeat all rows with only single values from the multi-valued attribute

DEPARTMENT

Dname	<u>Dnumber</u>	Dmgr_ssn	<u>Dlocation</u>
Research	5	333445555	Bellaire
Research	5	333445555	Sugarland
Research	5	333445555	Houston
Administration	4	987654321	Stafford
Headquarters	1	888665555	Houston

Is this a good solution?

If not, then why?

Our typical issue: redundancy!



1NF

DEPARTMENT

A non-normalized relation

Dname	<u>Dnumber</u>	Dmgr_ssn	Dlocations
Research	5	333445555	{Bellaire, Sugarland, Houston}
Administration	4	987654321	{Stafford}
Headquarters	1	888665555	{Houston}

Method # 2: Add multiple attributes for each possible value of the multi-valued attribute

DEPARTMENT

Dname	<u>Dnumber</u>	Dmgr_ssn	<u>Dlocation1</u>	<u>Dlocation2</u>	<u>Dlocation3</u>
Research	5	333445555	Bellaire	Sugarland	Houston
Administration	4	987654321	Stafford	NULL	NULL
Headquarters	1	888665555	Houston	NULL	NULL

Is this a good solution? If not, then why? Too many NULLs...



DEPARTMENT

A non-normalized relation

Dname	<u>Dnumber</u>	Dmgr_ssn	Dlocations
Research	5	333445555	{Bellaire, Sugarland, Houston}
Administration	4	987654321	{Stafford}
Headquarters	1	888665555	{Houston}

Method # 3: Remove the multi-valued attribute and add to a new relation with the PK of this relation

DEPT_LOCATIONS

<u>Dnumber</u>	<u>Dlocation</u>
1	Houston
4	Stafford
5	Bellaire
5	Sugarland
5	Houston

Is this a good solution?

Yes, it's a great solution!

FK from DEPT relation and Dlocation combined will be PK



1NF

EMP_PROJ

Ssn	Ename	Pnumber	Hours
123456789	Smith, John B.	1	32.5
		2	7.5
666884444	Narayan, Ramesh K.	3	40.0
453453453	English, Joyce A.	1	20.0
		2	20.0
333445555	Wong, Franklin T.	2	10.0
		3	10.0
		10	10.0
		20	10.0

A non-normalized relation

Relation inside a relation!
(nested relation)

How to convert this into 1NF? Decompose it into two relations...

EMP_PROJ1

<u>Ssn</u>	Ename
------------	-------

EMP_PROJ2

<u>Ssn</u>	<u>Pnumber</u>	Hours
------------	----------------	-------



- A relation R is in 2NF if
 - I. R is in 1NF, and
 - II. Every nonprime attribute A in R is fully functionally dependent on the primary key of R


- Note: A relation which is in 1NF and which only has simple primary key automatically falls in 2NF...



2NF

Non-prime attributes

EMP_PROJ



<u>Ssn</u>	<u>Pnumber</u>	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

(Ssn, Pnumber) → Hours



2NF

Non-prime attributes



EMP_PROJ

<u>Ssn</u>	<u>Pnumber</u>	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

(Ssn, Pnumber) → Hours

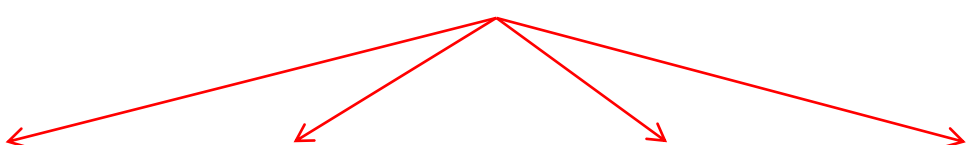
(Ssn, Pnumber) → Ename



2NF

Non-prime attributes

EMP_PROJ



<u>Ssn</u>	<u>Pnumber</u>	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

(Ssn, Pnumber) → Hours

(Ssn, Pnumber) → Ename



2NF

Non-prime attributes

EMP_PROJ

<u>Ssn</u>	<u>Pnumber</u>	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

(Ssn, Pnumber) → Hours

(Ssn, Pnumber) → Ename

(Ssn, Pnumber) → {Pname, Plocation}



2NF

Non-prime attributes

EMP_PROJ

<u>Ssn</u>	<u>Pnumber</u>	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

(Ssn, Pnumber) → Hours

(Ssn, Pnumber) → Ename

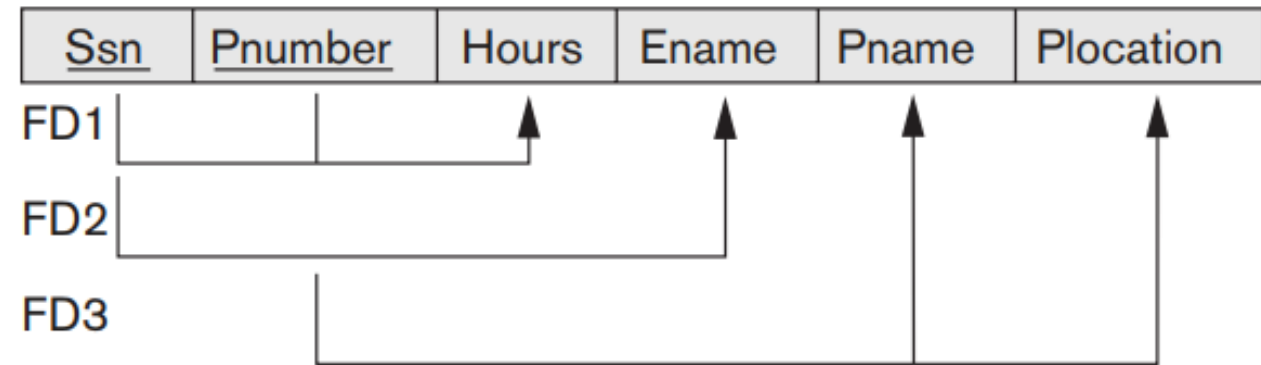
(Ssn, Pnumber) → {Pname, Plocation}



2NF

- A relation R is in 2NF if
 - I. R is in 1NF, and
 - II. Every nonprime attribute A in R is fully functionally dependent on the primary key of R
- FDs in EMP_PROJ
 - $(Ssn, Pnumber) \rightarrow Hours$
 - $Ssn \rightarrow Ename$
 - $Pnumber \rightarrow \{Pname, Plocation\}$
- Primary key is $\{Ssn, Pnumber\}$
- This relation is NOT in 2NF. Why?
 - Because Ename is a non-prime attribute, and it is not fully functionally dependent on the primary key...
 - It is ONLY dependent on Ssn!

EMP_PROJ

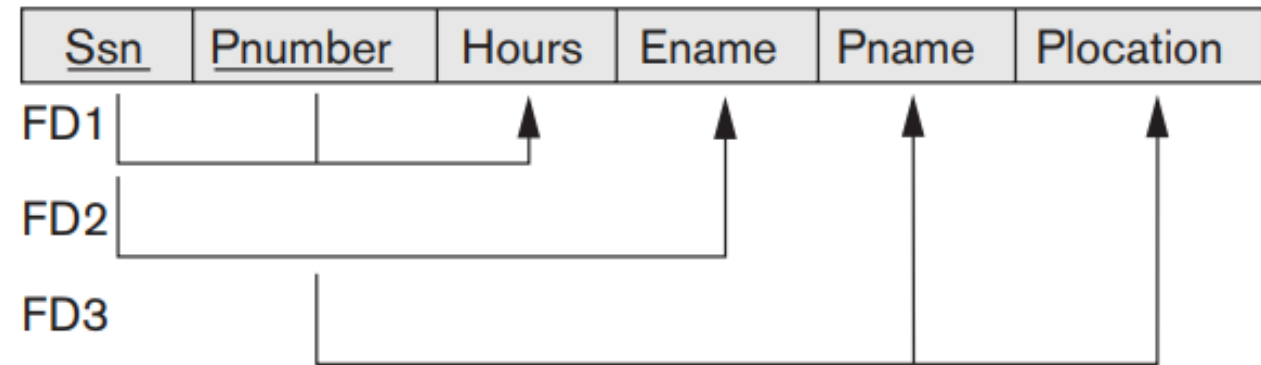




2NF

- A relation R is in 2NF if
 - I. R is in 1NF, and
 - II. Every nonprime attribute A in R is fully functionally dependent on the primary key of R
- FDs in EMP_PROJ
 - $(Ssn, Pnumber) \rightarrow Hours$
 - $Ssn \rightarrow Ename$
 - $Pnumber \rightarrow \{Pname, Plocation\}$
- Primary key is $\{Ssn, Pnumber\}$
- This relation is NOT in 2NF. Why?
 - Similarly, Pname and Plocation are only dependent on Pnumber...

EMP_PROJ

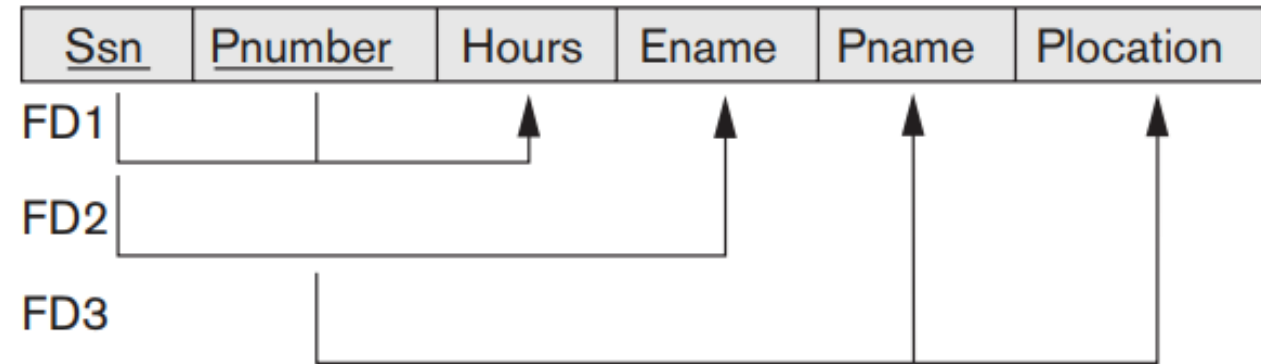




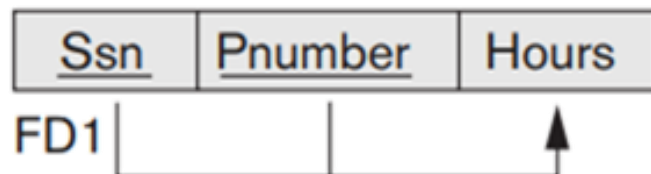
2NF

- So how to convert it into 2NF?
- Create separate relations for non-prime attributes and include only that part of primary key on which they are fully dependent

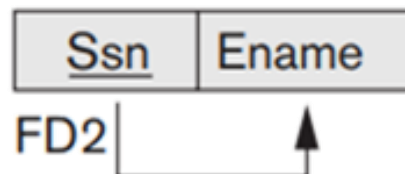
EMP_PROJ



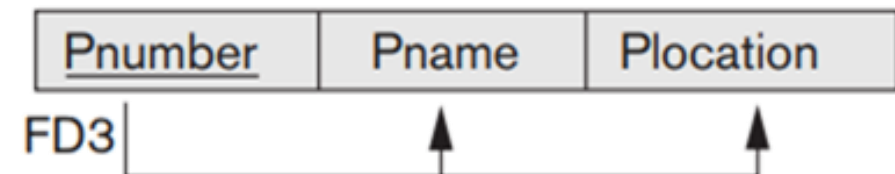
EP1



EP2



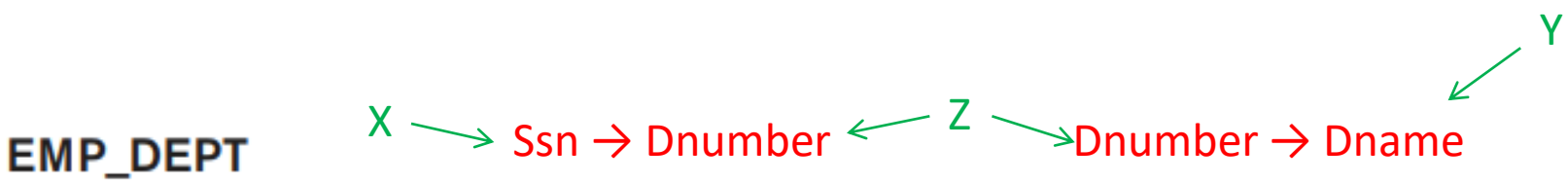
EP3





3NF

- Transitive dependency: A FD $X \rightarrow Y$ is transitive if there exists a set of attributes Z which are not candidate key nor part of a candidate key, and $X \rightarrow Z$ and $Z \rightarrow Y$



Ename	<u>Ssn</u>	Bdate	Address	Dnumber	Dname	Dmgr_ssn
Smith, John B.	123456789	1965-01-09	731 Fondren, Houston, TX	5	Research	333445555
Wong, Franklin T.	333445555	1955-12-08	638 Voss, Houston, TX	5	Research	333445555
Zelaya, Alicia J.	999887777	1968-07-19	3321 Castle, Spring, TX	4	Administration	987654321
Wallace, Jennifer S.	987654321	1941-06-20	291 Berry, Bellaire, TX	4	Administration	987654321
Narayan, Ramesh K.	666884444	1962-09-15	975 FireOak, Humble, TX	5	Research	333445555
English, Joyce A.	453453453	1972-07-31	5631 Rice, Houston, TX	5	Research	333445555
Jabbar, Ahmad V.	987987987	1969-03-29	980 Dallas, Houston, TX	4	Administration	987654321
Borg, James E.	888665555	1937-11-10	450 Stone, Houston, TX	1	Headquarters	888665555



3NF

- A relation R is in 3NF if
 - I. R is in 2NF, and
 - II. No non-prime attribute is transitively dependent on the primary key

EMP_DEPT

Ename	<u>Ssn</u>	Bdate	Address	Dnumber	Dname	Dmgr_ssn
Smith, John B.	123456789	1965-01-09	731 Fondren, Houston, TX	5	Research	333445555
Wong, Franklin T.	333445555	1955-12-08	638 Voss, Houston, TX	5	Research	333445555
Zelaya, Alicia J.	999887777	1968-07-19	3321 Castle, Spring, TX	4	Administration	987654321
Wallace, Jennifer S.	987654321	1941-06-20	291 Berry, Bellaire, TX	4	Administration	987654321
Narayan, Ramesh K.	666884444	1962-09-15	975 FireOak, Humble, TX	5	Research	333445555
English, Joyce A.	453453453	1972-07-31	5631 Rice, Houston, TX	5	Research	333445555
Jabbar, Ahmad V.	987987987	1969-03-29	980 Dallas, Houston, TX	4	Administration	987654321
Borg, James E.	888665555	1937-11-10	450 Stone, Houston, TX	1	Headquarters	888665555



3NF

- As seen on previous slide, this relation is not in 3NF
- So how do we convert it into 3NF?

EMP_DEPT

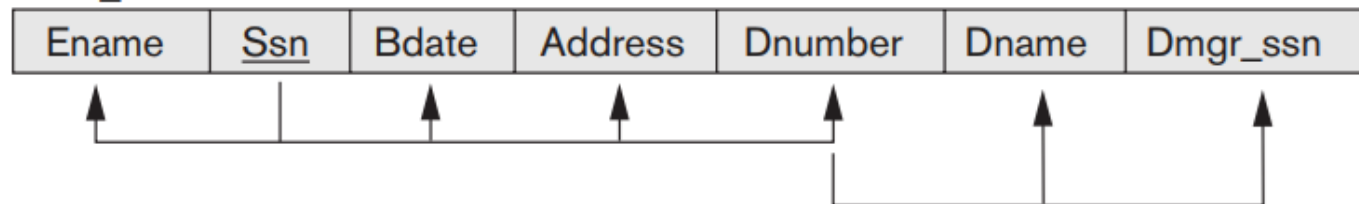
Ename	<u>Ssn</u>	Bdate	Address	Dnumber	Dname	Dmgr_ssn
Smith, John B.	123456789	1965-01-09	731 Fondren, Houston, TX	5	Research	333445555
Wong, Franklin T.	333445555	1955-12-08	638 Voss, Houston, TX	5	Research	333445555
Zelaya, Alicia J.	999887777	1968-07-19	3321 Castle, Spring, TX	4	Administration	987654321
Wallace, Jennifer S.	987654321	1941-06-20	291 Berry, Bellaire, TX	4	Administration	987654321
Narayan, Ramesh K.	666884444	1962-09-15	975 FireOak, Humble, TX	5	Research	333445555
English, Joyce A.	453453453	1972-07-31	5631 Rice, Houston, TX	5	Research	333445555
Jabbar, Ahmad V.	987987987	1969-03-29	980 Dallas, Houston, TX	4	Administration	987654321
Borg, James E.	888665555	1937-11-10	450 Stone, Houston, TX	1	Headquarters	888665555



3NF

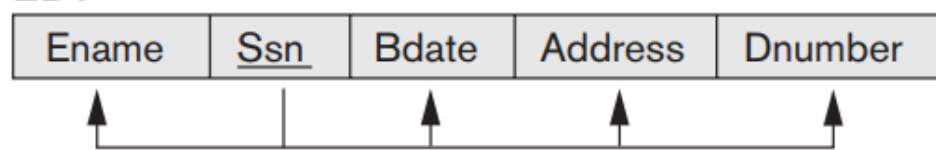
- As seen on previous slide, this relation is not in 3NF
- So how do we convert it into 3NF?
- Decompose the relation into 3NF relations by removing the transitivity
 - For a FD $X \rightarrow Z \rightarrow Y$, make separate relations for $X \rightarrow Z$ and $Z \rightarrow Y$

EMP_DEPT

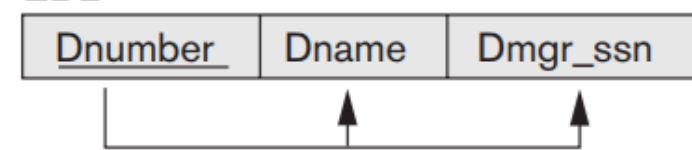


3NF Normalization

ED1



ED2





3NF revisited

- A relation R is in 3NF if
 - i. R is in 2NF, and
 - ii. No non-prime attribute is transitively dependent on the primary key

$Ssn \rightarrow Dnumber$

$Dnumber \rightarrow Dname$ ← The problematic FD

- If Dname was prime attribute, or Dnumber was key, then there will not be a problem, and the relation would be in 3NF, right?
- So 3NF can be re-written in general form as:
 - A relation R is in 3NF if, whenever a non-trivial FD $X \rightarrow Y$ exists in R, then either (a) X is a super key of R, or (b) Y is a prime attribute



BCNF (Boyce-Codd Normal Form)

- A stricter form of 3NF
- A relation R is in BCNF if, whenever a non-trivial FD $X \rightarrow Y$ exists in R, then X is a superkey of R
- Non-trivial FD
 - Any FD $X \rightarrow Y$ such that Y is NOT a subset of X
 - $\{S_Reg_Id, S_Name, S_C_GPPA\} \rightarrow \{S_Name, S_CGPA\}$ ← Trivial FD
 - $S_RegID \rightarrow S_Name$ ← Non-trivial FD
 - $S_RegID \rightarrow S_Address$ ← Non-trivial FD



- Consider the relation EMP
- Pname and Dname are two different multivalued attributes, determined by Ename
 - Pname and Dname are independent of each other...
- This is called multivalued dependency
 - This situation arises when two independent 1:N relationships are put into same relation
 - Results in redundancy!

EMP

<u>Ename</u>	<u>Pname</u>	<u>Dname</u>
Smith	X	John
Smith	Y	Anna
Smith	X	Anna
Smith	Y	John

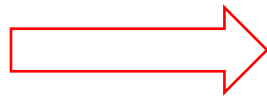


4NF

- A relation is in fourth normal form (4NF) if it is in BCNF and contains no non-trivial multi-valued dependencies
- EMP is not in 4NF
- How do we convert it into 4NF?
- Make separate relations for each multivalued attribute

EMP

<u>Ename</u>	<u>Pname</u>	<u>Dname</u>
Smith	X	John
Smith	Y	Anna
Smith	X	Anna
Smith	Y	John



EMP_PROJECTS

<u>Ename</u>	<u>Pname</u>
Smith	X
Smith	Y

EMP_DEPENDENTS

<u>Ename</u>	<u>Dname</u>
Smith	John
Smith	Anna



- A relation is in 5NF if it has no join dependency
- Join dependency: The situation where joining the decomposed relations produces spurious or inaccurate tuples
- In simpler words, a relation is in 5NF if breaking it down results in relations which cannot be naturally joined to get original relation
- A relation representing a ternary relation can be thought of as being in 5NF
 - If you break it down into three relations representing binary relationships, then joining the three relations cannot reproduce the original relation!



Normalization review

- Normalization goals:
 - Minimizing redundancy
 - Minimizing anomalies



Normalization review

- The process of normalization through decomposition must also ensure the existence of additional properties that the resulting schemas should possess:
 - The **nonadditive join or lossless join property**, which guarantees that the spurious tuple generation problem does not occur with respect to the relation schemas created after decomposition
 - The **dependency preservation property**, which ensures that each functional dependency is represented in some individual relation resulting after decomposition
- The **nonadditive** join property is extremely critical and **must be achieved at any cost**, whereas the dependency preservation property is sometimes sacrificed



Normalization review

- Most database projects acquire the existing designs, called legacy design, and improve upon those designs
- Existing designs are evaluated by applying the tests for normal forms
- Normalization is carried out so that the resulting designs are of high quality and meet the desirable properties stated previously



Normalization review

- The practical utility of 4NF and 5NF normal forms is questionable
 - The constraints on which they are based are rare
 - It is hard for the database designers to understand and detect these constraints
- Database design pays particular attention to normalization only up to 3NF, BCNF, or at most 4NF



Normalization review

- ER modeling can be used to produce an initial relational schema which can then be normalized to remove any remaining redundancies
- If a relational schema is designed using proper ER modeling, then normalization is not usually required
- There is always a tradeoff in normalization
 - Normalization reduces redundancy but may create too many smaller relations which may be too expensive for frequent querying
 - Some redundancy may help in improving the performance
 - This is the typical time versus space tradeoff!



Thanks a lot

